MATH 3339
Statistics for the Sciences

Wendy Wang, Ph.D.
wwang60@central.uh.edu

Lecture 3 - 3339

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outline

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Chapter 2: Descriptive and Graphical Statistics

A Data Set: Course Grades From Previous Semesters

https://www.math.uh.edu/~cathy/Math3339/data/grades.txt

Student	Score	Grade	Tests	Quiz	HW	Opt-out	Session
1	100.707	A	99.233	87.308	101.270	yes	Sp16
2	81.310	B	75	98.231	64.444	yes	Sp16
3	8.194	F	14.667	12.769	3.175	no	Sp16
4	90.449	A	91.533	77.231	82.222	yes	Sp16
5	68.461	D	65.783	81.769	68.571	no	Sp16
6	103.955	A	103.32	97.923	101.905	yes	Sp16
7	92.889	A	95.6	85.923	75.556	no	Sp16
8	84.805	B	83.2	79.385	75.238	yes	Sp16
9	91.640	A	89.967	91.231	85.079	yes	Sp16
10	22.316	F	17.433	40.615	44.444	no	Sp16
11	98.363	A	94.167	99.231	101.587	yes	Sp16
12	49.250	F	43.917	73.077	78.095	no	Sp16
13	16.967	F	15.5	20.077	29.841	no	Sp16
14	50.747	F	45.533	67.385	57.460	no	Sp16
15	43.184	F	72.983	47.462	38.413	no	Sp16
16	100.845	A	98.667	96.231	100.317	yes	Sp16
17	84.195	B	77.5	87.154	95.556	yes	Sp16
18	84.400	B	78.733	78.615	82.540	yes	Sp16
19	67.170	D	74.3	68.538	72.063	no	Fal15
20	87.413	B	92	82.077	77.778	yes	Fal15
21	67.899	D	71.8	71.077	84.127	no	Fal15
22	74.676	C	70.083	83.308	73.016	no	Fal15
23	40.054	F	44.133	21.308	33.333	no	Fal15
24	101.014	A	101.08	98.923	95.873	no	Fal15
25	11.972	F	17.1	10.385	3.810	no	Fal15
26	79.831	B	86.233	71.923	46.667	no	Fal15
27	83.301	B	94.6	69.692	60.317	no	Fal15
28	72.299	C	64.967	67.615	99.394	no	Sum16
29	83.821	B	77.2	80.923	83.030	yes	Sum16
30	90.703	A	83.617	87.923	80.000	no	Sum16

R studio cloud

Distributions

- When observing a data set, one of the first things we want to know is how each variable is *distributed*.
- The **distribution** of a variable tells us what values it takes and how often it takes these values based on the individuals.
- The distribution of a variable can be shown through tables, graphs, and numerical summaries.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Describing distributions

- An initial view of the distribution and the characteristics can be shown through the graphs.
- Then we use numerical descriptions to get a better understanding of the distributions characteristics.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Distributions for categorical variables

- Lists the categories and gives either the count or the percent of cases that fall in each category.
- One way is a **frequency table** that displays the different categories then the count or percent of cases that fall in each category.
- Then we look at the graphs (bar or pie) to determine the distribution of a categorical variable.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Frequency Tables

Opt-out	Percent
Yes	40%
No	60%

Grade	Percent
A	30%
B	26.67%
C	6.67%
D	10%
F	26.67%

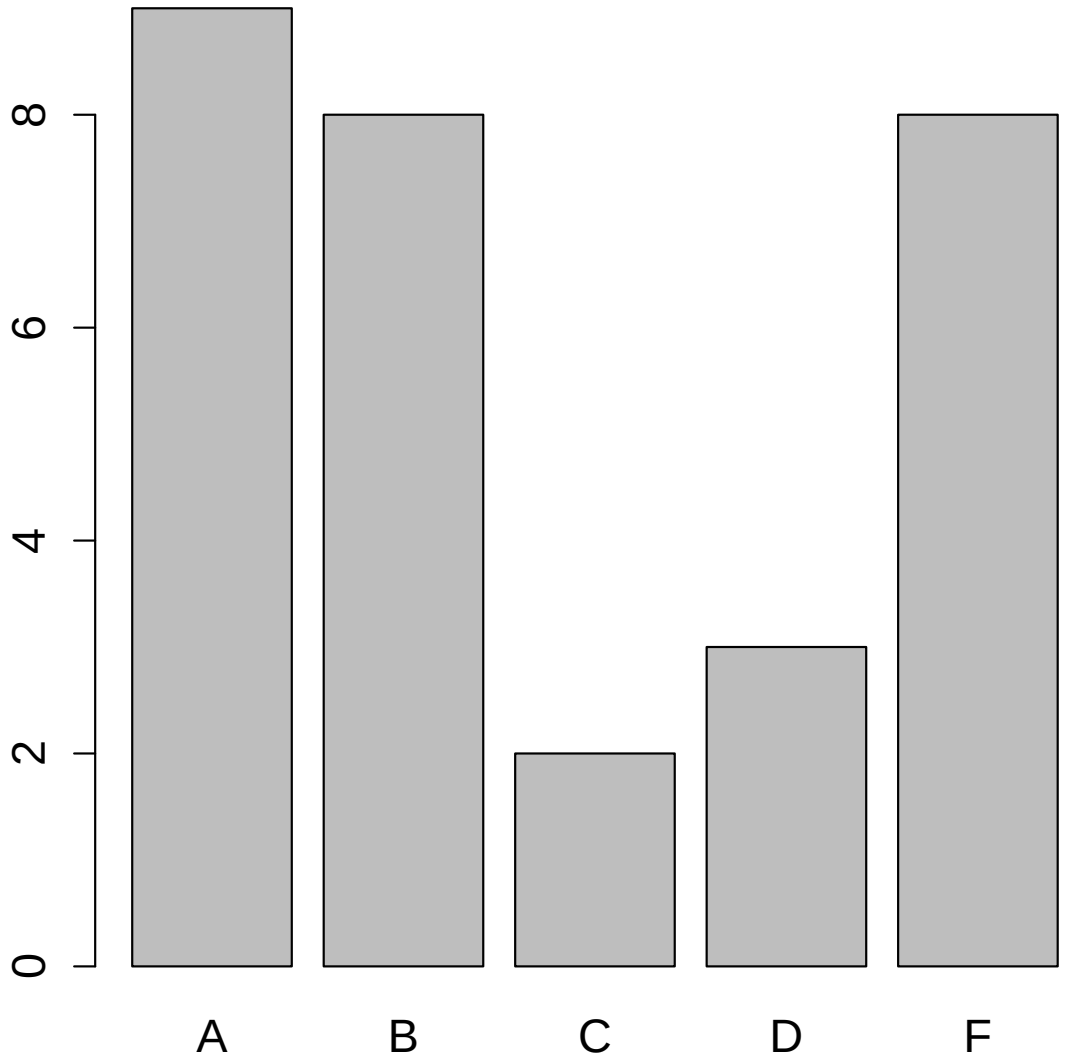
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Describing Data By Graphs

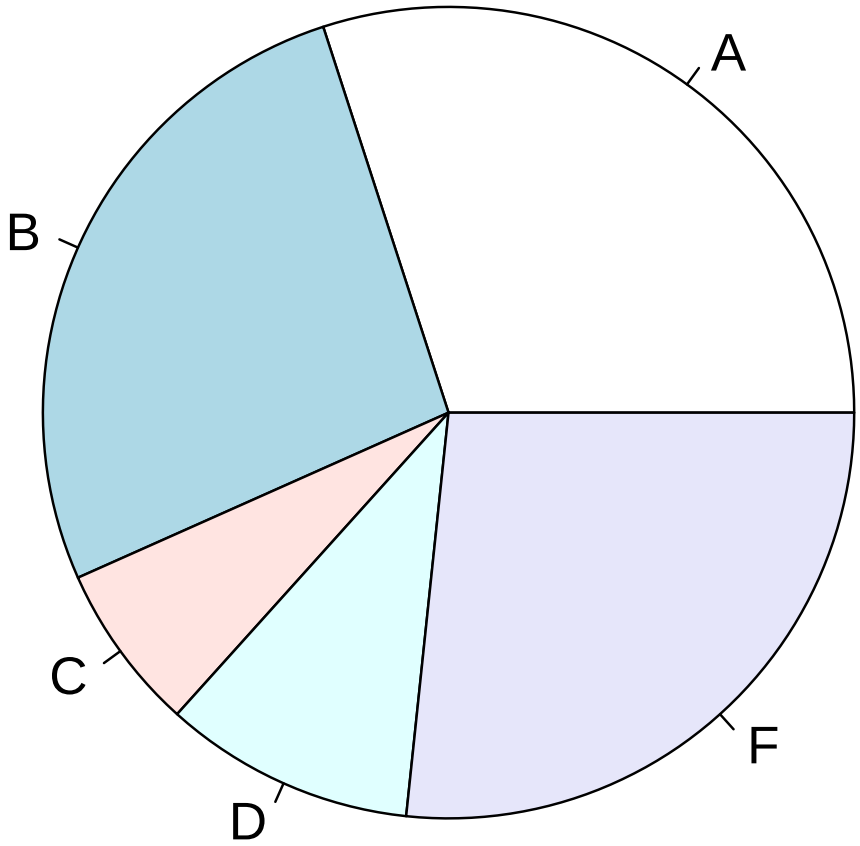
- Graphs are an easy and quick way to describe the data.
- Types of graphs that we use depends on the type of data that we have.
- Graphs for **categorical** variables.
  - **Bar graphs**: Each individual bar represents a category and the height of each of the bars are either represented by the count or percent.
  - **Pie charts**: Helps us see what part of the whole each group forms.
- Graphs for **quantitative** variables.
  - Dotplot
  - Stemplot
  - Histogram
  - Boxplot

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Bar Graph of Letter Grades



Pie Chart of Letter Grades



R code

- First create a table: counts = table(grades$Grade)
- For bar graph: barplot(counts)
- For pie chart: pie(counts)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Describing distributions of quantitative variables

- The **distribution** of a variable tells us what values it takes and how often it takes these values.
- There are four main characteristics to describe a distribution:
  1. Shape
  2. Center
  3. Spread
  4. Outliers

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

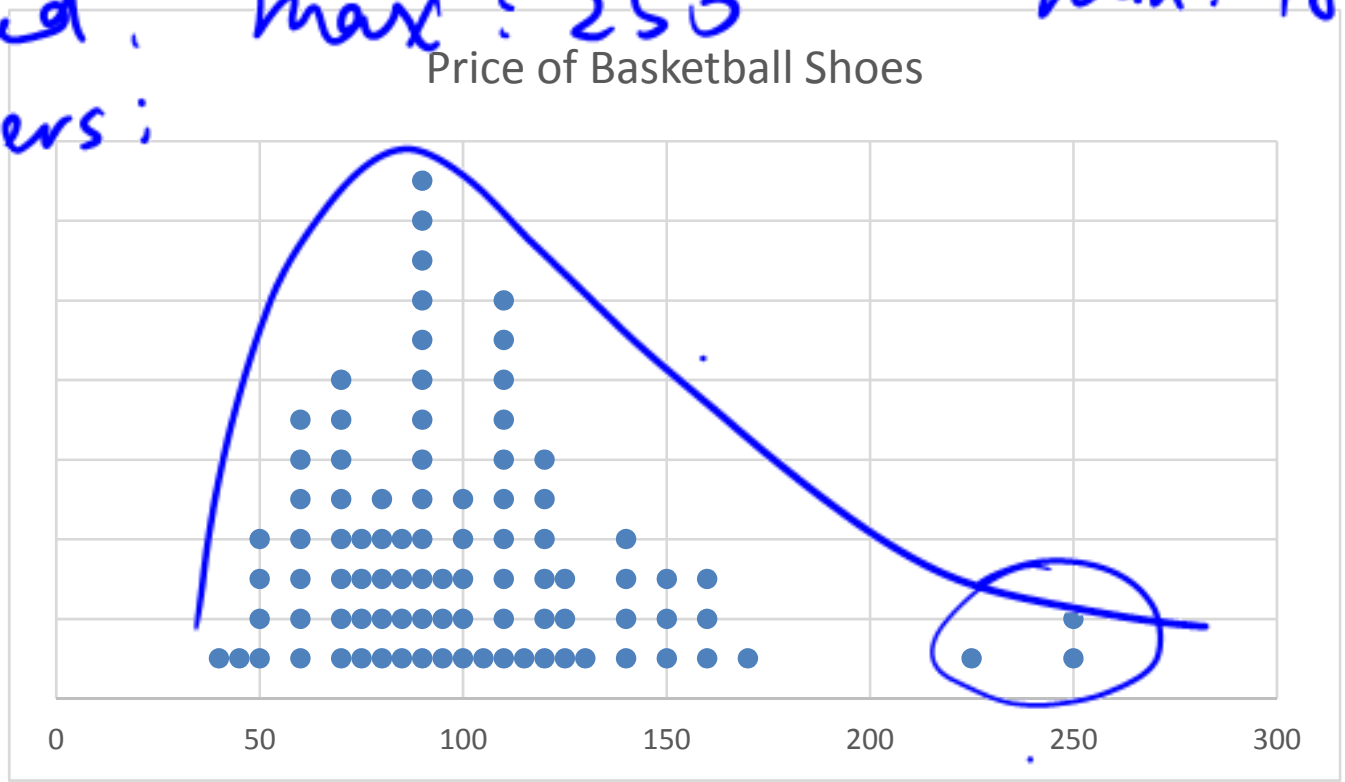
Describing a distribution

- Shape
  - A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
  - A distribution is **skewed to the right** if the right side (higher values) of the graph extends much farther out than the left side.
  - A distribution is **skewed to the left** if the left side (lower values) of the graph extends much farther out than the right side.
  - A distribution is **uniform** if the graph is at the same height (frequency) from lowest to highest value of the variable.
- Center - the values with roughly half the observations taking smaller values and half taking larger values.
- Spread -from the graphs we describe the spread of a distribution by giving *smallest and largest values*.
- Outliers - individual values that falls outside the overall pattern.

Dot plots

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

shape: right-skewed,
center: ~ 90
spread: max: 250 min: 48
outliers:



$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Stem - and - leaf plot

1. Separate each observation into a **stem** consisting of all but the final rightmost digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Rcode: stem(**dataset name$variable name**)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

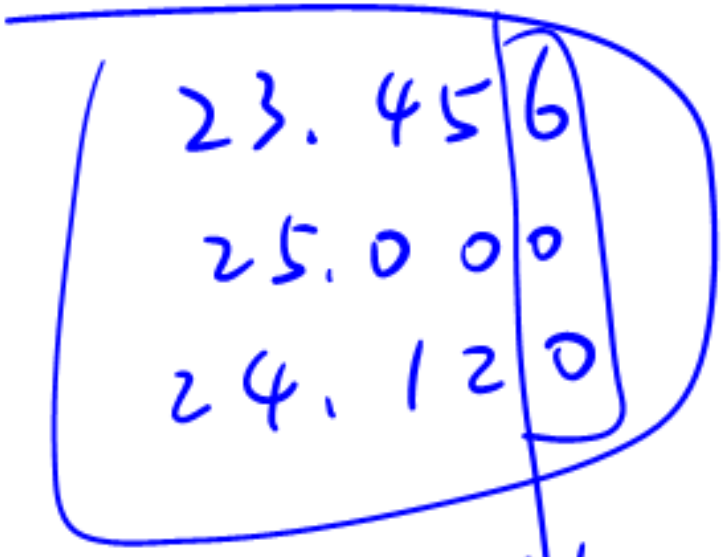
Stem-and-leaf Plot

This is the number of wins out of the 2015 baseball season that each pitcher won.

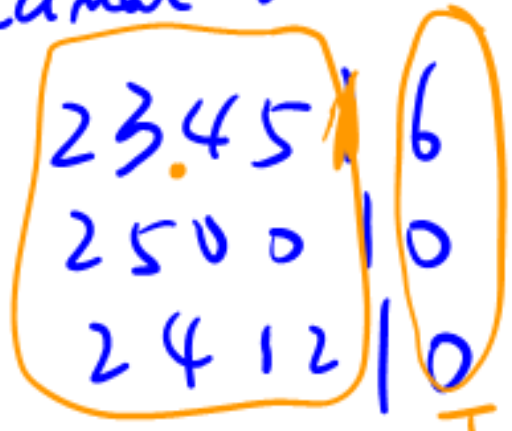
https://www.math.uh.edu/~cathy/Math3339/data/Era.txt

```
> stem(Era$Wins)

  The decimal point is 1 digit(s) to the right of the |

   2 | 679
   3 | 8
   4 | 1246678
   5 | 022223345677889
   6 | 1234577
   7 | 0019
   8 | 6
```

stem leaf



decimal will be 2 digits to the left of "1"



Stem-and-leaf Plot of ERA

```
> stem(Era$ERA)

  The decimal point is at the |

  1 | 78
  2 | 1
  2 | 567889
  3 | 00023344
  3 | 67777889
  4 | 0001111233
  4 | 579
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example of Stem-and-leaf Plot

```
> stem(grades$Score)

  The decimal point is 1 digit(s) to the right of the |

   0 | 8
   1 | 27
   2 | 2
   3 |
   4 | 039
   5 | 1
   6 | 788
   7 | 25
   8 | 01344457
   9 | 01238
  10 | 1114
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Better Plot

```
> stem(grades$Score,scale=0.5)

  The decimal point is 1 digit(s) to the right of the |

   0 | 827
   2 | 2
   4 | 0391
   6 | 78825
   8 | 0134445701238
  10 | 1114
```

1. What is the "shape" of this distribution?

   a) skewed left b) skewed right
   c) symmetric d) uniform

2. What is the approximate center of this distribution?

   a) 50 b) 82 c) 8.5 d) 4

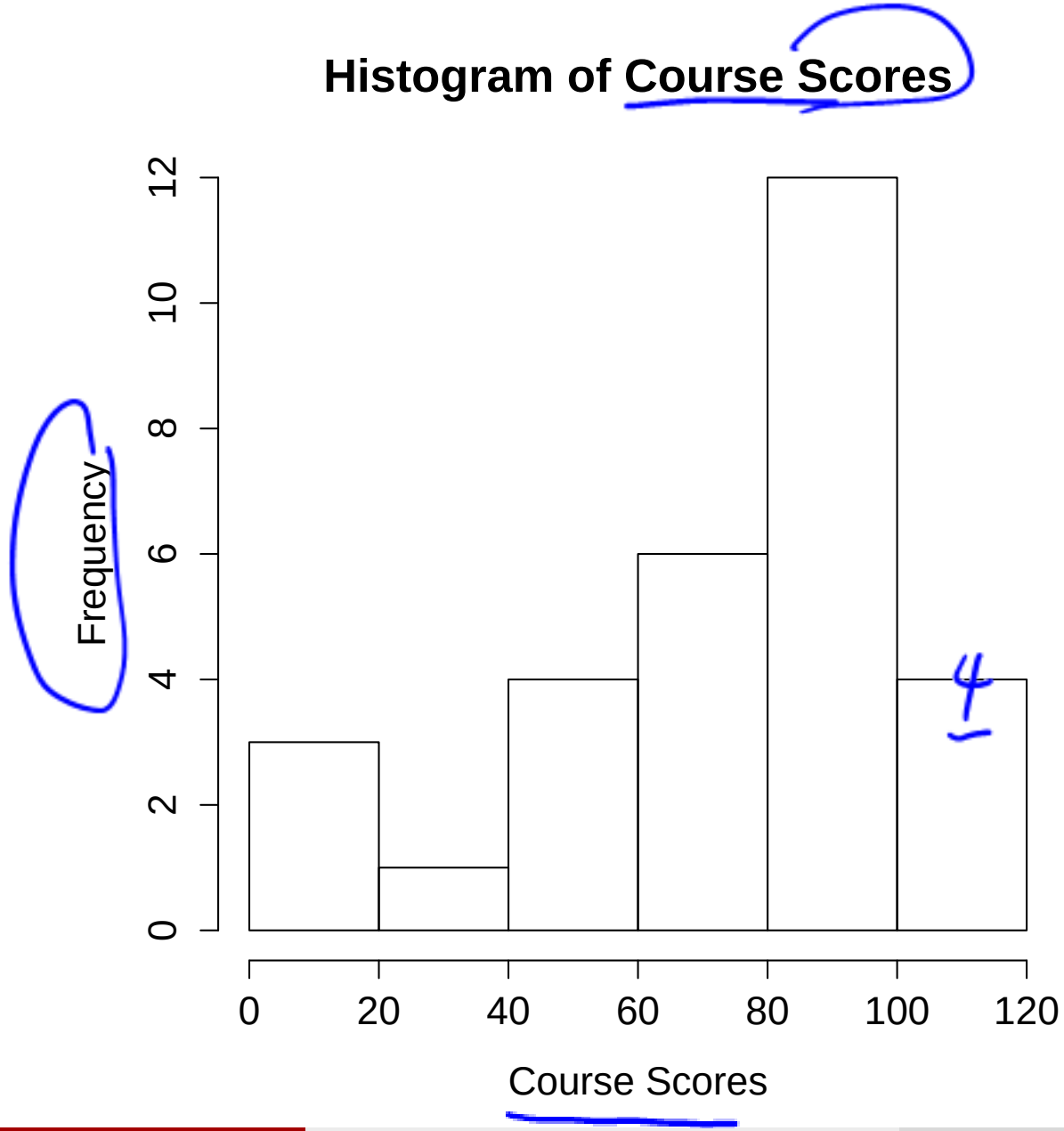
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Histograms

- Bar graph for quantitative variables.
- Values of the variable are grouped together.
- The width of the bar represents an interval of values (range of numbers) for that variable.
- The height of the bar represents the number of cases within that range of values.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Histogram of Course Score



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

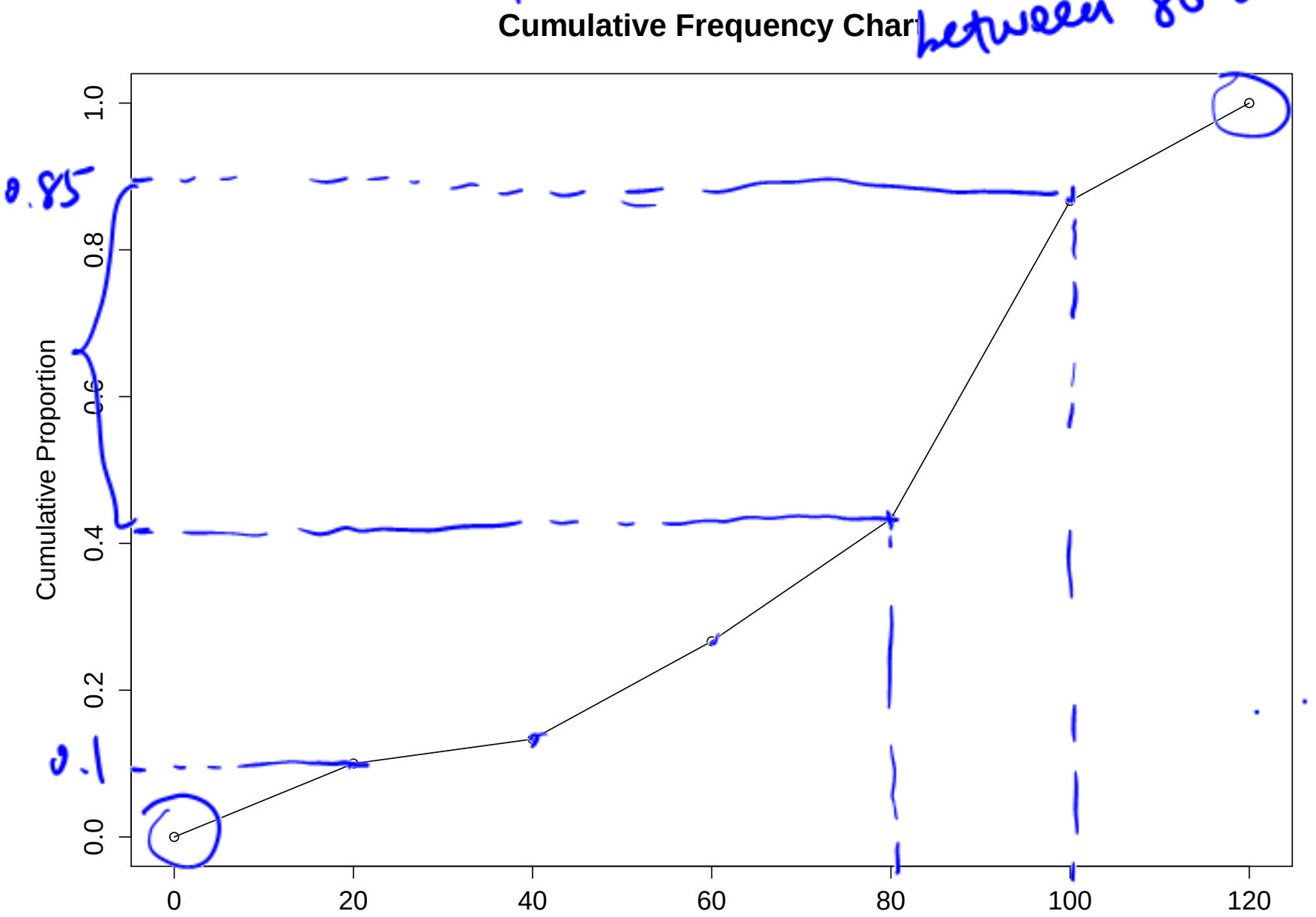
Cumulative Frequency Polygon

- Plot a point above each upper class boundary at a height equal to the cumulative frequency of the class.
- Connect the plotted points with line segments.
- A similar graph can be used with the cumulative percents.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Cumulative Percent Polygon

~ 45% of students with scores between 80 and 100



Describing Quantitative Variables with Numbers

- Center - mean, median or mode
- Spread - range, interquartile range, variance, or standard deviation
- Location - percentiles or standard scores

IQR

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Parameters and Statistics

- A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we usually do not know its value.
- A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample.
- The purpose of sampling or experimentation is usually to use statistics to make statements about unknown parameters, this is called **statistical inference**.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Notation of Parameters and Statistics

Name	Statistic	Parameter
mean	$\bar{x}$	μ mu
standard deviation	s	σ sigma
correlation	r	ρ rho
regression coefficient	b	β beta
proportion	$\hat{p}$	p

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Measuring center: The mean

- Most common measure of center.
- To calculate the mean of a set of observations $x_1, x_2, \ldots, x_n$, add their values and divide by the number of observations ***n***.
- Denoted: $\bar{x}$ called x-bar if the data is from a sample, μ, called "mu" if the data is from the entire population.

n: sample size

sample:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

population:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

N: population size

- Where ***n*** is the size of the sample and ***N*** is the size of the population.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Measuring center: The Median

The **median** M is the midpoint of a data set such that half of the observations are smaller and the other half are larger.

1. Arrange all observations in order of size, from smallest to largest.
2. Find the middle value of the arranged observations by counting $(n+1)/2$ from the bottom of the list.
   - If the number of observations ***n*** is odd, the median M is the the center observation in the ordered list.
   - If the number of observations ***n*** is even, the median M is the mean of the two center observation in the ordered list.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Measuring Center: The Mode

1 2 2 2 3 3 3 5 6

mode: "2" & "3"

1 2 3 4 5 6

- The **mode** of a data set is the numerical value that appears the most frequently.
- The data set can have one mode, two or more modes.
- A data set may not have any mode.

1 2 3 3 3 4 5 | "3"

1 2 3 3 3 3 5 | "3"

Cacluate the mean, median and mode

The following is a stem-and-leaf plot of the course scores. Determine, the mean, medain and mode of the course scores.

```
The decimal point is 1 digit(s) to the right of the |

0  |  8
1  |  27
2  |  2
3  |
4  |  039
5  |  1
6  |  788
7  |  25
8  |  01344457
9  |  01238
10 |  1114
```

(8, 12, 17, 22, 40, 43, ... , 101, 104)

mode(s): 84, 101

```
> scores=c(8,12,17,22,40,43,49,51,67,68,68,72,75,80,8
1,83,84,84,84,85,87,90,91,92,93,98
+       ,101,101,101,104)
> mean(scores)
[1] 71.03333
> median(scores)
[1] 82
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Finding mean and median in R

```
scores=c(8,12,17,22,40,43,49,51,67,68,68,72,75,80,81,
83,84,84,84,85,87,90,91,92,93,98,101,101,101,104)
mean(scores)
[1] 71.03333
median(scores)
[1] 82
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example: Test Scores

The test scores of a class of 20 students have a mean of 71.6 and the test scores of another class of 14 students have a mean of 78.4. Find the mean of the combined group.

$$\frac{(20 * 71.6) + (14 * 78.4)}{(20 + 14)}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

The following are ages of automobiles.

8	3	6	5	5	2	
10	9	8	2	3	2	2

Determine the mean, median and mode of this set.

```
> age=c(8,3,6,5,5,2,10,9,8,2,3,2,2)
> age
 [1]  8  3  6  5  5  2 10  9  8  2  3  2  2
> mean(age)
[1] 5
> median(age)
[1] 5
> table(age)
age
 2  3  5  6  8  9 10
 4  2  2  1  2  1  1
```

mean: 5

median: 5

mode: 2.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Mean vs. Median

- If the mean and the median are both numbers that describe the center of the values then why do we have different values?
- If the data has values that are **outliers** values that are beyond the range of the others, the mean is going toward these outliers.
- The median is resistant to extreme values (outliers) in the data set.
- **Trimmed Mean**
  Trimmed means are obtained by finding the mean of the values of the data excluding a given percentage of the largest and smallest values. For example, the 5% trimmed mean is the mean of the values of the data excluding the largest 5% of the values and the smallest 5% of the values.

  R commands:

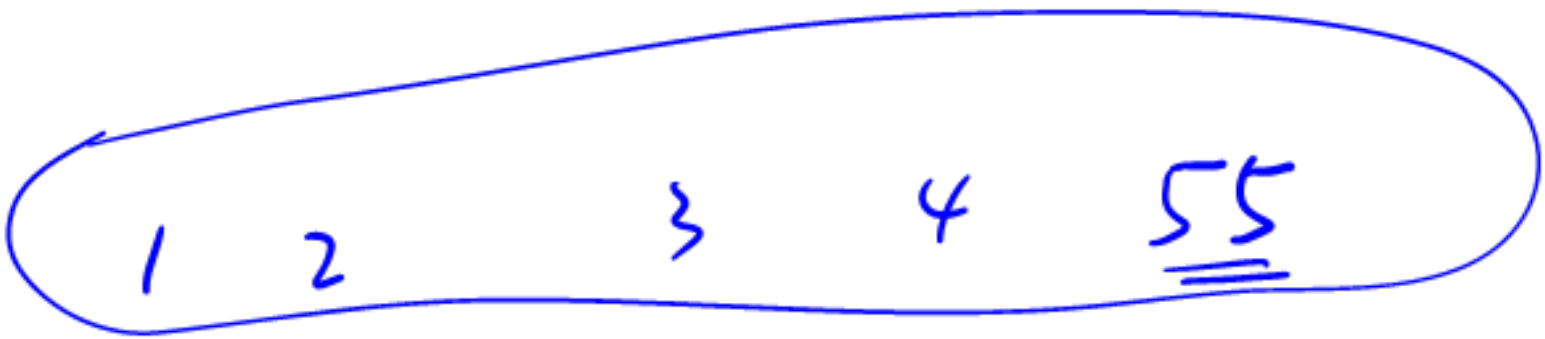
  >mean(data, trim=0.1)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

1 2 3 4 5

mean: 3

median: 3



mean: ? 13

median: 3

Average Test Scores?

What is the mean and median for each of these sections' test scores?

mean (A) = 71.5

mean (B) = 71.5

median (A) = 72

median (B) = 72

Section A	Section B
65	42
66	54
67	58
68	62
71	67
73	77
74	77
77	85
77	93
77	100

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Types of Measurements for the Spread

- Range
- Percentiles
- Quartiles
- IQR; Interquartile range
- Variance
- Standard deviation
- Coefficient of Variation

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

The Range

- The range is the difference between the highest and lowest values.
- Section A: Range = 77 - 65 = 12
- Section B: Range = 100 - 42 = 58

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Percentiles

- The *p***th percentile** of data is the value such that p percent of the observations fall at or below it.
- The median is also a percentile value - the 50^{th} percentile.
- The use of percentiles to report spread when the median is our measure of center.
- If you are looking for the measurement that has a desired percentile rank, the $100P^{th}$ 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.

  For example, in a collection of 30 data measurements, which measurement represents the 25^{th} percentile?

$n = 30 \quad p = 0.25$

$np + 0.5 = 8$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Determine Percentiles

- Suppose you know the position (order) of a value and want to know what percentile it is ranked at.
- If you have n data measurements, x_i, represents the $100(i-0.5)/n^{th}$ percentile.
- Example: Determine the percentile of the 4^{th} order statistic for sample size of $n = 15$.

$i = 4$

$$\frac{100(4-0.5)}{15} = 23.3\%.$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

The Quartiles

Q_1 Q_2 Q_3

- The first quartile is 25th percentile, Q_1.
- The second quartile is the median and the 50th percentile, Q_2.
- The third quartile is the 75th percentile, Q_3.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

A manufacturer claims that his fabric consists of 80 percent cotton. To check his claim, we take a small swatch from each bolt of fabric and determine its cotton content. The results of 25 such measurements are as follows:

77 81 76 76 79 79 80 77 89 77 78 85 80
75 79 88 81 78 82 80 76 83 81 85 79

Determine the percentile of the 4th order statistic. $\frac{100(4-0.5)}{25}\%$
Determine the 50th percentile. median
Determine the first and third quartiles.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

A manufacturer claims that his fabric consists of 80 percent cotton. To check his claim, we take a small swatch from each bolt of fabric and determine its cotton content. The results of 25 such measurements are as follows:

75 76 76 76 77 77 77 78 78 79 79 79
79 80 80 80 81 81 81 82 83 85 85 88 89

Determine the percentile of the 4th order statistic.
Determine the 50th percentile.
Determine the first and third quartiles.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Quantiles

R commands:

>quantile(fabric)

>quantile(fabric,.95)

>quantile(fabric,c(.3,.6,.9))

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

R-code for finding Q_1, Q_2, & Q_3

The values: Minimum, Q_1, Median (Q_2), Q_3, and Maximum are called the **Five Number Summary**

```
> shoeprice=c(100,110,120,120,140,140,140,150,
185,185,215,215,250,250,290)
> fivenum(shoeprice)
[1] 100 130 150 215 290
```

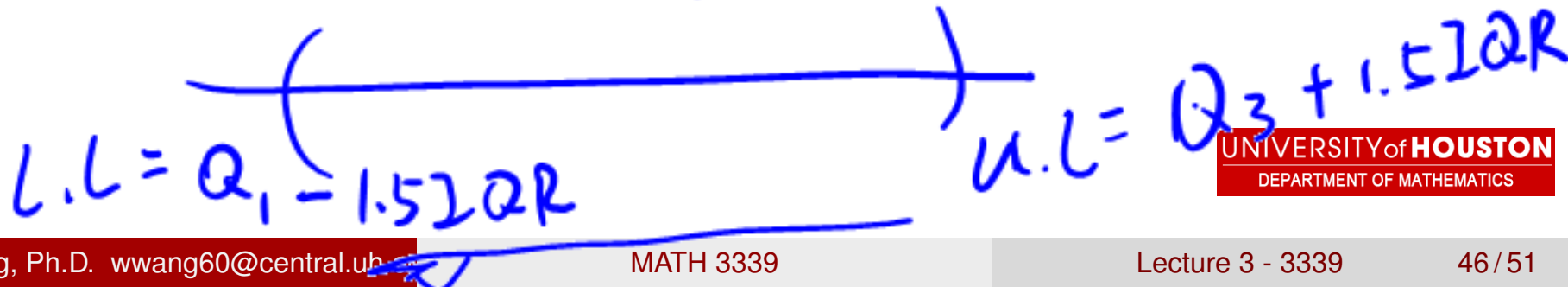
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Detecting Outliers: 1.5IQR Rule

- Interquartile range, **IQR**, is the difference between Q_3 and Q_1

$$IQR = Q_3 - Q_1$$

- An **outlier** is an observation that is "distant" from the rest of the data.
- Outliers can occur by chance or by measurement errors.
- Any point that falls outside the interval calculated by $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$ is considered an outlier.



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outliers for Basketball Shoe Prices?

- Recall: $Q_1 = 130$, $Q_3 = 215$, So IQR = 215 - 130 = 85.
- $Q_1 - 1.5(IQR) = 130 - 1.5(85) = 2.5$
- $Q_3 + 1.5(IQR) = 215 + 1.5(85) = 342.5$
- Any price that is below \$2.50 or above \$342.50 is considered an outlier.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outliers?

The following is information from 91 pairs of basketball shoes:

```
> fivenum(shoes$Price)
[1]  40  75  90 120 250
```

The highest four numbers in the dataset is $\ldots, 170, 225, 250, 250$. Are there any prices that are considered an outlier?

$IQR = Q_3 - Q_1 = 120 - 75 = 45$

Lower limit $= L.L. = Q_1 - 1.5\,IQR$

$= 75 - 1.5(45) = 7.5$

Upper limit $= U.L = Q_3 + 1.5\,IQR$

$= 120 + 1.5\,IQR = 187.5$

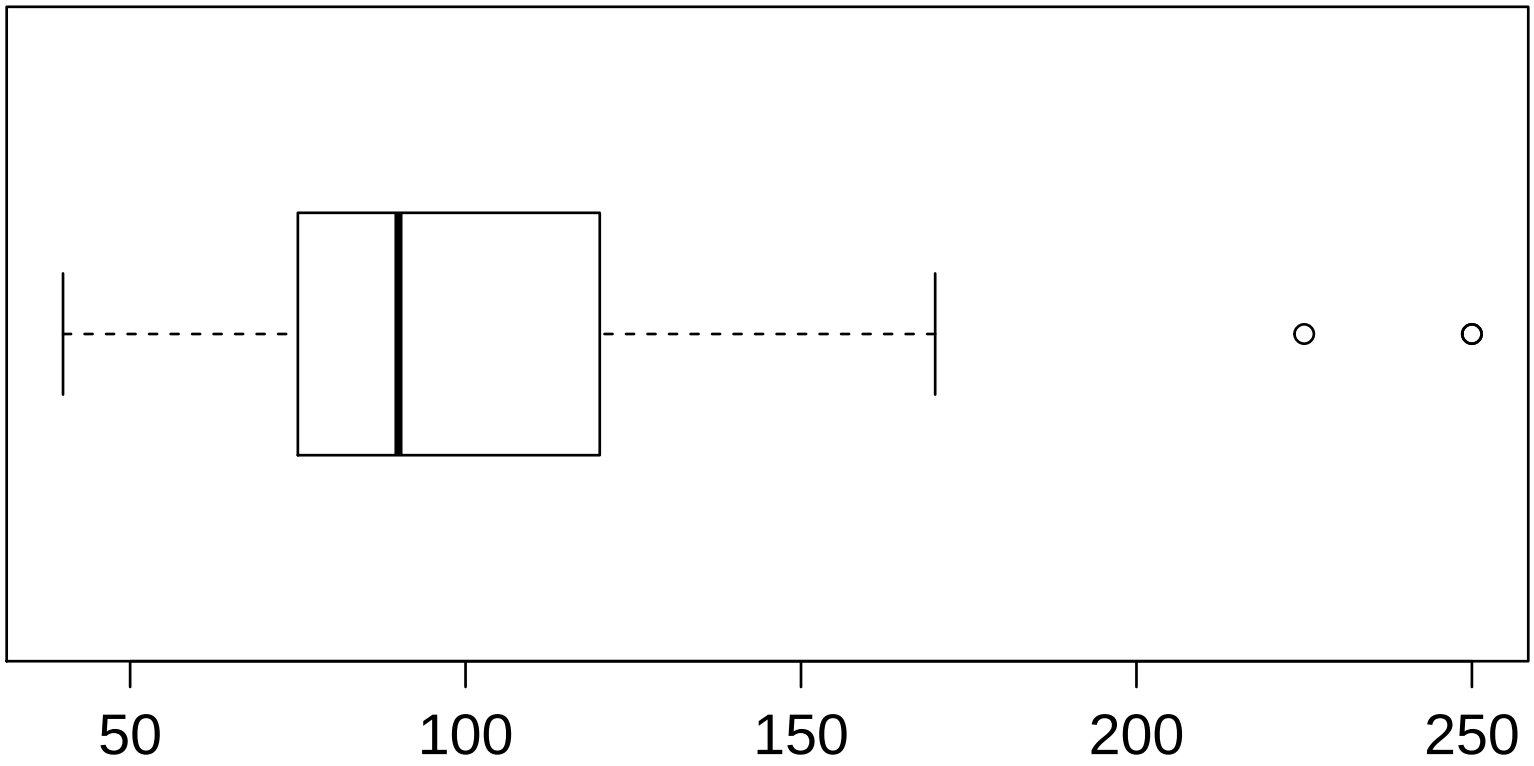
UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

A Graph of the Five Number Summary: Boxplot

- A central box spans the quartiles.
- A line inside the box marks the median.
- Lines extend from the box out to the smallest and largest observations.
- Asterisks represents any values that are considered to be outliers.
- Boxplots are most useful for side-by-side comparison of several distributions.
- Rcode: boxplot(**dataset name$variable name**)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

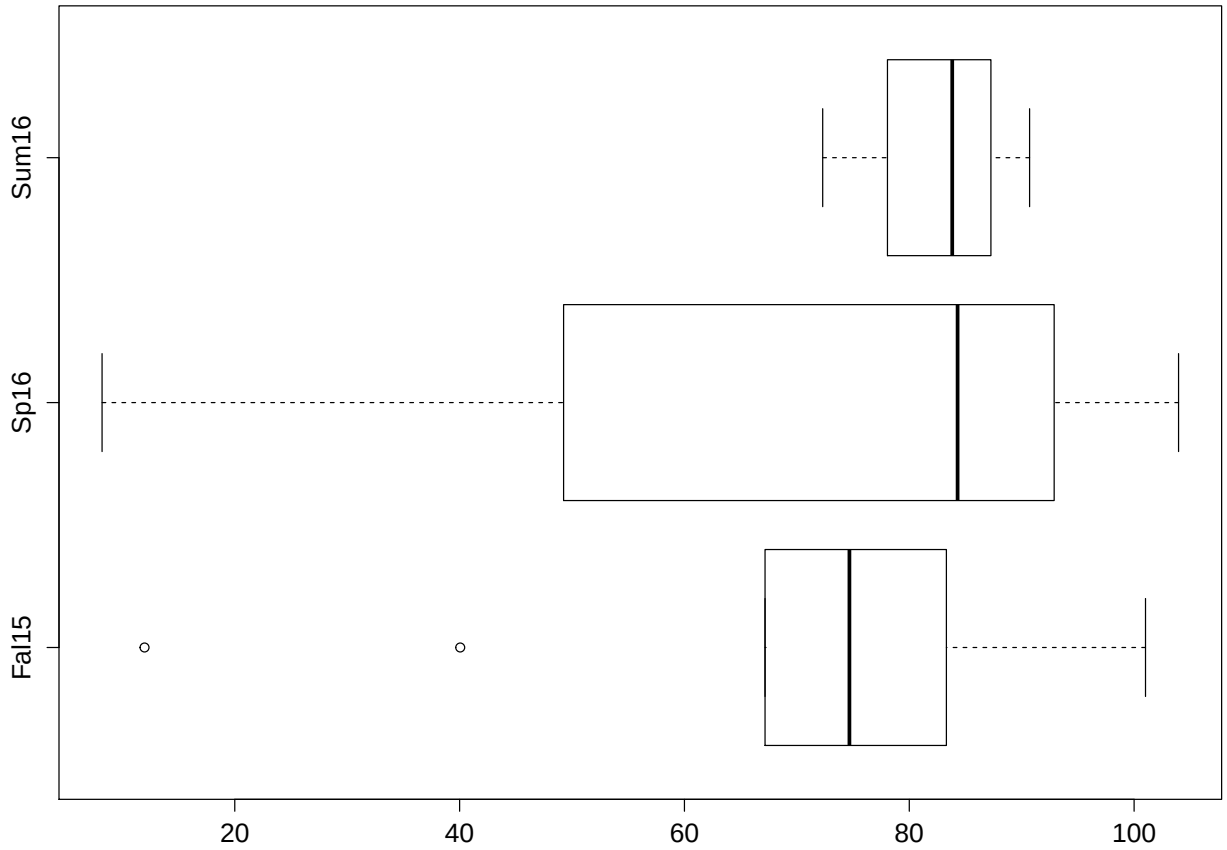
Boxplot of Prices




```
boxplot(shoes$Price,horizontal = T)
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Boxplot of Course Scores by Session




```
boxplot(grades$Score~grades$Session,horizontal=TRUE)
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS