Math 2311

Bekki George – bekki@math.uh.edu

Office Hours: MW 11am to 12:45pm in 639 PGH

Online Thursdays 4-5:30pm

And by appointment

Class webpage: http://www.math.uh.edu/~bekki/Math2311.html

Math 2311
Class Notes for Section 1.3 - 1.5

$var = (st.\ dev)^2$

$st.\ dev = \sqrt{var}$

Section 1.3:

Another important question we want to answer about data is about its spread or dispersion. Roughly speaking, the **population standard deviation**, σ, tells the average distance that data values fall from the mean. The standard deviation is the square root of the **population variance**, σ^2. So, what is the variance? The variance is the average of the squared differences of the data values from the mean.

If N is the number of values in a population with mean μ, and x_i represents each individual value in the population, then the variance is found by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

And the population standard deviation is $\sigma = \sqrt{\sigma^2}$

Most of the time we are not working with the entire population. Instead, we are working with a sample.

Σ = add up

- Sample variance - $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Sample standard deviation - $s = \sqrt{s^2}$

Example:

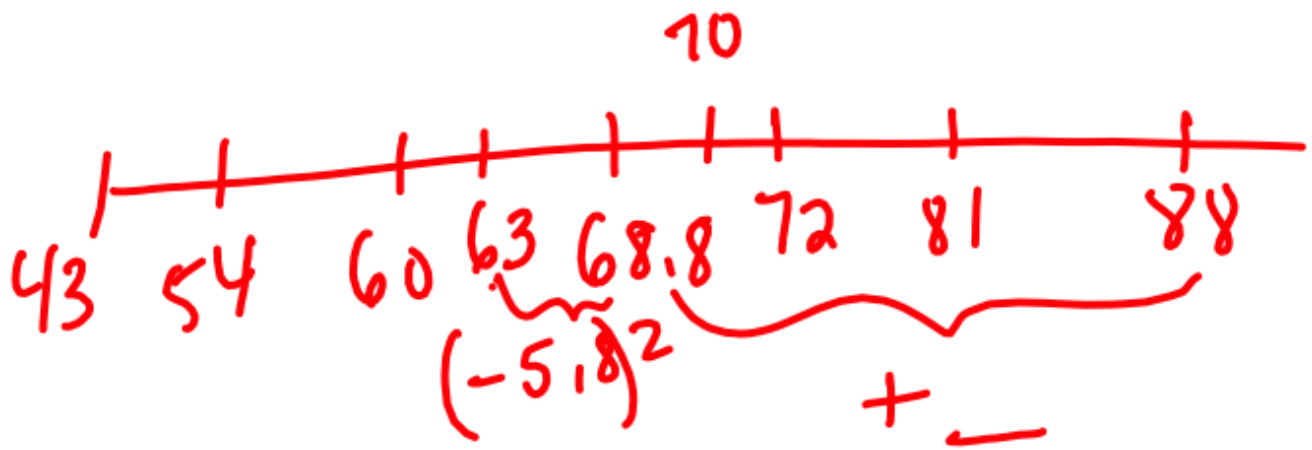
1. A statistics teacher wants to decide whether or not to curve an exam. From her class of 300 students, she chose a sample of 10 students and their grades were:

72, 88, 85, 81, 60, 54, 70, 72, 63, 43

Find the mean, variance and standard deviation for this sample.

$\bar{x} = \dfrac{72 + 88 + \cdots + 43}{10} = 68.8$ $\quad s^2 = \dfrac{(72-68.8)^2 + (88-68.8)^2 + \cdots}{10-1}$

$s = \sqrt{199.7} = 14.13$ $\quad s^2 = 199.7$


70
43
54
60
63
68.8
72
81
88
$(-5.8)^2$
+
—

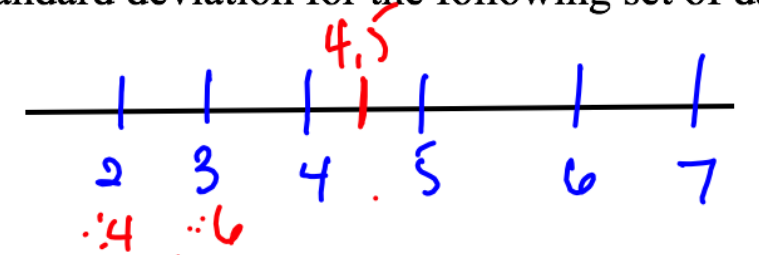
2. Suppose the statistics teacher decides to curve the grades by adding 10 points to each score. What is the new mean, variance and standard deviation?

```
> ex2=ex1+10
> ex2
 [1] 82 98 95 91 70 64 80 82 73 53
> mean(ex2)
[1] 78.8
> sd(ex2)
[1] 14.1327
```

We can see from example 2 that adding the same value to all elements does not affect the variance (or standard deviation) of a set of data. What about multiplying?

3. Find the variance and the standard deviation for the following set of data (whose mean is 4.5)

3, 6, 2, 7, 4, 5



Now, multiply each value by 2. What is the new variance and the new standard deviation?

6, 12, 4, 14, 8, 10

```
> sd(ex3)
[1] 1.870829
```

```
> ex3b=ex3*2
> ex3b
[1]  6 12  4 14  8 10
> sd(ex3b)
[1] 3.741657
```

$$\text{mean}(a x_i) = a\bar{x}$$

$$\text{sd}(a x_i) = a \cdot s_x$$

$$\text{var}(a x_i) = a^2 \cdot s_x^2$$

adding only changes mean

Sometimes we want to compare the variation between two groups. The **coefficient of variation** can be used for this. The coefficient of variation is the ratio of the standard deviation to the mean. A smaller ratio will indicate less variation in the data.

Example:

4. The following statistics were collected on two different groups of stock prices:

	Portfolio A	Portfolio B
Sample size	10	15
Sample mean	$52.65	$49.80
Sample standard deviation	$6.50	$2.95

What can be said about the variability of each portfolio?

$A: \frac{6.50}{52.65} = .123$

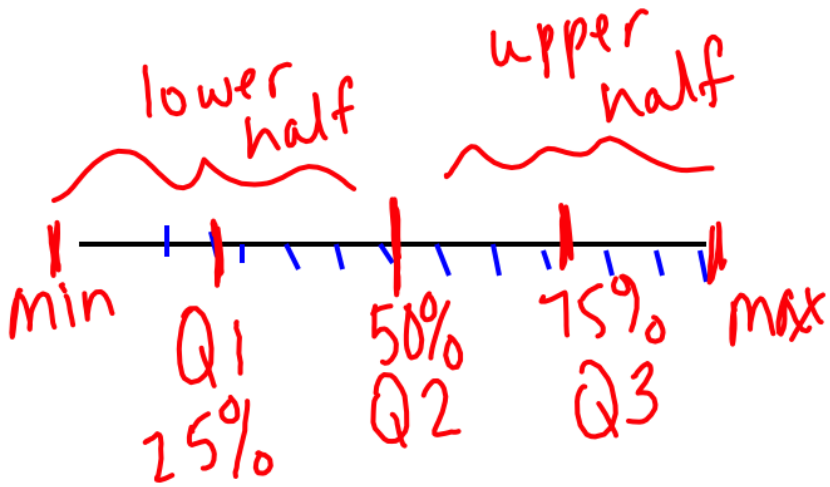
$B: \frac{2.95}{49.80} = .059 \Rightarrow$ less variation

Section 1.4:

More measures of spread (or dispersion):

- Range – highest value minus lowest value

Drawbacks of range: sensitivity to **outliers**



- Percentiles:
  - 25th percentile, Q1 – median of the lower half
  - 50th percentile, Median or Q2 – middle
  - 75th percentile, Q3 – median of the upper half

The values of the minimum, Q1, Q2, Q3 and the maximum make up what is called our **five number summary**.

- IQR – Interquartile range Q3 – Q1 middle 50% of the data

Example:

1. Twelve babies spoke for the first time at the following ages (in months):

8 9 10 11 12 13 15 15 18 20 20 26

1 2 3 4 5 6 7 8 9 10 11 12

Find Q1, Q2, Q3, the range and the IQR.

babies= c(8,9,10,...)

```
> fivenum(babies)
[1]  8.0 10.5 14.0 19.0 26.0
```

Q2 = 14

Q1 = 10.5

Q3 = 19

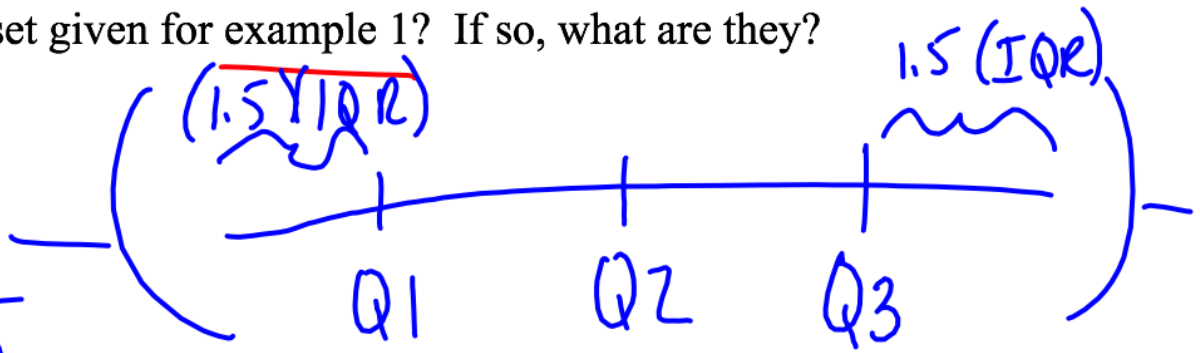
Range = 26 − 8 = 18

IQR = 19 − 10.5 = 8.5

The IQR is used to determine data classified as **outliers**. An outlier is an observation that is "distant" from the rest of the data. Outliers can occur by chance or be measurement errors so it is important to identify them. Any point that falls outside the interval calculated by Q1- 1.5(IQR) and Q3 + 1.5(IQR) is considered an outlier.

Example:

2. Are there any outliers in the data set given for example 1? If so, what are they?



IQR = 8.5

1.5 (IQR) = 12.75

Q1 = 10.5

Q3 = 19

10.5 − 12.75 = −2.25

19 + 12.75 = 31.75

[−2.25, 31.75]

no outliers

There are other percentiles as well. The ***k*th percentile** means that k% of the ordered data values are at or below that data value. For example, if the median is 100, then 50% of the ordered data values fall at or below 100. Also, (100-k)% represents the amount of ordered data that falls above the percentile data value.

If you are looking for the measurement that has a desired percentile rank, the 100Pth percentile, is the measurement with rank (or position in the list) of nP+0.5, where n represents the number of data values in the sample.

Example:

3. In a collection of 30 data measurements, which measurement represents the 30th percentile?

$n = 30$

$P = .30$

$30(.30) + .5 = 9.5$

between 9th + 10th position in list

Suppose you know the position (the order) of a value and want to know what percentile it is ranked at. In general, if you have n data measurements, x_1 represents the $100(1-0.5)/n^{th}$ percentile, x_2 represents the $100(2-0.5)/n^{th}$ percentile, and x_i represents the $100(i-0.5)/n^{th}$ percentile.

Example:

4. Using the data in example 1, determine the percentile of the 4^{th} order statistic (x_4).

$n = 12 \quad i = 4$

$100(4-.5)/12 = 29.2$

$X_4 = 11$ is at the 29.2^{th} percentile

Section 1.5:

Graphs and Describing Distributions

Lets start with two examples, one for categorical data and one for quantitative data:

Example 1: Suppose we have found the eye color for 85 people. 40 have brown eyes, 25 have blue eyes, 10 have green eyes and 10 have hazel eyes. Display the data in a bar graph.

Example 2: Height measurements for a group of people were taken. The results are recorded below (in inches):

66, 68, 63, 71, 68, 69, 65, 70, 73, 67, 62, 59, 63, 68, 71, 63, 63, 60, 64, 66, 58

We will organize these data sets using different graphs:

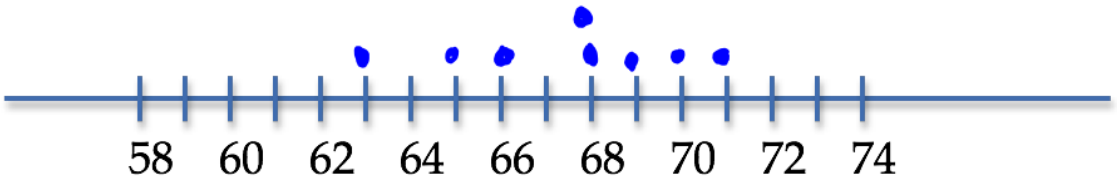
A **bar graph** is created by listing the categorical data along the *x*-axis and the frequencies along the *y*-axis. Bars are drawn above each data value.

To create this in R Studio, enter your "names" list and your "numeric" list:

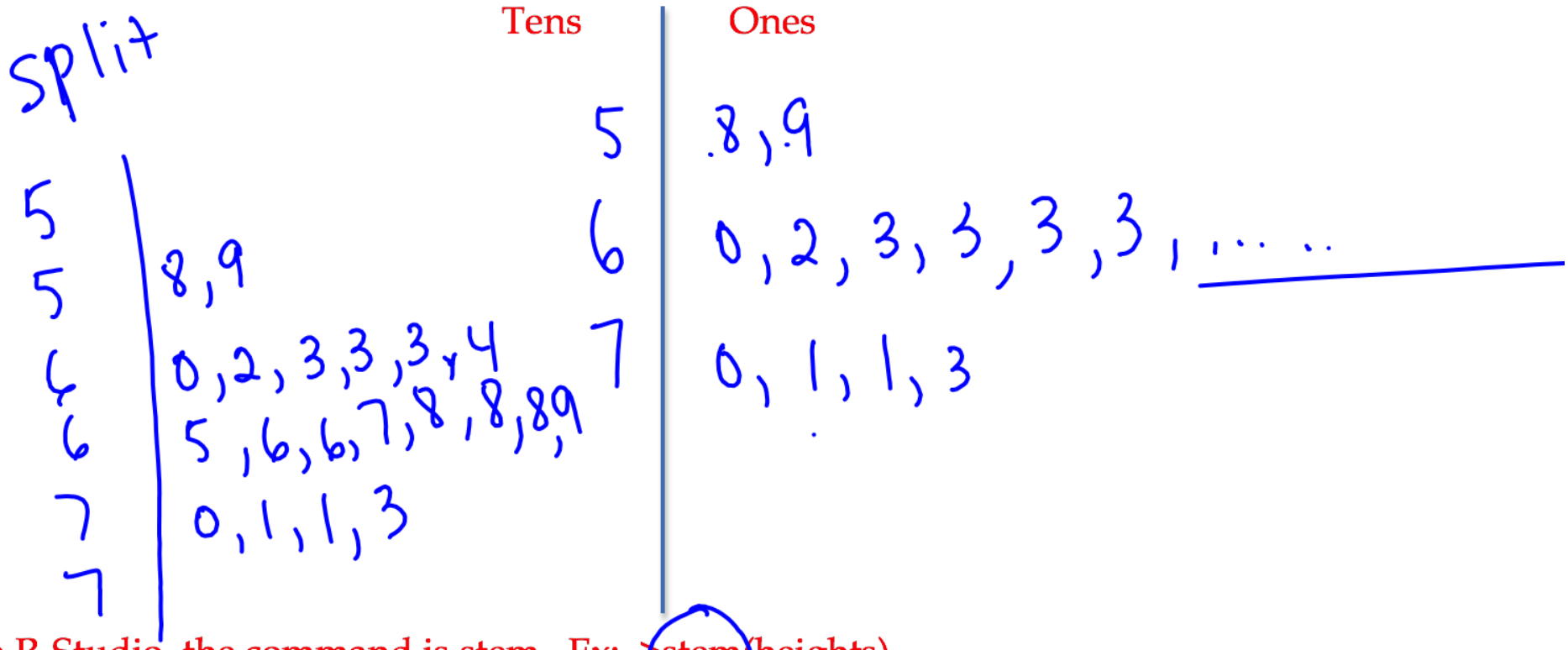
```
>colors=c("brown", "blue", "green", "hazel")
>numcol=c(40, 25, 10, 10)
>barplot(numcol,names=colors)
```

A **dot plot** is made simply by putting dots above the values listed on a number line.

Create a dot plot of example 2



A **stem and leaf plot**, the data is arranged by values. The digits in the largest place are referred to as the stem and the digits in the smallest place are referred to as the leaf (leaves). The leaves are displayed to the right of the stem. A **split stemplot** divides up the stems into equal groups. **Back-to-back stempots** can be used when comparing two sets of data. (using example 2 data)

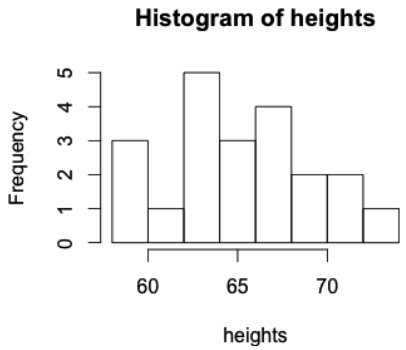


In R-Studio, the command is stem. Ex: >stem(heights)

We can even change scale: >stem(heights, scale=0.5)

Histograms are created by first dividing the data into classes, or bins, of equal width. Next, count the number of observations in each class. The horizontal axis will represent the variable values and the vertical axis will represent your frequency or your relative frequency.

>hist(heights)



>hist(heights, breaks=5)

this means the width of each bin is 5

Boxplots not only help identify features about our data quickly (such as spread and location of center) but can be very helpful when comparing data sets.

How to make a box plot:

1. Order the values in the data set in ascending order (least to greatest).
2. Find and label the median.
3. Of the lower half (less than the median—do not include), find and label Q1.
4. Of the upper half (greater than the median—do not include), find and label Q3.
5. Label the minimum and maximum.
6. Draw and label the scale on an axis.
7. Plot the five number summary.
8. Sketch a box starting at Q1 to Q3.
9. Sketch a segment within the box to represent the median.
10. Connect the min and max to the box with line segments.

Note: If data contains outliers, a **box and whiskers plot** can be used instead to display the data. In a box and whiskers plot, the outliers are displayed with dots above the value and the segments begin (or end) at the next data value within the outlier interval.

>boxplot(heights)

A **pie chart** is a circular chart, divided into sectors, indicating the proportion of each data value compared to the entire set of values. Pie charts are good for categorical data.

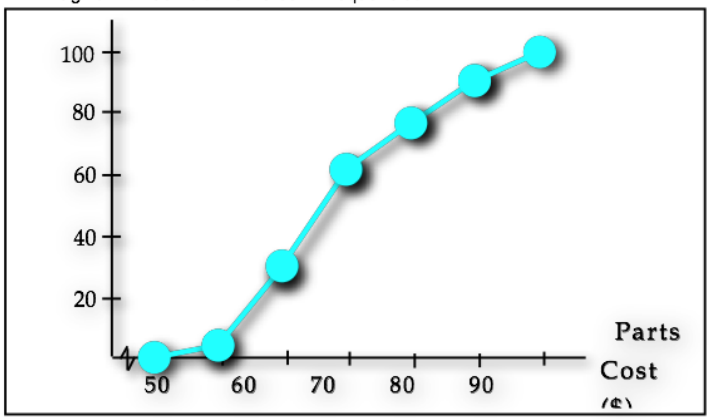
```
>pie(numcols, labels=colors)
```

A **cumulative frequency plot** of the percentages (also called an **ogive**) can be used to view the total number of events that occurred up to a certain value.

Example: Here is an ogive for Hudson Auto Repair's cost of parts sold:

Example: Hudson Auto Repair

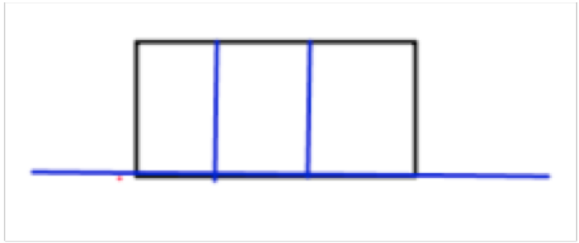
- Ogive with Cumulative Percent Frequencies



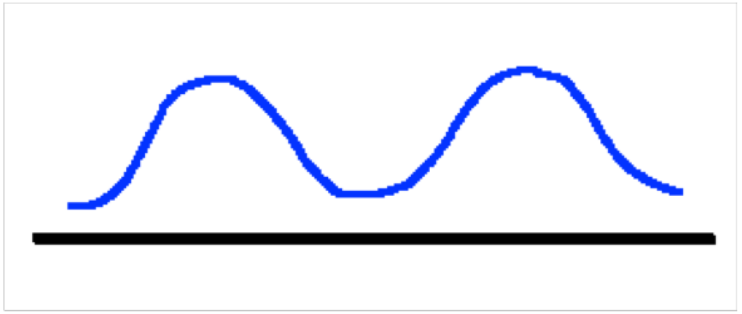
Where is the median of this data?

Patterns and shapes:

Uniform graphs

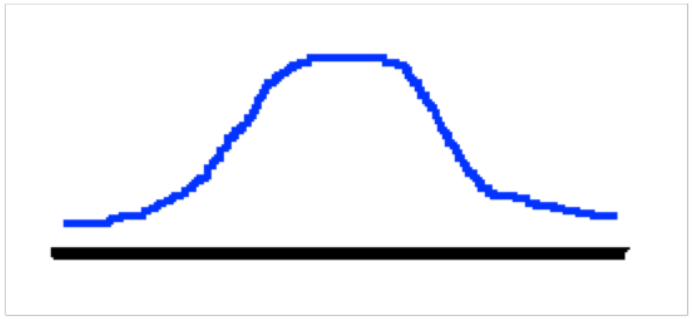


Symmetric graphs

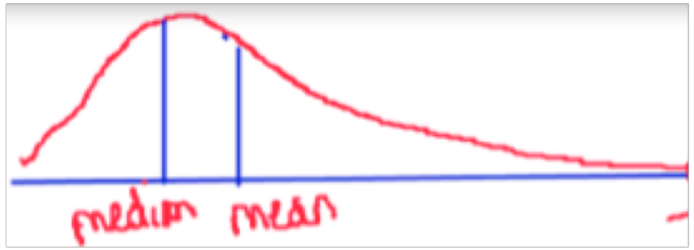


Some other features

Bell Shaped



Skewed right



Skewed left