Sec 1.1

Success Criteria

- Recognize categorical vs quantitative variables
- Calculate marginal distributions (%s) and conditional distributions (%s)
- Justify potential relationships between 2 categorical variables

Variables

- Characteristic of an "individual"

people, animals or things

- Something that varies

Types of Variables

1) Categorical / Qualitative
   - No Averages Possible
   - Eye Color, Gender, Zip Code ...

2) Quantitative Variables
   - Averages Possible
   - Height, Age, GPA, SAT Score ...

Distribution of A Variable

Tells what values a variable takes and how often it takes these values

- Pie Charts
- Bar graphs

} p. 9

- Dot Plots (p. 27)
- Stemplots (p. 34)
- Histograms (p. 37)

Relationships Between Categorical Variables

	Student Smokes	Student Not Smokes	
Both Parents Smoke	400	1380	1780
One Parent Smokes	416	1823	2239
No Parent Smokes	188	1168	1356
	1004	4371	5375

Relationships Between Categorical Variables

	Student Smokes	Student Not Smokes	
Both Parents Smoke	400	1380	1780
One Parent Smokes	416	1823	2239
No Parent Smokes	188	1168	1356
	1004	4371	5375

Marginal Distribution

$$\%\text{ of students smoke} = \frac{1004}{5375} = 19\%$$

Tells you nothing about any relationship

Relationships Between Categorical Variables

	Student Smokes	Student Not Smokes	
Both Parents Smoke	400	1380	1780
One Parent Smokes	416	1823	2239
No Parent Smokes	188	1168	1356
	1004	4371	5375

Conditional Distribution

a) % of students smoke when both parents smoke $= \frac{400}{1780} = 22\%$

b) % of students smoke when neither parent smokes $= \frac{188}{1356} = 14\%$

} compare

Conclusions

- When both parents smoke, student smoking goes from 14% to 22%
- 8% more students smoke when both parents smoke
- When both parents smoke, student smoking increases by 57% $\left|\frac{14-22}{14}\right|$

Simpson's Paradox

- "Lurking variables" which can change/reverse a relationship
- Most likely to occur when several groups are combined into single categorical group

Ex Where should I have surgery?

	Hospital A	Hospital B
Died After Surgery	63	16
Survived Surgey	2037	784
	2100	800

Hospital A loses $63/2100 = 3\%$ of patients

Hospital B loses $16/800 = 2\%$ of patients

So you choose Hospital B for your surgery

But previous data ignored the condition of the patient:

Good Condition

	A	B
Died	6	8
Survived	594	592
	600	600

Poor Condition

	A	B
Died	57	8
Survived	1443	192
	1500	200

Hospital A only loses $6/600 = 1\%$ of patients in good condition (like you) and Hospital B loses $8/600 = 1.3\%$ so... Hospital A is actually safer for you!

Organizing A Statistical Problem

"Statistic Problems Demand Consistency"

- **State** Question you're trying to answer
- **Plan** Statistical techniques needed
- **Do** Needed calculations / graphs
- ★ **Conclude** In context of problem

Sec 1.2

Success Criteria

- Construct dotplots, stemplots and histograms
- Describe distributions using SOCS

Graphs of Quantitative Variables

1) Dotplots (P. 27)
2) Stemplots (P. 35)
3) Histograms (P. 37)

Shape < Symmetric / Skewed
Center (Median)
Spread
Outliers?

↓

SOCS

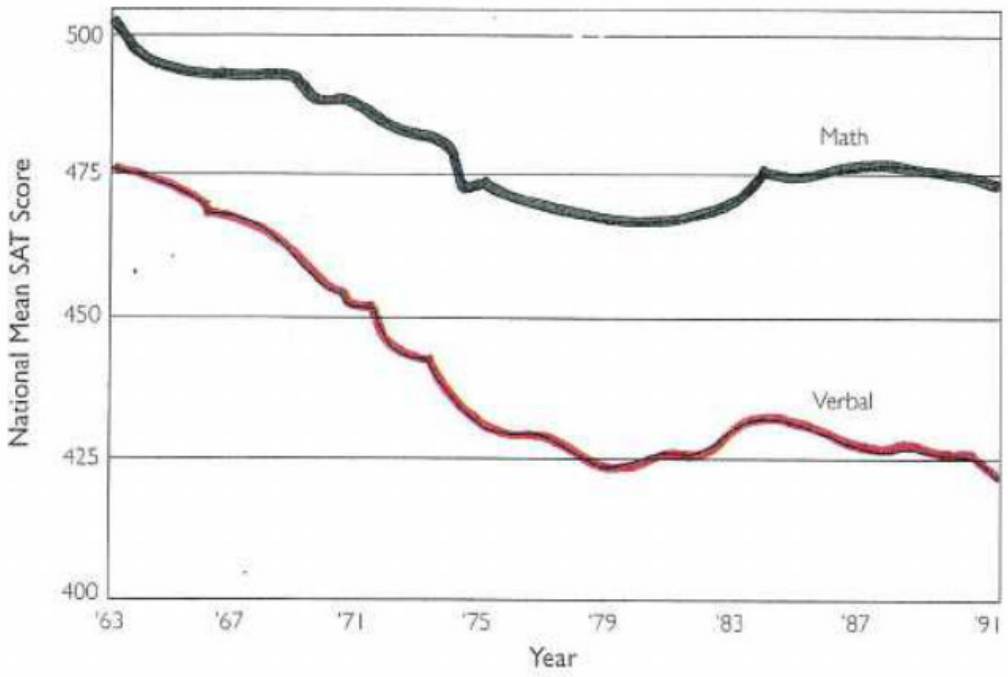
Making Graphs

1) Title

2) Label / scale axes for dotplots and histograms; include "key" for stemplots

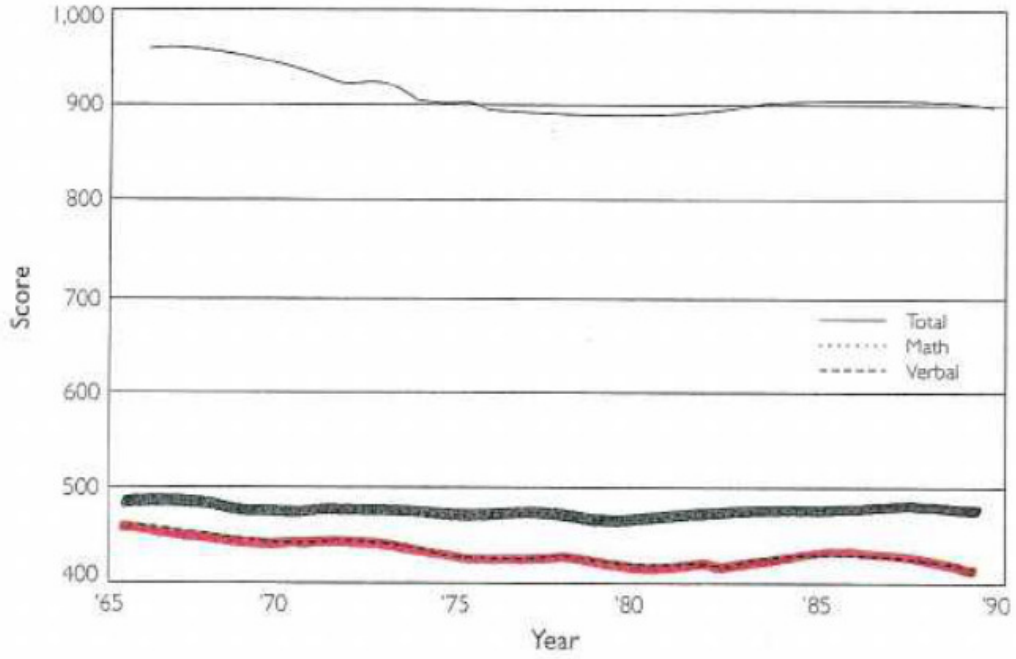
Caution

- Note scales used to distort data
- SAT Example...



Source: *The Public Interest*, No. 6 (Winter 1992): 33. © 1999 by National Affairs, Inc.

FIG. 25a Decline in SAT Scores, Graphed on a One-Hundred-Point Axis



Source: C. C. Carson, R. M. Huelskamp, and T. D. Woodall. 1992. "Perspectives on Education in America." *Journal of Educational Research* 86: 259–310.

FIG. 25b Decline in SAT Scores, Graphed on a Six-Hundred-Point Axis

Dotplot (P. 27)

Shape - Skewed Right

Center ≈ 4

Spread = 0 to 8

Outliers - 6, 7 or 8 maybe

Stemplot

16 Test Scores:

stem | 1 digit (leaf)

9|0, 8|0, 9|6, 5|4, 8|0, 9|5, 10|0, 7|5,

8|7, 6|2, 6|5, 8|5, 9|2, 8|7, 7|4, 8|9

5	4
6	2 5
7	5 4
8	0 0 7 5 7 9
9	0 6 5 2
10	0

→

5	4
6	2 5
7	4 5
8	0 0 5 7 7 9
9	0 2 5 6
10	0

86%

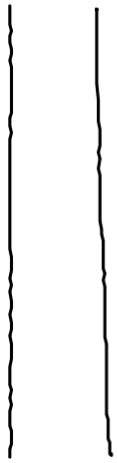
5|4 means 54%

Misc

1) Stems can be split (0-4, 5-9)

2) Use side by side stem plots to compare

Male Female



Histograms

- By Hand
- Using GDC

} IQ Scores (P. 39)

Histograms By Hand . . .

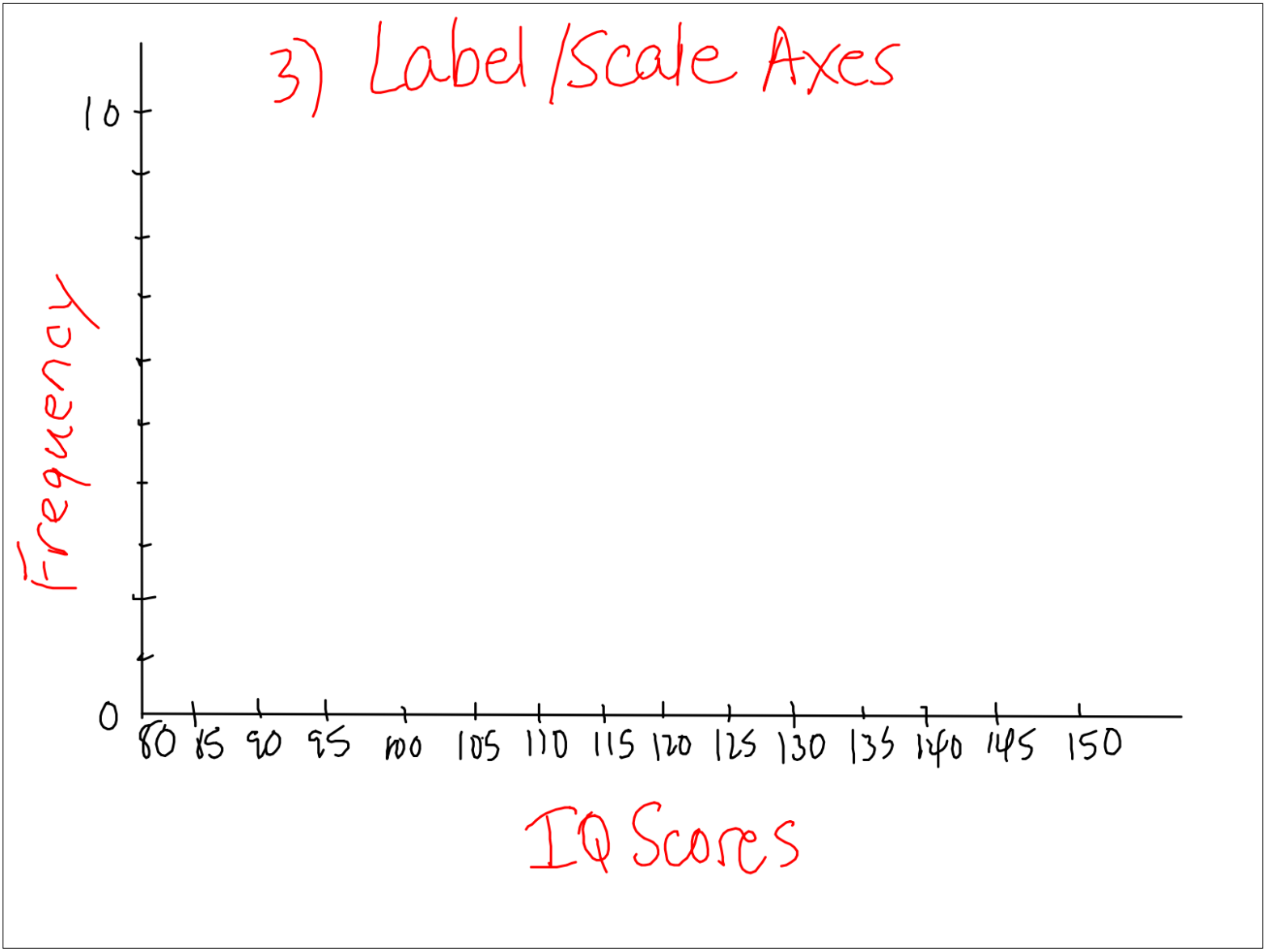
1) Make Frequency Distribution

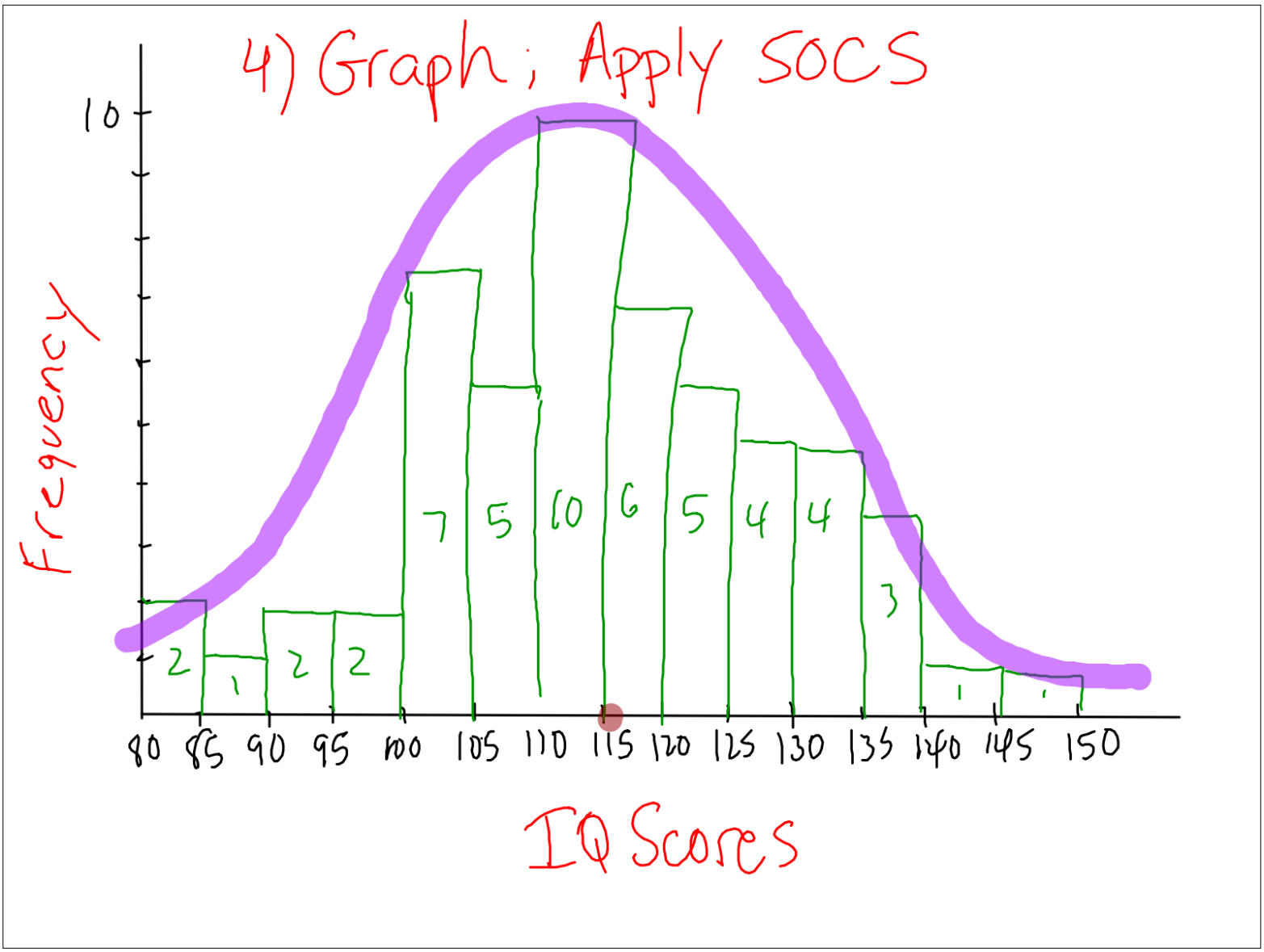
80	90 - 1	100 - 1	110 - III	120	130 - 1	140
81 - 1	91	101 - IIII	111	121	131 - 1	141
82 - 1	92	102 - III	112 - II	122 - II	132	142 - 1
83	93	103 - 1	113 - II	123 - 1	133 - 1	143
84	94 - 1	104	114 - III	124 - III	134 - II	144
85	95	105 - 1	115 - 1	125 - 1	135	145 - 1
86	96 - 1	106 - 1	116 - 1	126 - 1	136 - 1	146
87	97 - 1	107	117 - III	127 - II	137 - 1	147
88	98	108 - 1	118 - II	128 - 1	138	148
89 - 1	99	109 - III	119	129	139 - II	149

2) Choose Appropriate Classes (5?)

80	90 -1	100 -1	110 -111	120	130 -1	140
81 -1	91	101 -111	111	121	131 -1	141
82 -1	92	102 -111	112 -11	122 -11	132	142 -1
83	93	103 -1	113 -11	123 -1	133 -1	143
84	94 -1	104	114 -111	124 -111	134 -11	144
85	95	105 -1	115 -1	125 -1	135	145 -1
86	96 -1	106 -1	116 -1	126 -1	136 -1	146
87	97 -1	107	117 -111	127 -11	137 -1	147
88	98	108 -1	118 -11	128 -1	138	148
89 -1	99	109 -111	119	129	139 -11	149

3) Label/Scale Axes




4) Graph; Apply SOCS
10
Frequency
2
1
2
2
7
5
10
6
5
4
4
3
1
1
80 85 90 95 100 105 110 115 120 125 130 135 140 145 150
IQ Scores

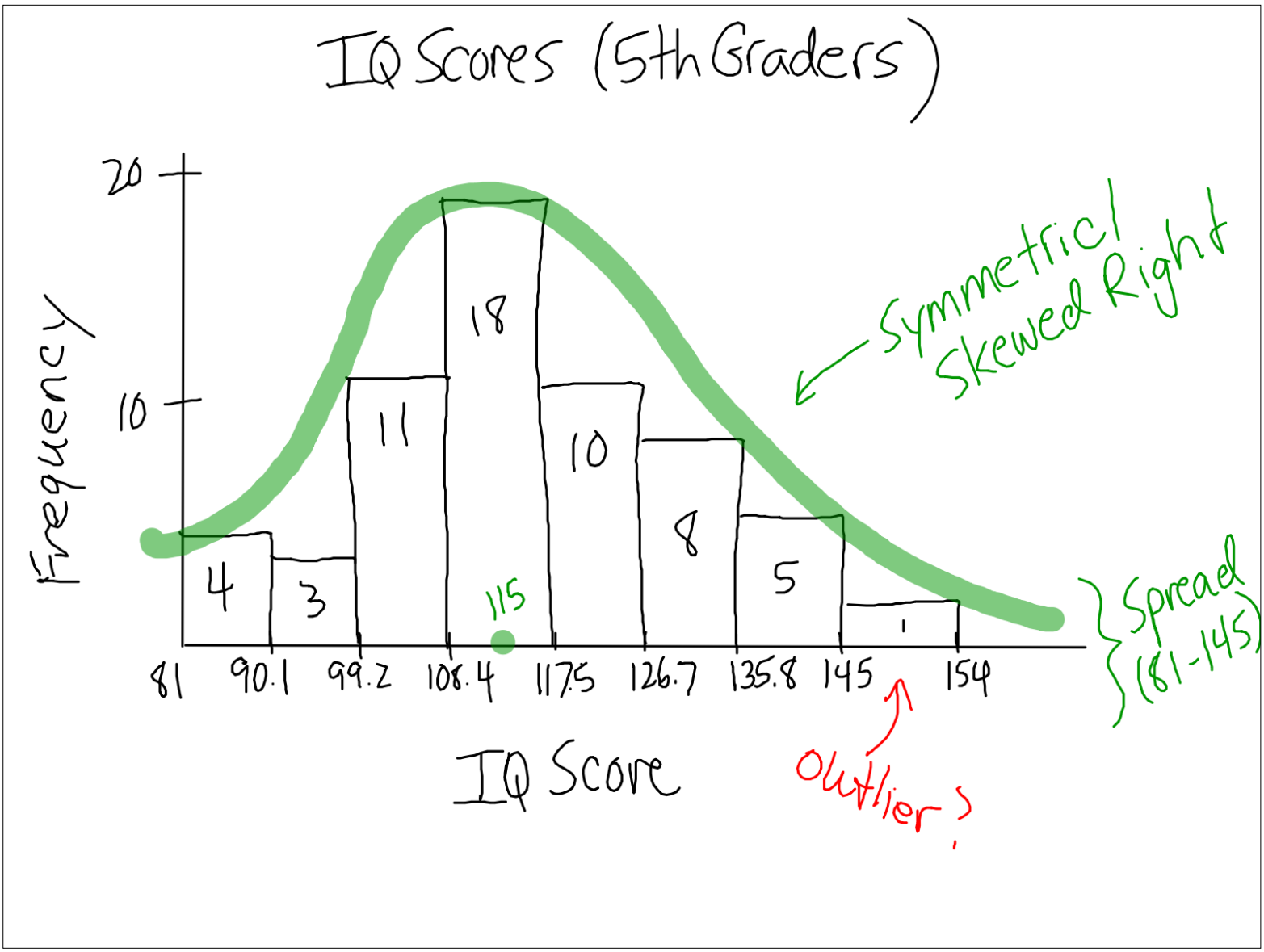
Using GDC

STAT → EDIT → ENTER → Data in L_1

STATPLOT → ENTER →  → ZOOM → 9

2nd → Y=

- sketch... Use TRACE for values


IQ Scores (5th Graders)
20
10
Frequency
4
3
11
18
10
8
5
1
115
81
90.1
99.2
108.4
117.5
126.7
135.8
145
154
IQ Score
Symmetric/
Skewed Right
Spread
(81-145)
Outlier?

Misc

1) Calculator histograms use actual frequencies
2) Histograms using relative frequencies (%) need to be done by hand
3) Histograms are not bar graphs
4) When comparing data where number of observations are different, use percents (p. 40)

Sec 1.3

Success Criteria

- Calculate center, spread and 5-number summaries of quantitative data
- Determine which measure of center and spread to use in a distribution
- Construct boxplots
- Apply IQR rule to determine outliers

Measure of Central Tendency (Center)

1) Mean ($\bar{x}$ for samples, μ for populations)

↑ "x bar" ↑ Mu

- Arithmetic average = $\frac{\sum_i x_i}{n}$
- Not resistant (affected by outliers)
- Best for symmetric distributions

2) Median (M)

- Midpoint of values

  Ex 2, 4, 6 ($M = 4$)

  2, 4, 6, 8 ($M = 5$)

- Very resistant

  Ex 10, 20, 30 ($\bar{X} = 20$, $M = 20$)

  10, 20, 100 ($\bar{X} = 43.3$, $M = 20$)

- Best for skewed distributions

Spread

Range = max − min

Variance (s^2)

$$s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i-\bar{x})^2$$

✡ Standard Deviation (s) = $\sqrt{\text{Variance}}$

- Not resistant
- Best for symmetric distributions

Ex Calculate standard deviation:

$$\{3, 4, 8\} \rightarrow \bar{x} = 5$$

$$s = \sqrt{\frac{(3-5)^2 + (4-5)^2 + (8-5)^2}{3-1}}$$

$$= \sqrt{7}$$

$$= 2.6457$$

Interquartile Range (IQR)

- Middle 50% = $Q_3 - Q_1$
- Best for skewed distributions

5-Number Summary + 2

Ex Ages of people at family reunion:

16 19 24 25 25 33 33 34 34 37 37 40 42 46 49 73

min = 16

$Q_1 = 25$

M = 34

$Q_3 = 41$

Max = 73

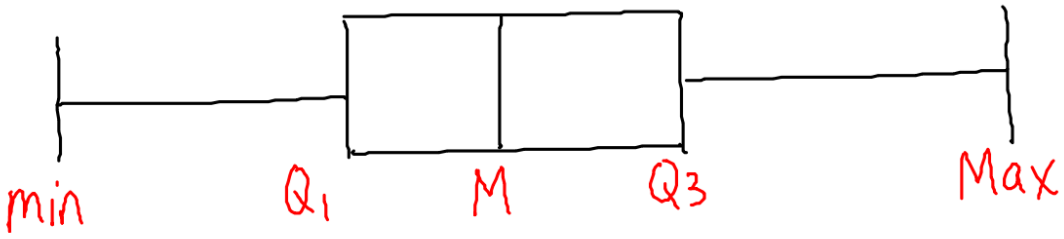
IQR = 16 yrs

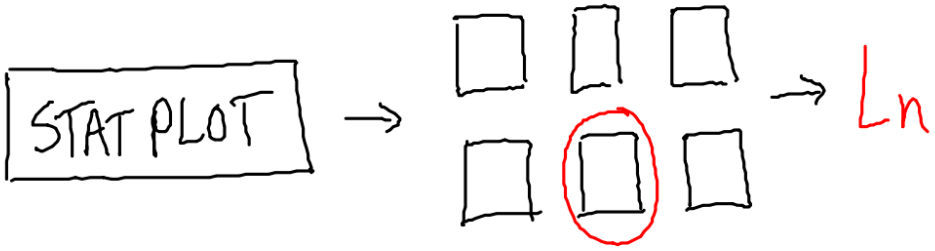
STAT → CALC → ENTER → 1-Var Stats Ln

$\bar{x} = 35.44$ yrs $s = 13.63$ yrs

Boxplots

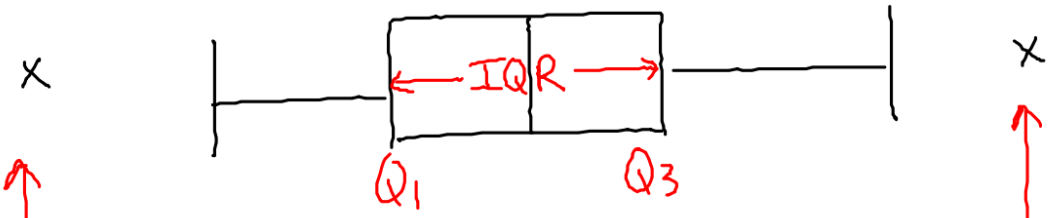
Used to display 5-number summary





Modified Boxplot

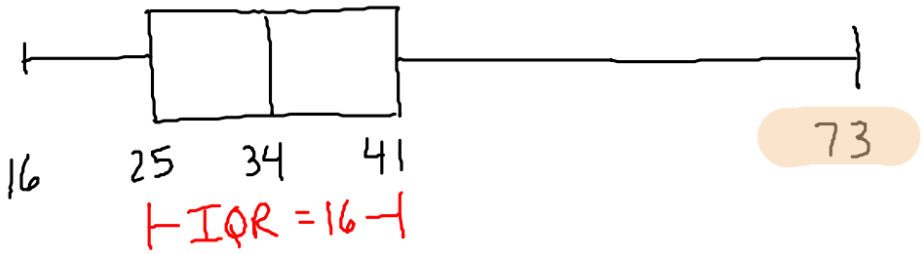
- Displays possible outliers



Outlier $< Q_1 - 1.5(IQR)$

Outlier $> Q_3 + 1.5(IQR)$

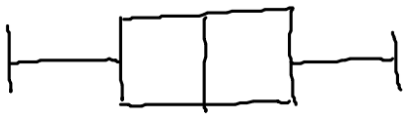
Ex Ages at family reunion



Outlier $< 25 - 1.5(16) < 25 - 24 < 1$

Outlier $> 41 + 1.5(16) > 41 + 24 > 65$

Modified Box Plot

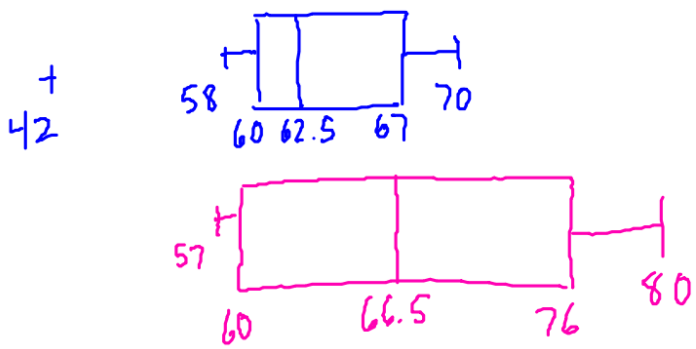


x

Side By Side Boxplots

L_1 = Male Pulse Rates

L_2 = Female Pulse Rates



Frequency Distributions

Age	Frequency
14	10
15	25
16	40
17	15
18	10

} Find mean & standard deviation

1) Let L_1 = all 100 values → CALC → 1 VAR STATS L_1

2) Let L_1 = values, L_2 = frequency → CALC → 1 VARSTATS L_1, L_2

Test Taking Strategies

- Review (online) Notes / Homework
- Read Section Summaries (Pp 21, 42, 69)
- Practice Solving Problems
  - Practice Test
  - Chapter Review Exercises (Pp 75-78)

All answers in back of book :)

* Work With A Study Buddy / Group
- Attend Review Session Before School