To Do List

Read Chapter 1
Do Problems 1-20

Assignment 1 is due Friday
September 23
Tutorial Test 1 is on Wednesday
September 28

Last Class

We discussed the following statistical jargon:

Empirical study

Population, process

Variates and types of variates

Attributes

Types of empirical studies

Types of Empirical Studies

Sample Surveys

Observational Studies

Experimental Studies

We will see much later in the course that cause and effect conclusions can only be made if an experimental study has been conducted.

Today and Friday Lectures:

Ways of Summarizing Data:

1) Numerical Summaries (today)

2) Graphical Summaries (Friday)

Note: Numerical and graphical summaries will be useful in identifying a suitable probability model for the data.

Numerical Measures for Summarizing Univariate Data

Types of numerical measures:

1) **Measures of location (sample mean, median, and mode)**

2) **Measures of variability or dispersion (sample variance, sample standard deviation, range, and interquartile range (IQR))**

3) **Measures of shape (sample skewness and sample kurtosis)**

Measures of Central Tendency or Location

Let the data be represented as $\{y_1, y_2, \ldots, y_n\}$ where y_i is a real number.

Numerical measures of the "center" of the data:

1) Sample mean or or average

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

2) Sample median: $\hat{m}$

3) Mode

Order Statistics and the Median

We denote the ordered sample (also called the order statistic) as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$

where $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

and $y_{(1)}$ = minimum($y_1,\ldots,y_n$), $y_{(n)}$ = maximum($y_1,\ldots,y_n$).

For an odd number of observations:

$$\text{sample median} = \hat{m} = y_{\left(\frac{n+1}{2}\right)}$$

e.g. median{5,1,2} = median{1,2,5} = 2

Order Statistics and the Median

The median is not unique in the case of an even number of observations.

The average of the middle two observations is chosen for convenience.

$$\text{sample median} = \hat{m} = \frac{1}{2}\left(y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)} \right)$$

Sample Mode

The sample mode is the *most common value* in the set of data. If the values are all unique then the mode does not exist.

The sample mode is most useful for discrete or categorical data with a relatively small number of possible values.

For frequency or grouped data the group or class with the highest frequency is called the sample modal class.

STAT 231 W13 Marks (n=37)

64	71	73	73	75	76	76	77	77	79
80	80	81	81	81	81	82	83	83	85
86	87	87	87	87	88	89	89	90	90
91	91	92	92	92	94	96			

$$\bar{y} = \frac{3086}{37} = 83.4, \quad \hat{m} = y_{(19)} = 83$$

The mode is not unique since 81 and 87 both occur 4 times each.

STAT 231 W13 Marks – Frequency Table

Mark Interval	Frequency
[65,70)	1
[70,75)	3
[75,80)	6
[80,85)	9
[85,90)	9
[90,95)	8
[95,100]	1
Total	37

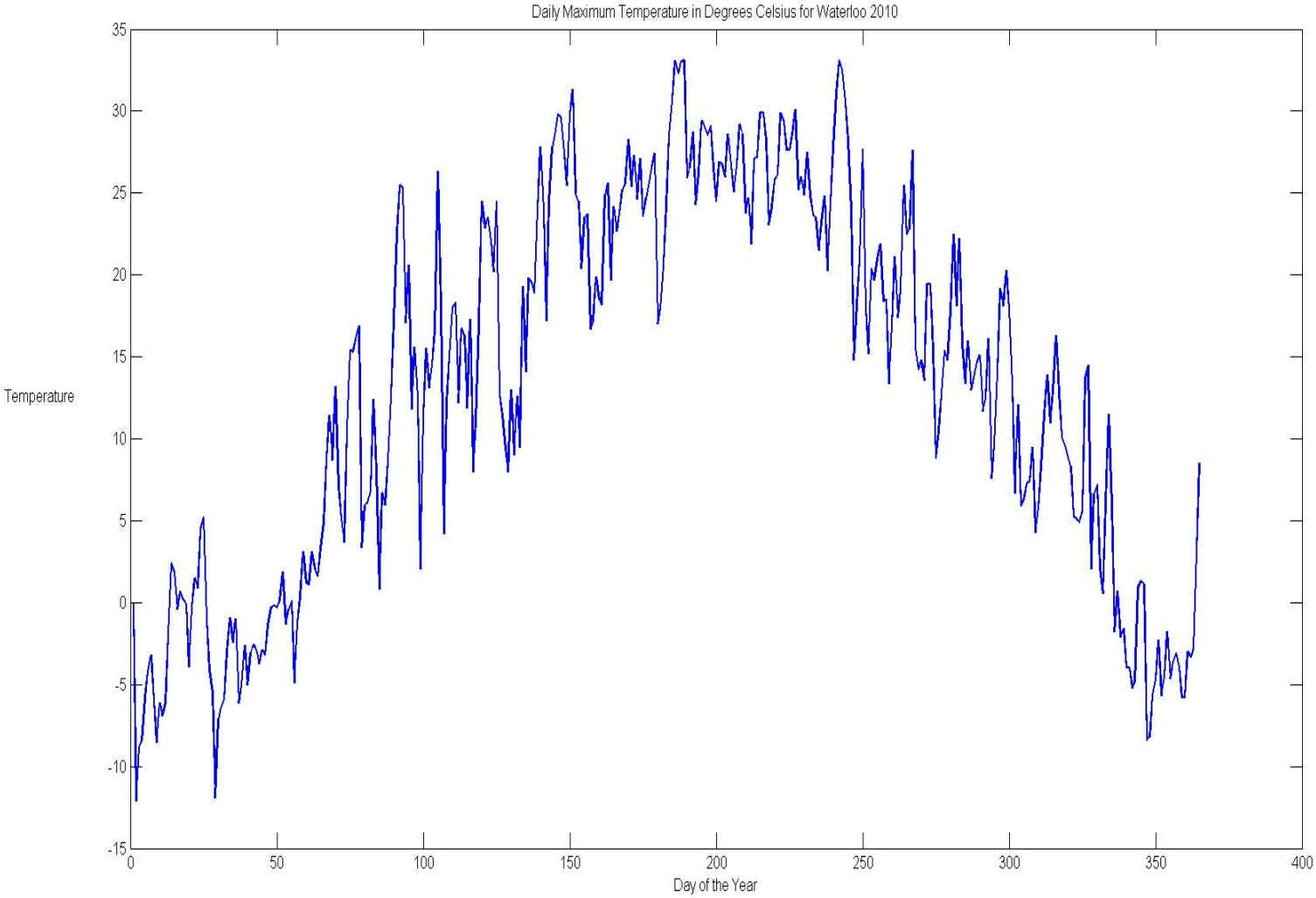
[80,85) and [85,90) are both modal classes. Note that the mean and median cannot be determined from these grouped data but we are able to determine in which interval the median must lie.

Weather Data from Waterloo 2010

http://weather.uwaterloo.ca/data.html

Date	Low temperature	High temperature	Precipitation
1-Jan-10	-11.9	-0.1	1.7
2-Jan-10	-15.4	-12.1	0.3
3-Jan-10	-14.7	-8.8	0.5
4-Jan-10	-11.5	-8.4	0
5-Jan-10	-9.8	-5.8	0.6
6-Jan-10	-7.9	-4.2	0.3
7-Jan-10	-6.9	-3.2	0.3
8-Jan-10	-10.6	-6	2.7
9-Jan-10	-20.7	-8.5	0.1
10-Jan-10	-21.5	-6.1	0.2
11-Jan-10	-10.2	-6.9	1.6
⋮	⋮	⋮	⋮
27-Dec-10	-12.2	-3	0
28-Dec-10	-6.7	-3.3	0
29-Dec-10	-4.8	-2.8	0
30-Dec-10	-6.5	3.1	0.2
31-Dec-10	3	8.5	1.6

Daily High Temperatures in Degrees Celsius for Waterloo 2010



Daily low temperatures in degrees Celsius in 2010:

$\bar{y} = 3.32, \quad \hat{m} = 2.7\,(\text{Oct. } 4), \quad \text{mode} = -3.1\,(\text{occurred } 5 \text{ times})$

Daily high temperatures in degrees Celsius in 2010:

$\bar{y} = 13.12, \quad \hat{m} = 14.2\,(\text{Sept. } 26), \quad \text{mode} = 24.5\,(\text{occurred } 5 \text{ times})$

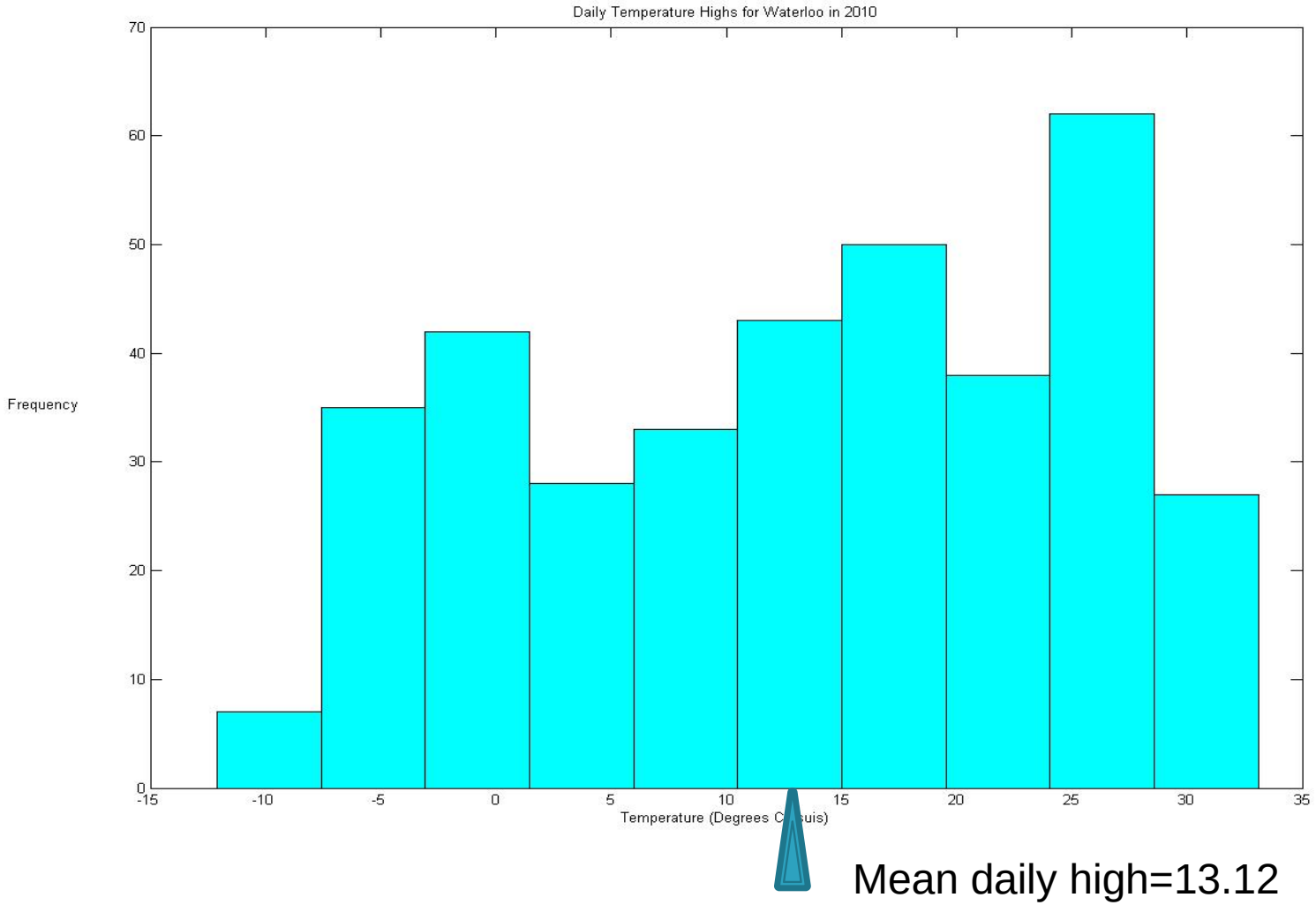
Daily precipitation in centimeters in 2010:

$\bar{y} = 2.41, \quad \hat{m} = 0.1\,(\text{occurred } 30 \text{ times}),$

$\text{mode} = 0\,(\text{occurred } 169 \text{ times})$

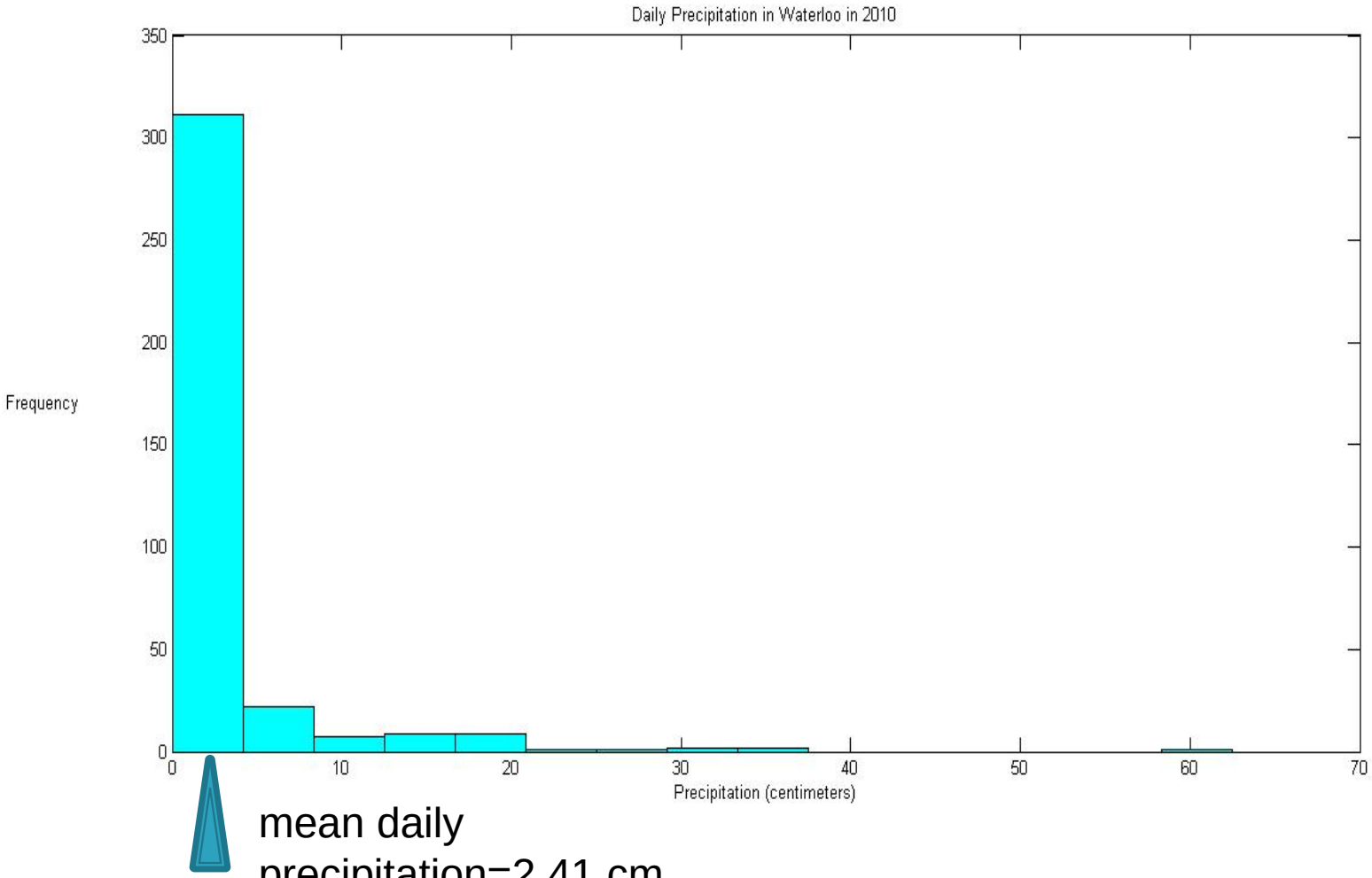
How good are these numerical summaries as measures of the "center" of these data?

Daily High Temperatures in Degrees Celsius for Waterloo 2010



Mean daily high=13.12

Daily Precipitation in Centimeters for Waterloo 2010



mean daily precipitation=2.41 cm

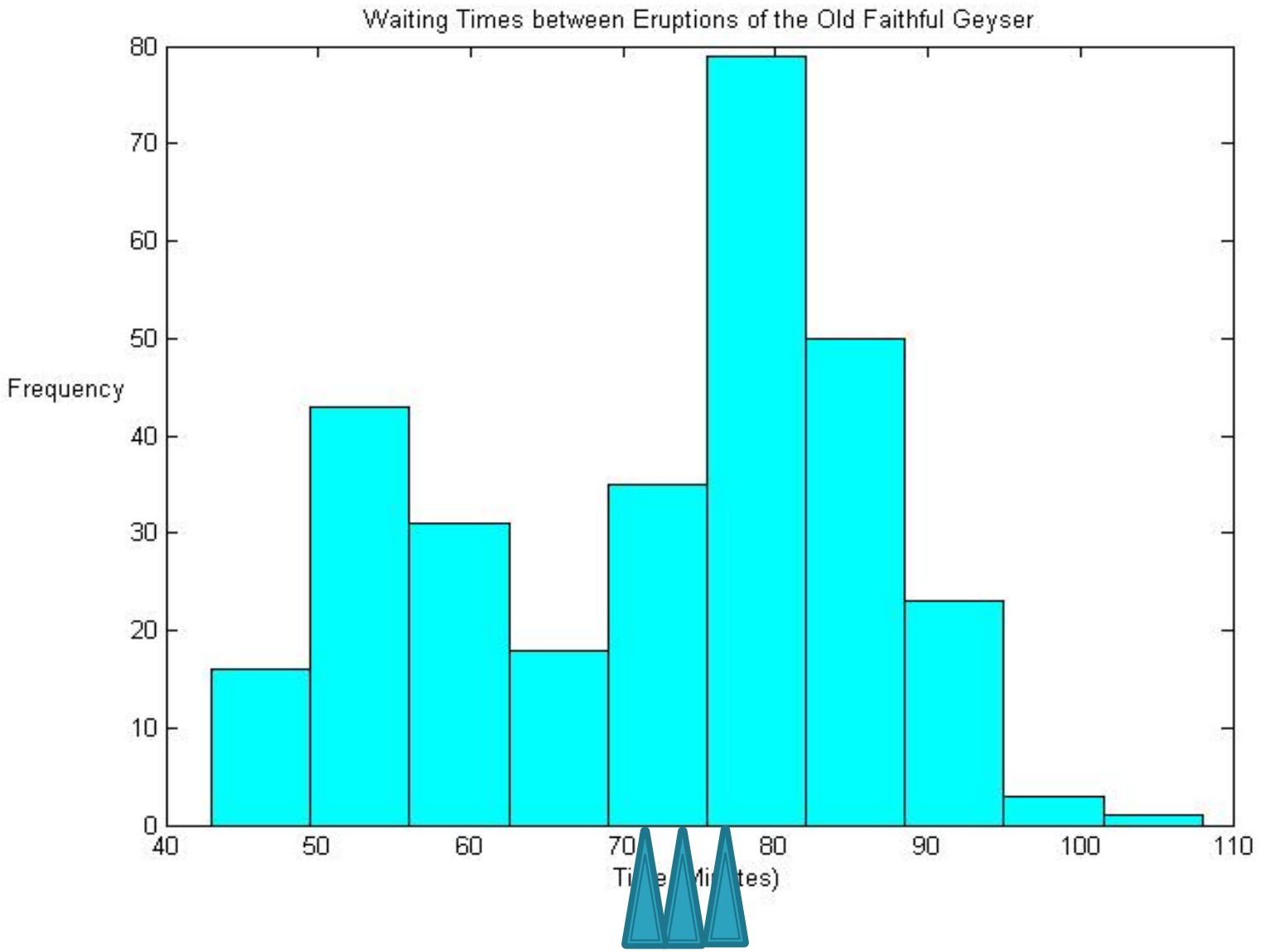
Example: Waiting times (in minutes) between 300 eruptions of the Old Faithful geyser between August 1 to 8, 1985

For the Old Faithful Geyser data the value 78 appeared the most (16 times) so the mode is 78.

$$\text{sample mean} = \bar{y} = 72.31 \text{ minutes}$$

$$\text{sample median} = \hat{m} = 76 \text{ minutes}$$

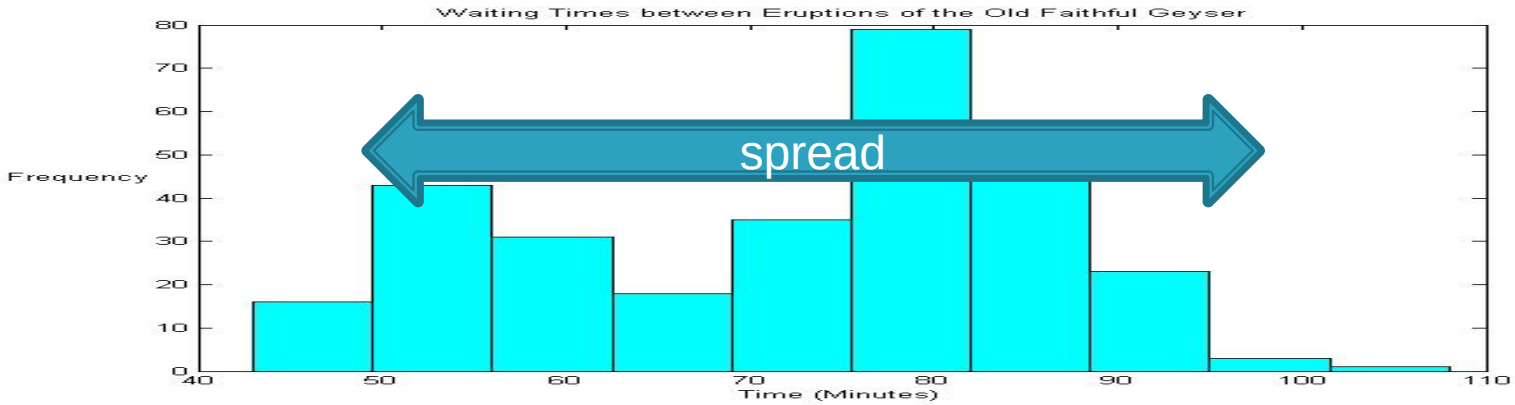
$$\text{sample mode} = 78 \text{ minutes}$$



mean median mode

Measures of variability or dispersion

Numerical measures which describe the variability or spread of the data:



a) Sample Variance and Sample Standard Deviation

b) Range

c) Interquartile Range (IQR)

Sample Variance and Sample Standard Deviation

Let the data be represented as $\{y_1, y_2, \ldots, y_n\}$ where y_i is a real number.

The sample variance is defined as

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 \right]$$

The sample standard deviation is s, the square root of s^2.

What are the units of measurement for sample variance and sample standard deviation?

If the data are unimodal and roughly symmetric then approximately 68% of the data will lie within one standard deviation of the mean, that is, approximately 68% of the data will lie in the interval

$$\left(\bar{y} - s,\ \bar{y} + s\right)$$

Approximately 95% of the data will lie within two standard deviation of the mean, that is, approximately 95% of the data will lie in the interval

$$\left(\bar{y} - 2s,\ \bar{y} + 2s\right)$$

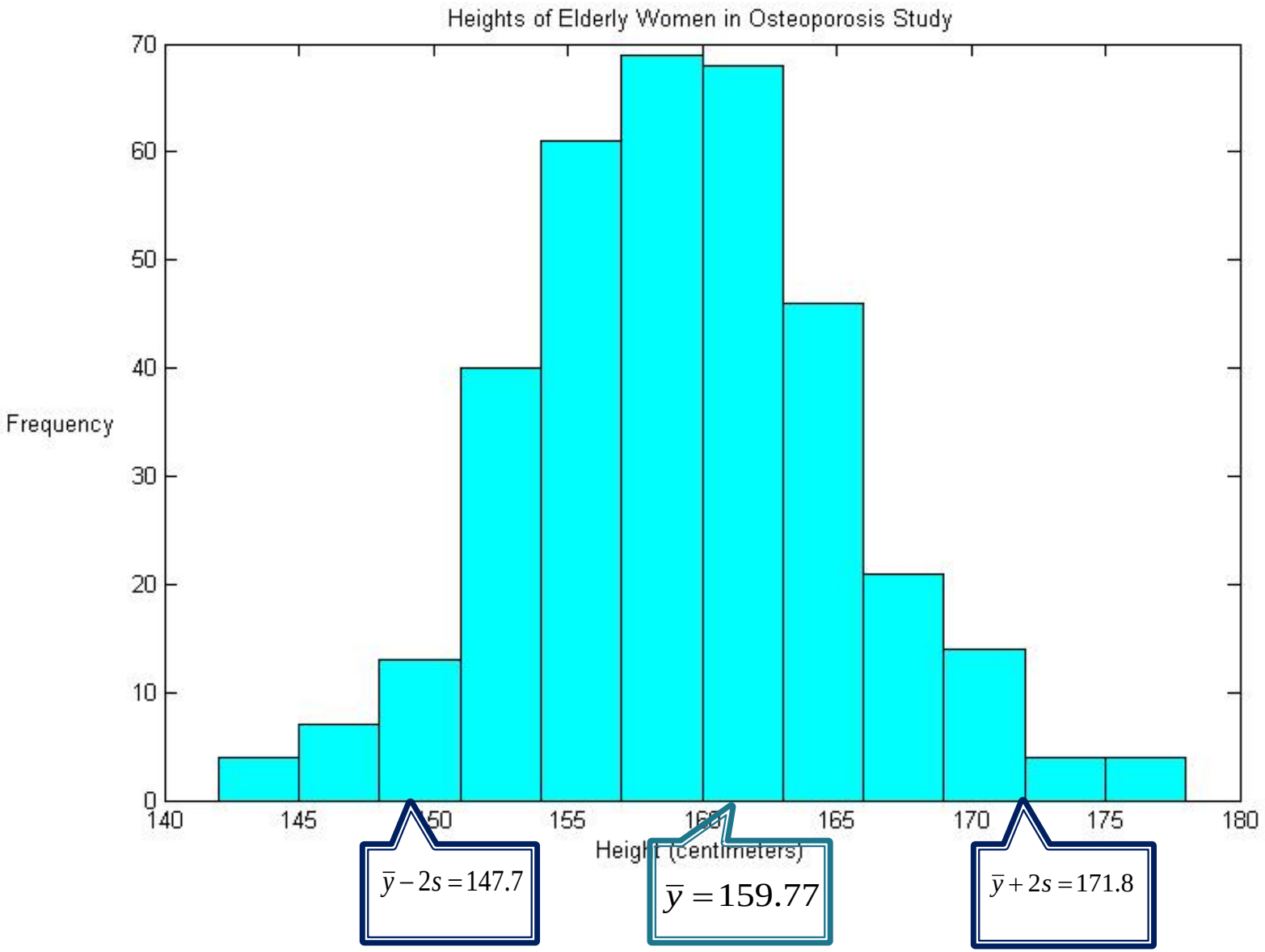
Question: What is the justification for these statements?

Hint: What do you know about the Normal distribution?

Example: Heights in centimeters of a sample of 351 elderly women randomly selected from a community in a study of osteoporosis.

156	163	169	161	154	156	163	164	156	166	177	158
150	164	159	157	166	163	153	161	170	159	170	157
156	156	153	178	161	164	158	158	162	160	150	162
155	161	158	163	158	162	163	152	173	159	154	155
164	163	164	157	152	154	173	154	162	163	163	165
160	162	155	160	151	163	160	165	166	178	153	160
156	151	165	169	157	152	164	166	160	165	163	158
153	162	163	162	164	155	155	161	162	156	169	159
159	159	158	160	165	152	157	149	169	154	146	156
157	163	166	165	155	151	157	156	160	170	158	165
167	162	153	156	163	157	147	163	161	161	153	155
166	159	157	152	159	166	160	157	153	159	156	152
151	171	162	158	152	157	162	168	155	155	155	161
157	158	153	155	161	160	160	170	163	153	159	169
155	161	156	153	156	158	164	160	157	158	157	156
160	161	167	162	158	163	147	153	155	159	156	161
158	164	163	155	155	158	165	176	158	155	150	154
164	145	153	169	160	159	159	163	148	171	158	158
157	158	168	161	165	167	158	158	161	160	163	163
169	163	164	150	154	165	158	161	156			
154	158	162	164	158	165	158	156	162			
157	167	142	166	163	163	151	163	153			
169	154	155	167	164	170	174	155	157			
155	168	152	165	158	162	173	154	167			
158	167	164	170	164	166	170	160	148			
150	165	165	147	162	165	158	145	150			
163	166	162	163	160	162	153	168	163			
158	155	168	160	153	163	161	145	161			
161	155	158	161	163	157	156	152	156			
160	152	153									

$$\bar{y} = 159.77\,\text{cm}, \quad s^2 = 36.36\,(\text{cm})^2, \quad s = 6.0\,\text{cm}$$



Old Faithful Data

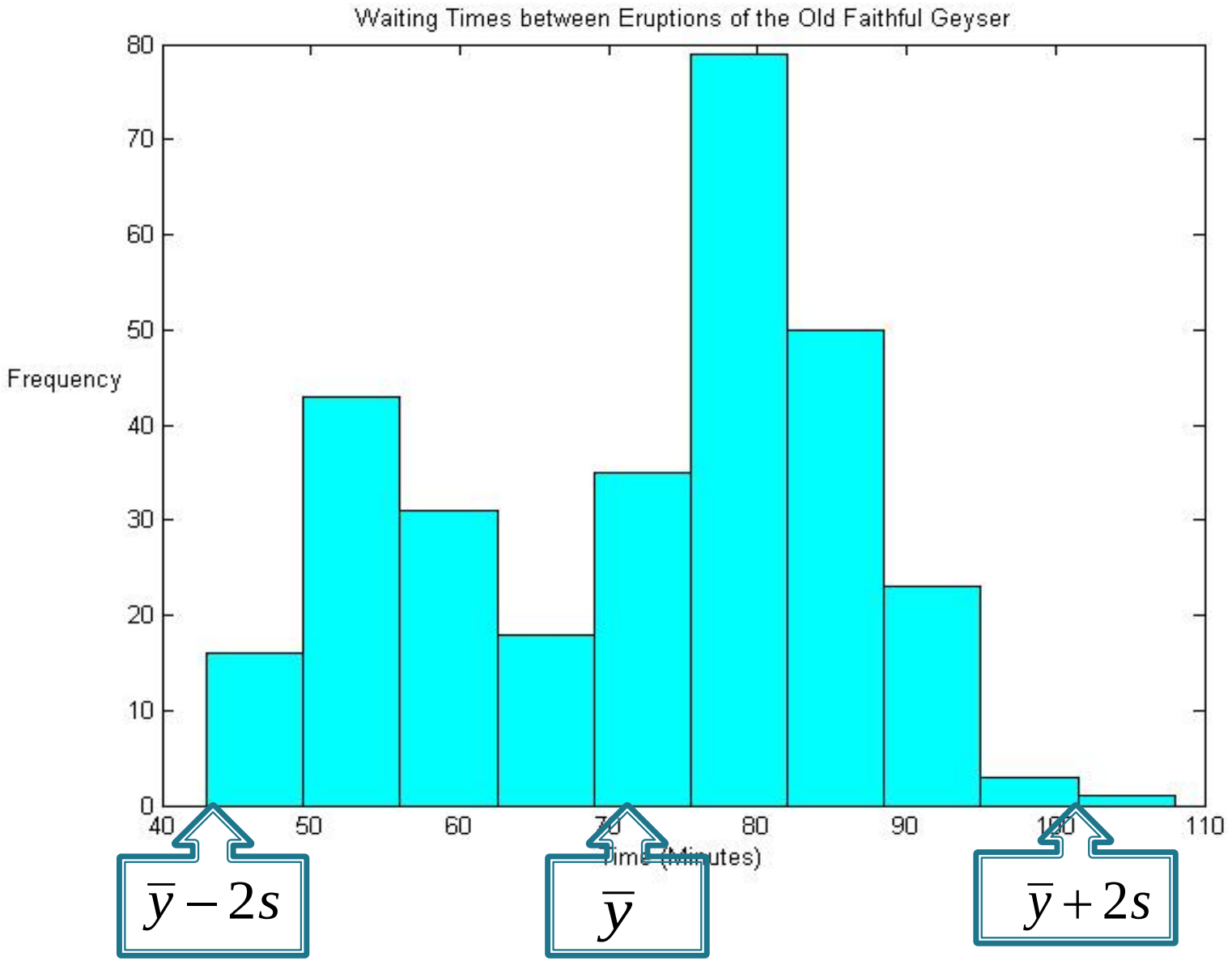
$$\bar{y} = 72.31\,\text{min}, \quad s^2 = 192.94\,(\text{min})^2, \quad s = 13.84\,\text{min}$$

$$(\bar{y} - s,\ \bar{y} + s) = (86.20, 58.42)$$

$$(\bar{y} - 2s,\ \bar{y} + 2s) = (44.53, 100.09)$$

The actual number of observations in the interval (86.20, 58.42) was 178 or 59.5%.

The actual number of observations in the interval (44.53, 100.09) was 297 or 99.23%.

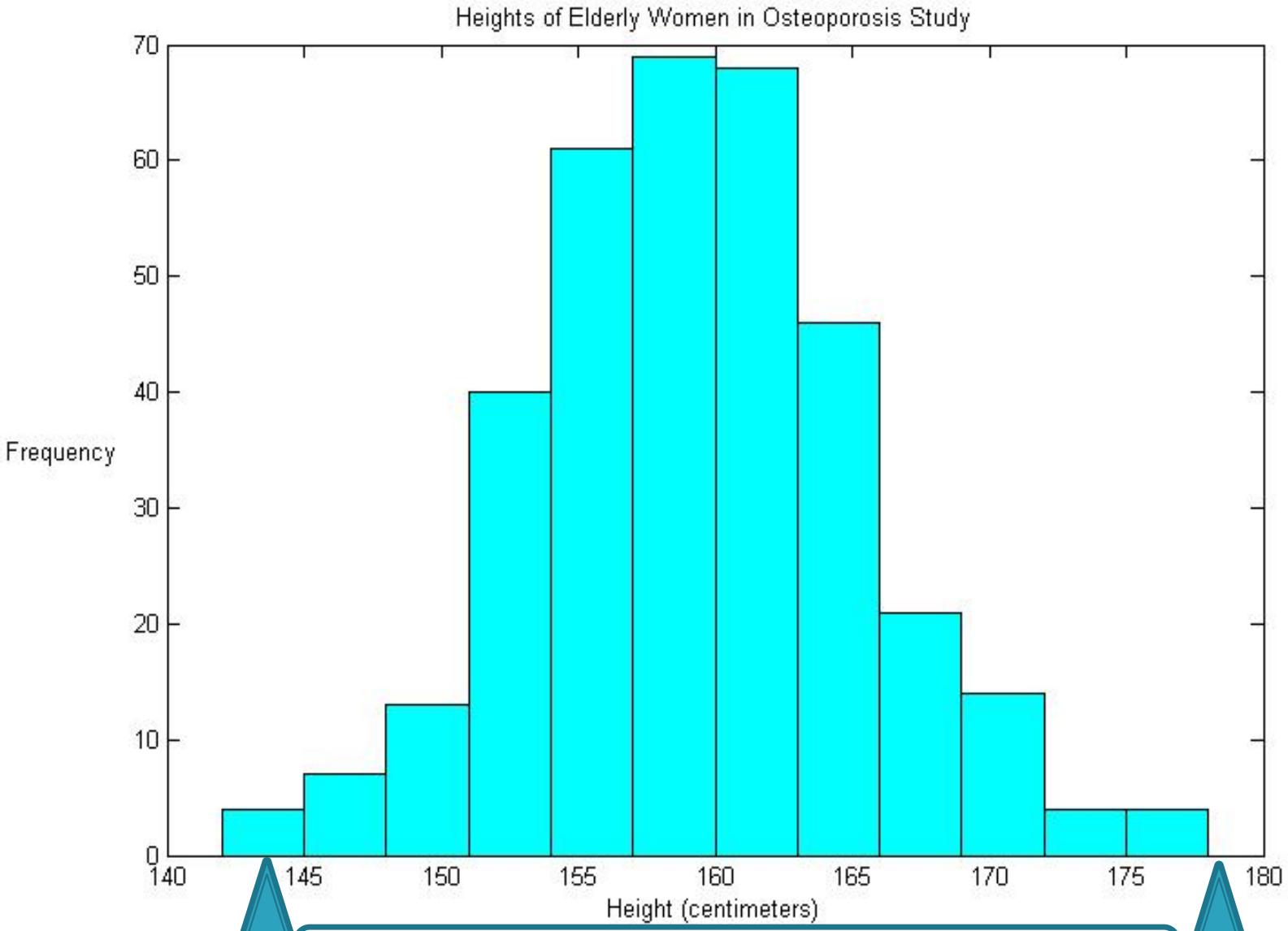

Waiting Times between Eruptions of the Old Faithful Geyser
Frequency
0
10
20
30
40
50
60
70
80
40
50
60
70
80
90
100
110
Time (Minutes)
$\bar{y} - 2s$
$\bar{y}$
$\bar{y} + 2s$

Range

The range is defined as:

$$\text{range} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$$
$$= y_{(n)} - y_{(1)}$$

The range is a very crude measure of the spread of the data.



Range = 178 – 142 = 36 cm

Percentiles and Quartiles

To define the interquartile range (IQR) we first define the p^{th}-percentile for a dataset.

The p^{th} percentile (also called the p^{th} quantile) is the value such that p percent of the data fall at or below this value.

Like the median we need a definition that works for all sizes of data sets.

Definition 1

The p^{th} quantile ($0 < p < 1$) is the value, call it $q(p)$, determined as follows:

- **Let $m = (n+1)p$ where n is the sample size.**
- **If m is and integer and $1 \le m \le n$ then take the m^{th} *smallest value* $q(p) = y_{(m)}$.**
- **If m is not an integer but $1 < m < n$ then determine the closest integer j, such that $j < m < j+1$ and take**

$$q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right].$$

Percentiles and Quartiles

The lower or first quartile, $q(0.25)$, is the 25^{th} percentile or the median of the observations below the median.

The median, $q(0.50)$, is the 50^{th} percentile.

The upper or third quartile, $q(0.75)$, is the 75^{th} percentile or the median of the observations above the median.

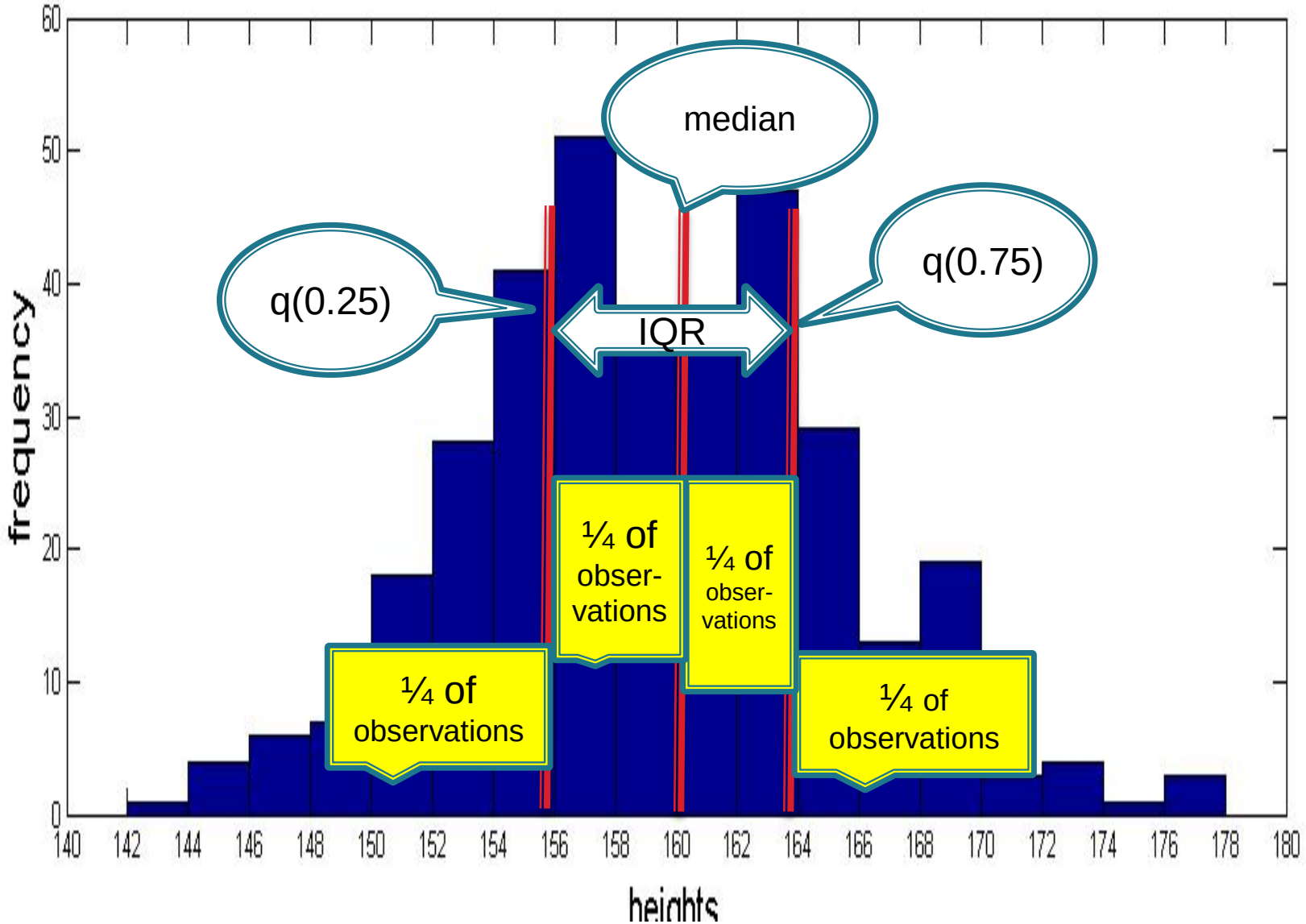
Interquartile Range (IQR)

The interquartile range (IQR) is defined as

$$\text{IQR} = q(0.75) - q(0.25)$$

The interquartile range is a more "robust" measure of spread since, unlike the range, it will not be affected by extreme values or outliers in the dataset.

Quartiles for Heights of Elderly Women



IQR for Old Faithful Geyser data