

STT200 – Lecture 1, Section 02,04.

Recitation 2

(September 11th. 2012)

TA: Zhen (Alan) Zhang

zhangz19@stt.msu.edu

Office hour: (C500 WH) 1:45 – 2:45PM Tuesday
(office tel: 432-3342)

Help-room: (A102 WH) 11:20AM-12:30PM, Monday, Friday

Class meet on Tuesday:

3:00 – 3:50PM A122 WH, Section **02**

12:40 – 1:30PM A322 WH, Section **04**

Reminders

- ❑ Recitation class webpage set up ([Click](#)):
 - Problem list for discussion will be posted every *Friday evening* for you to prepare for the recitation on the following *Tuesday*.
 - Notes and power points for recitations will be posted after *4PM on Tuesday*.

- ❑ Don't forget signing the attendance sheet.

- ❑ Pay attention to the different definitions of quartiles from the class note and a calculator.

Overview

For this recitation we are going to discuss:

□ **Chapter 4:** (Page 79) # 14, 24, 30, 31

□ **Chapter 5:** (Page 109) # 7, 8, 13, 14, 16, 20, 31

Chapter 4: #14

Super Bowl points:

25, 19, 9, 16, 3, 21, 7, 17, 10, 4, 18, 17, 4, 12, 17, 5, 10, 29, 22, 36, 19, 32, 4, 45, 1, 13, 35, 17, 23, 10, 14, 7, 15, 7, 27, 3, 27, 3, 3, 11, 12

Median?

Lower Quartile?

Upper Quartile?

Chapter 4: #14 *Strategy*:

□ First sorted from lowest value to largest:

1 3 3 3 3 4 4 4 5 7 7 7 9 10 10 10 11 12 12 13 14
15 16 17 17 17 17 18 19 19 21 22 23 25 27 27 29 32 35
36 45

□ Count the # of total observations: 41 which is odd. So there is a unique value lying in the “middle”, which is 21st number: **14 (median)**

□ For the first 21 numbers, there is a unique value lying in the “middle”, which is 11th number: **7 (lower quartile)**

□ Similarly, for the last 21 numbers, the “middle”: 31st number: **21 (upper quartile)**

Chapter 4: #24

Annual number of death

53,39,39,33,69,30,25,67,130,94,40

Mean?

Median?

Lower Quartile?

Upper Quartile?

Range?

IQR?

Chapter 4: #24

Annual number of death

53,39,39,33,69,30,25,67,130,94,40

Mean? (56.27273)

Median? (40)

Lower Quartile? (36)

Upper Quartile? (68)

Range? (105)

IQR? (32)

Chapter 4: #30

Standard Deviation

set 1

a) 4,7,7,7,10

b) 100,140,150,160,200

c) 10,16,18,20,22,28

set 2

4,6,7,8,10

10,50,60,70,110

48,56,58,60,62,70

Compare Standard Deviations?

Chapter 4: #30

□ Standard Deviation

set 1

set 2

a) 4,7,7,7,10 (2.12132)

4,6,7,8,10 (2.236)

b) 100,140,150,160,200 (36) 10,50,60,70,110 (36)

c) 10,16,18,20,22,28 (6.03) 48,56,58,60,62,70 (7.24)

Tips:

1. For (a), data are more alike, more similar to each other, or more centered around its mean level. Hence less spread, smaller SD.

2. Add/subtract a constant amount to all the data will not change its spread! I.e., SD will not change. This is (b).

3. By (b), first set has the same SD as 50,56,58,60,62,68, which is clearly more centered towards its mean level comparing to set 2: 48,56,58,60,62,70. Hence the SD of set 1 is less.

Chapter 4: #31

- Wage: 1200, 700, 400 (6), 500 (4)
- Mean and Median?
- How many above mean?
- Mean or Median, which better?
- Range, IQR or standard deviation?

Chapter 4: #31

- ❑ Wage: 1200, 700, 400 (6), 500 (4)
- ❑ Mean and Median? (\$525, \$450)
- ❑ How many above mean? (2)
- ❑ Mean or Median, which better? (Median, due to outlier)
- ❑ Range, IQR or standard deviation? (IQR, least sensitive to the outlier of \$1200)

*Tips: Don't forget the **unit** for center measures (mean, median), quartiles, min, max, range, IQR and SD when available*

Chapter 5 (Page 109): #7

Lengths of all the golf courses

count = 45,

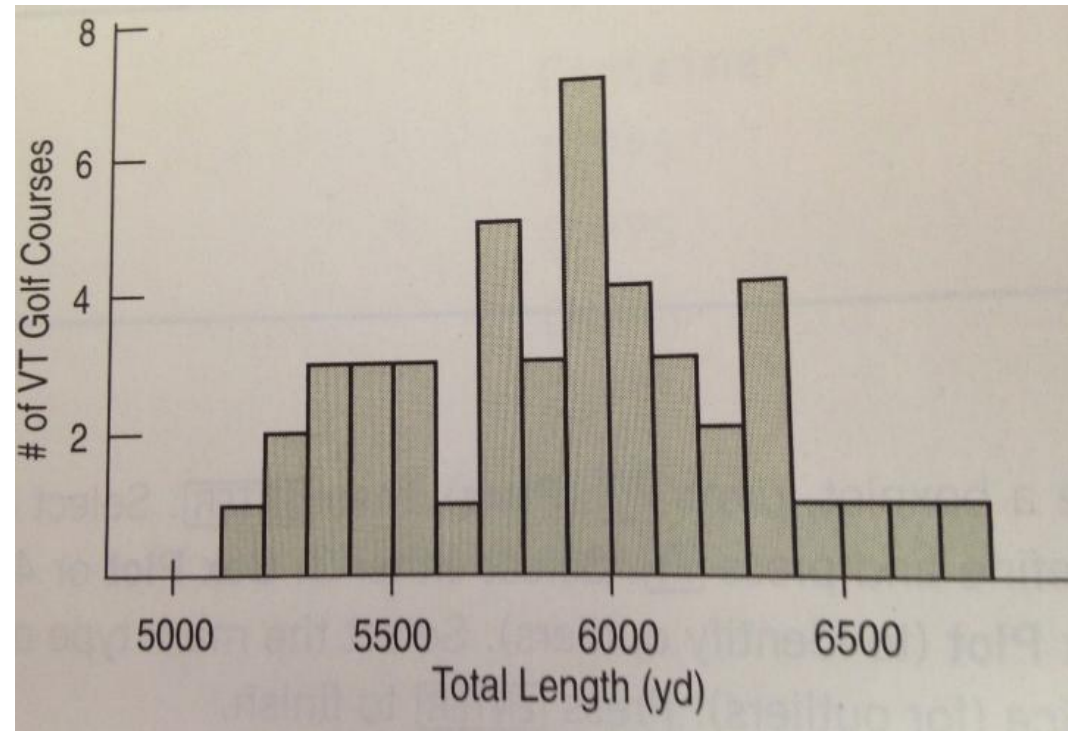
mean = 5892.91yd,

sd=386.59

min = 5185,

Q1 = 5585.75, Median=5928

Q3 = 6131, Max=6796



Range?

Between what lengths do the central 50% lie?

What summary statistics would you use?

Brief description of data (shape, center, and spread)

Chapter 5: #7

❑ Lengths of all the golf courses

count = 45, mean = 5892.91yd, sd=386.59

min = 5185, Q1 = 5585.75, Median=5928

Q3 = 6131, Max=6796

❑ Range? ($6796 - 5185 = 1611$ yd)

❑ Between what lengths do the central 50% lie? (Q1-Q3)

❑ What summary statistics would you use? (mean, sd.

Since unimodal and symmetric)

❑ Brief description of data (shape, center, and spread):

symmetric, unimodal, mean and median are roughly the same due to symmetry, indicating a center around 5900yd.

The standard deviation, the range.

Chapter 5: #8

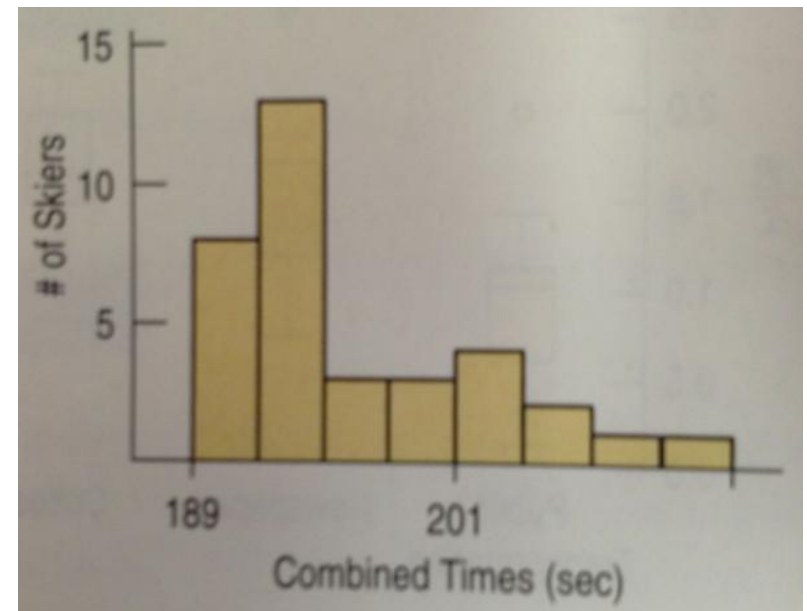
- Histograms of combined time for skiers

count = 35, mean = 196.079, sd = 5.80009

min = 189.35, Q1 = 192.238, Median = 193.270

Q3 = 200.625, Max = 211.890

- Range?
- Between what lengths do the central 50% lie?
- What summary statistics would you use?
- Brief description of data (shape, center, and spread)



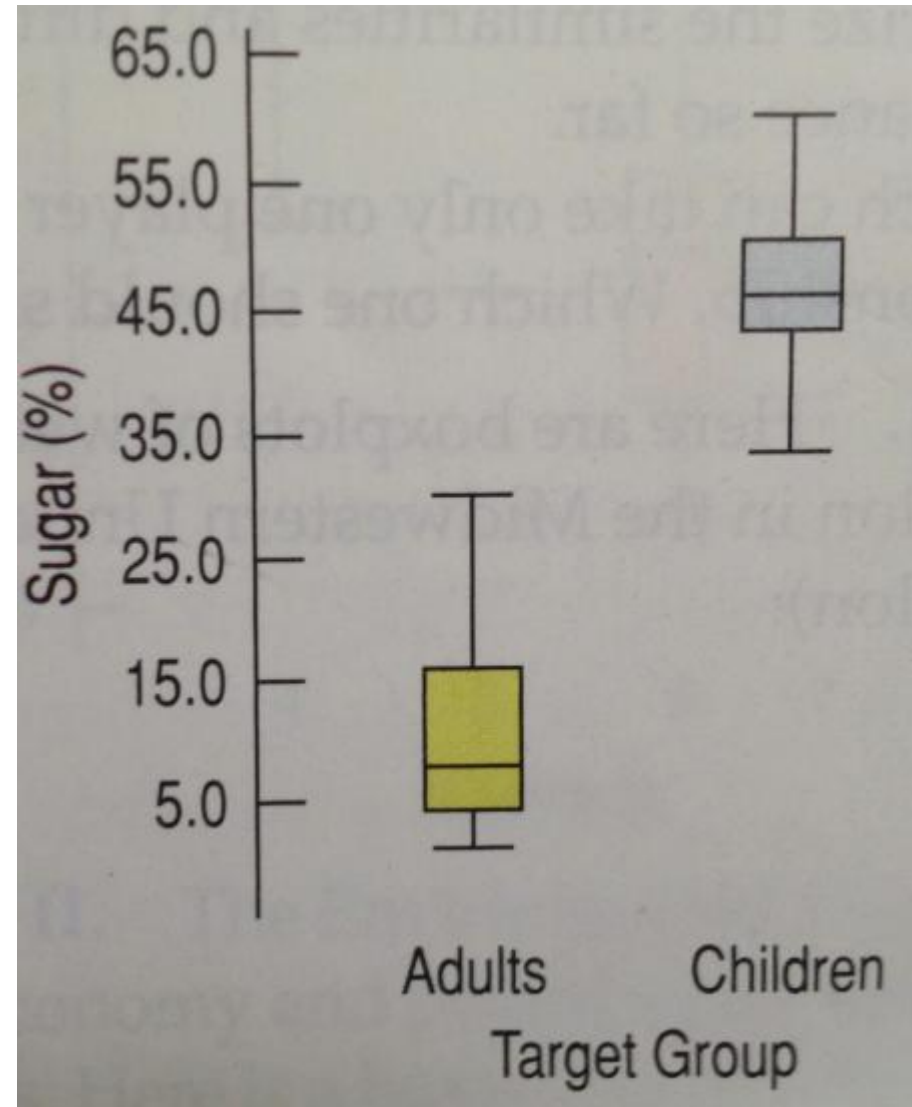
Chapter 5: #8

- ❑ Histograms of combined time for skiers
count = 35, mean = 196.079, sd = 5.80009
min = 189.35, Q1 = 192.238, Median = 193.270
Q3 = 200.625, Max = 211.890
- ❑ Range? (22.54 sec)
- ❑ Between what lengths do the central 50% lie? (Q1-Q3)
- ❑ What summary statistics would you use? (median and IQR)
- ❑ Brief description of data (shape, center, and spread)
Skewed to the right, unimodal (possibly 2). The median, IQR(central 50% of times), how about the winning time comparing to Q1? Etc..

Chapter 5: #13

Boxplot of Sugar content of cereals for children and adults

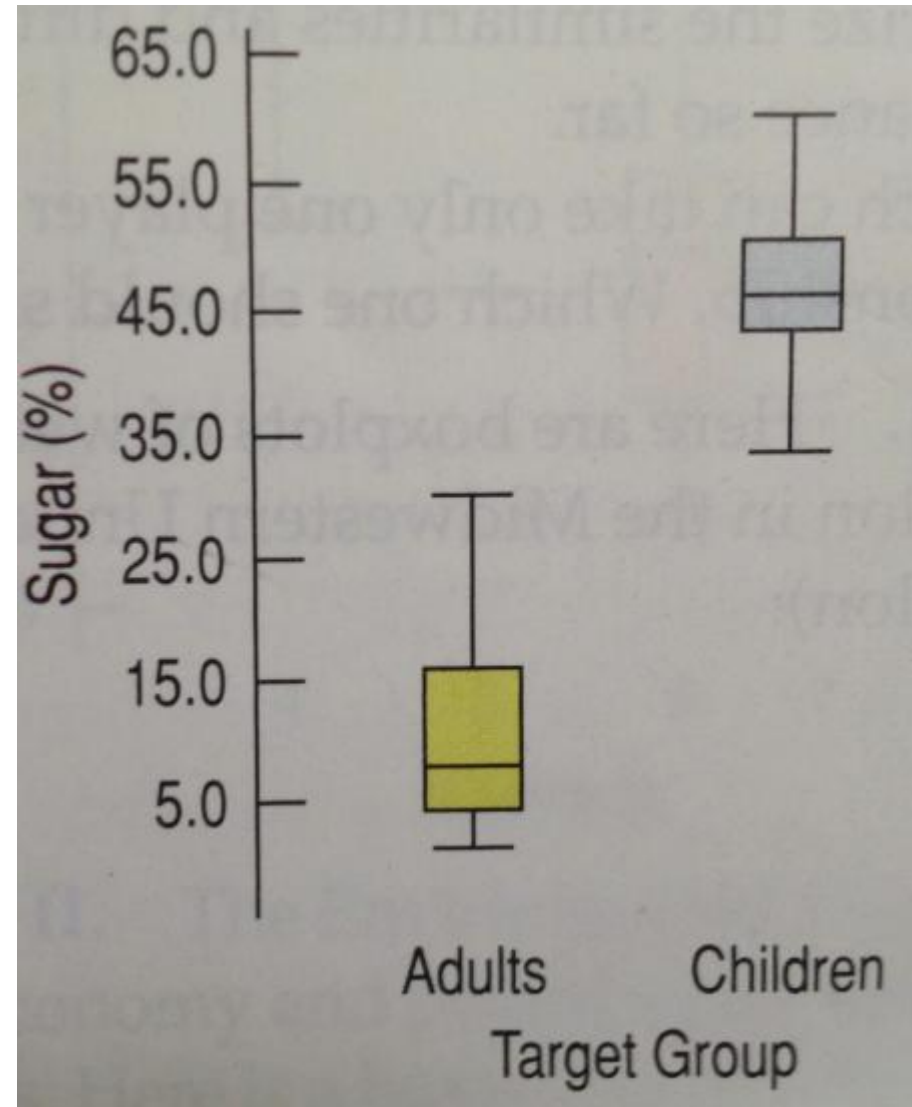
- All children's cereals higher than adult cereals?
- Which group varies more? Explain



Chapter 5: #13

Boxplot of Sugar content of cereals for children and adults

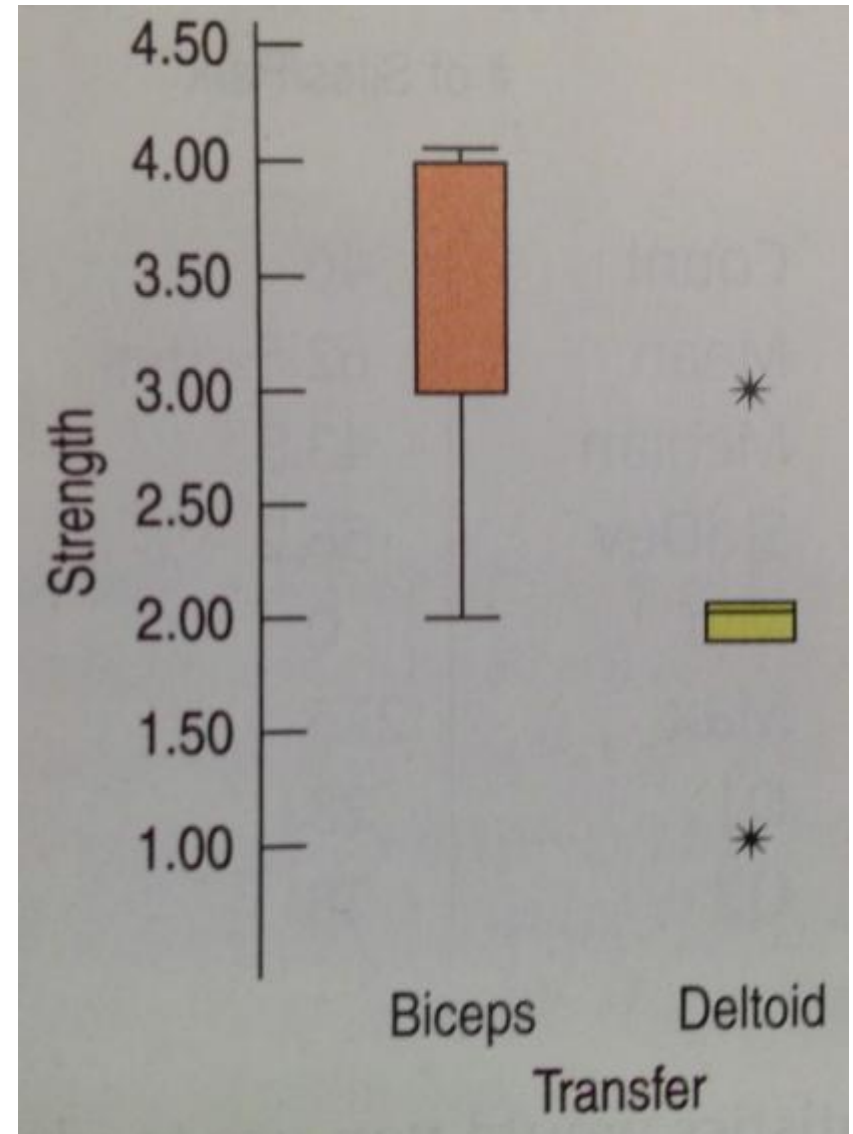
- All children's cereals higher than adult cereals? (yes)
- Which group varies more? Explain (adult, due to larger IQR -> more variability)

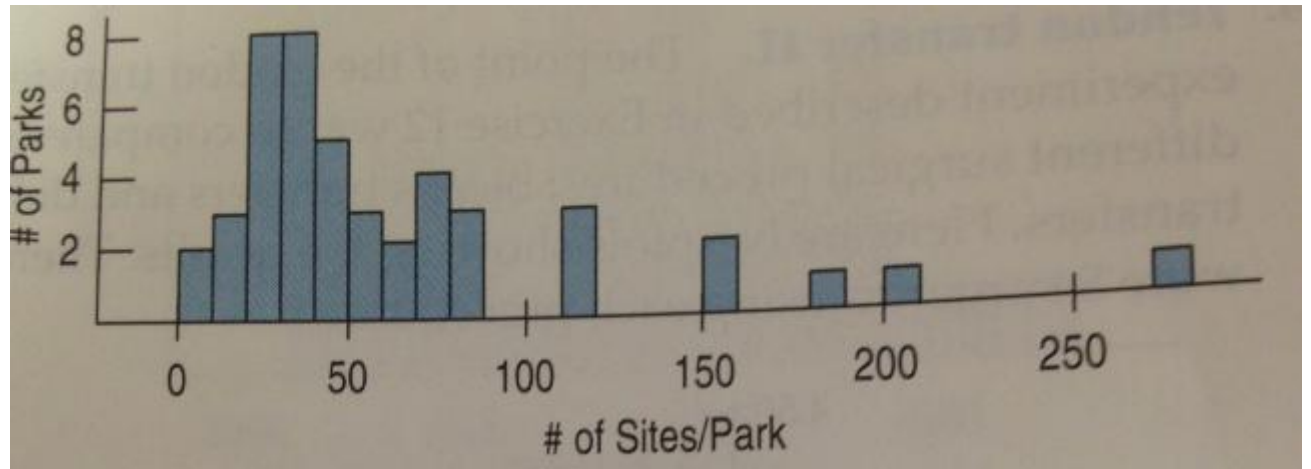


Chapter 5: #14

Boxplot of strength for biceps and deltoid

- Which method higher (better) median score?
- Was that method always best?
- Which method produces the most consistent results?
Explain. (hint: IQR smaller -> more consistent)



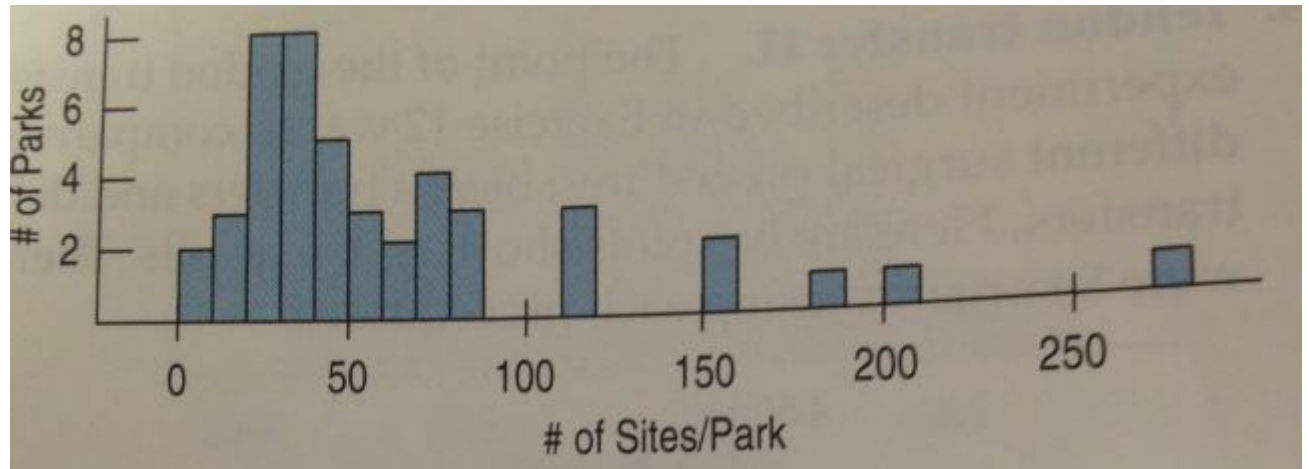


Chapter 5: #16

Histogram of # of parks versus number of sites/park

- Which statistics?
- How many peaks are outliers?
- Create a boxplot for these data
- Describe the distribution.

Count	46
Mean	62.8 sites
Median	43.5
StdDev	56.2
Min	0
Max	275
Q1	28
Q3	78



Chapter 5: #16

Histogram of # of parks versus number of sites/park

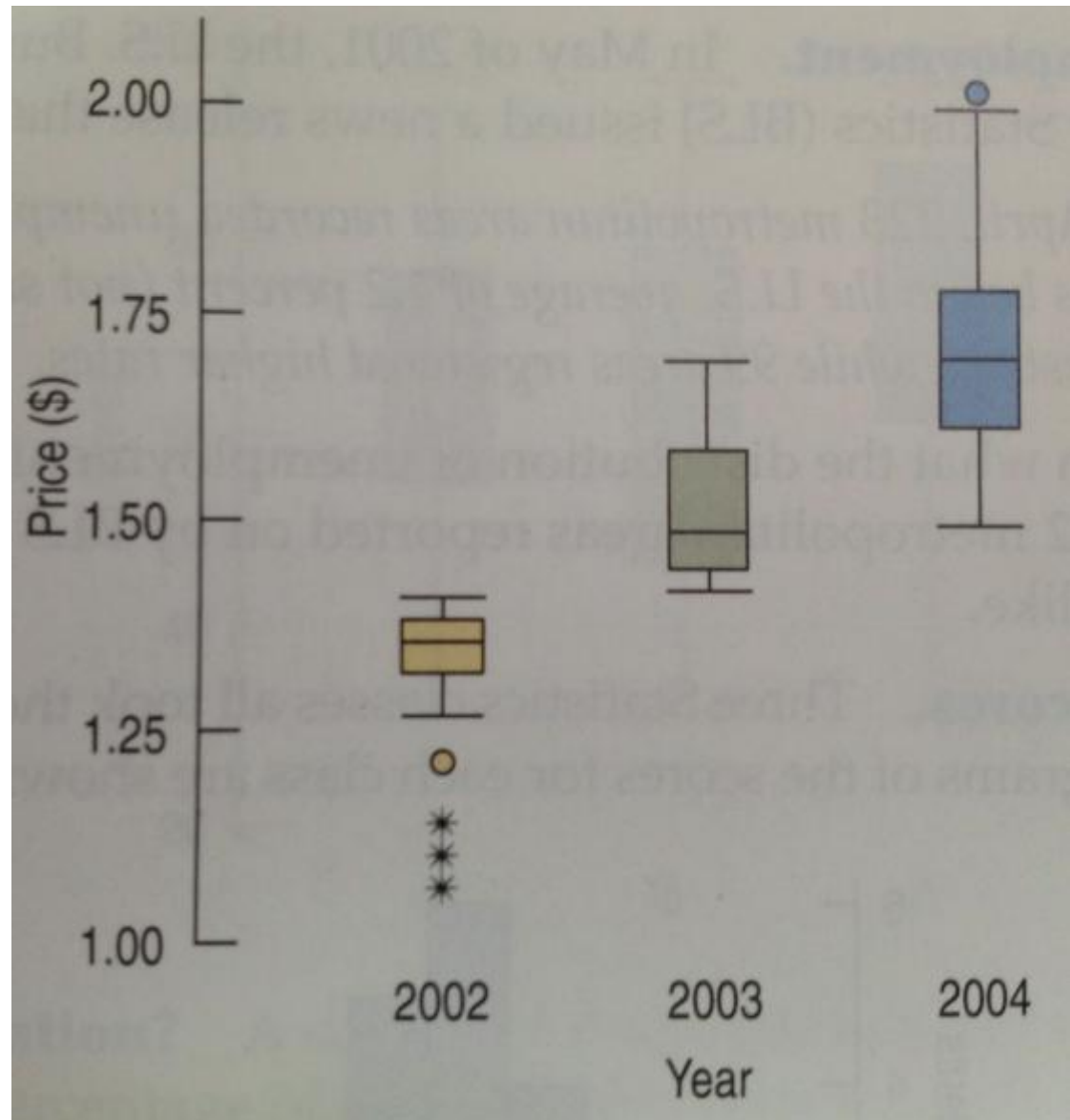
- Which statistics? (median and IQR due to the skewness to the right)
- How many peaks are outliers? (upper fence = $78 + 1.5 * 50 = 153$. possibly 4 peaks)
- Create a boxplot for these data (note **$78 + 3 * 50 = 228$**)
- Describe the distribution.

Count	46
Mean	62.8 sites
Median	43.5
StdDev	56.2
Min	0
Max	275
Q1	28
Q3	78

Chapter 5: #20

Boxplot of prices for year 2002 - 2004

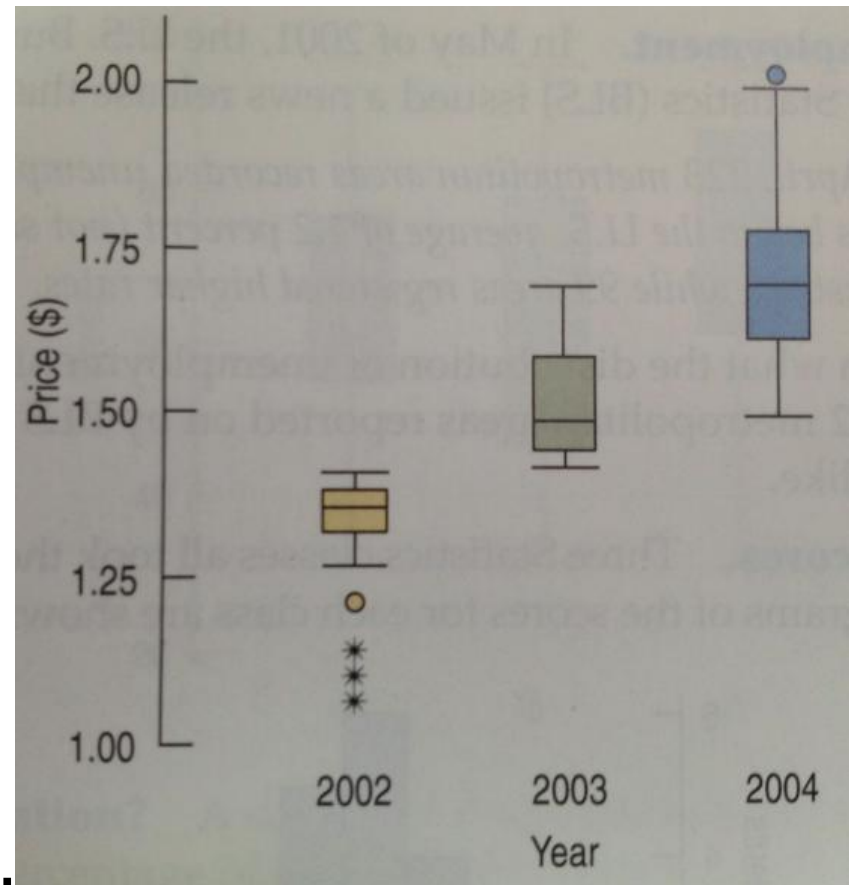
- Compare the three distributions.
- In which year were the prices least stable? Explain.



Chapter 5: #20

Boxplot of prices for year 2002 - 2004

- ❑ Compare the three distributions. (price has been increasing on average, and in spread. 2002: skewed to the left with outliers, then increasingly skewed to the right, with high outlier in 2004.)
- ❑ In which year were the prices least stable? Explain. (2004, larger range and IQR)



Chapter 5: #31

Summary statistics of percentage of graduation on time

- Symmetric or skewed? Explain.
- Any outliers? Explain.
- Create boxplot
- Write a few sentences about the graduation rates.

	% on time
Count	48
Mean	68.35
Median	69.90
StdDev	10.20
Min	43.20
Max	87.40
Range	44.20
25th %tile	59.15
75th %tile	74.75

Chapter 5: #31

Summary statistics of percentage of graduation on time

Symmetric or skewed? Explain.

(slightly skewed to the left:
mean < median, Q1 farther from
the median than Q3)

Any outliers? Explain.

Create boxplot

Write a few sentences about the graduation rates.
(mean, range, IQR)

Tips: read Page 90 – 91 in the textbook

	% on time
Count	48
Mean	68.35
Median	69.90
StdDev	10.20
Min	43.20
Max	87.40
Range	44.20
25th %tile	59.15
75th %tile	74.75

Suggestions for boxplot:

- 1. read Page 90 – 91 in the textbook carefully.*
- 2. Don't forget drawing the **axis** on the left as calibrations, with equal-spaced ticks and suitable range that covers the range of your data*
- 3. You will need to find the upper fence and lower fence during drawing the plot. But they are not necessary component of a boxplot. Note the endpoints of whiskers are **maximal (minimal) value falling in upper (lower) fences**, not the fences themselves!*
- 4. Understand outliers (circle, beyond $1.5IQR$) and **far outliers**(star, beyond $3IQR$).*
- 5. In the box, median in the middle indicates symmetry; median **closer to Q1** indicates possible **right-skewness**, and closer to Q3 indicates possible left-skewness. This should be judged along with outliers.*
- 6. It would be better if you add the labels for both x-y axis.*

Tips about distributions:

- 1. Measure of spread: range, SD, IQR. Larger SD, IQR: more spread, large variability, varies more, more dispersed, less stable, less consistent, less centered.*
- 2. Symmetric and unimodal: use mean and SD as measure of center and spread; skewed and unimodal: use median and IQR as measure of center and spread cause they are less sensitive, robust.*
- 3. Skewed to right: typically, mean > median. Skewed to left: typically, mean < median.*