To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, x_3, …, x_n, then the mean is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \ldots x_n}{n}$$

or in more compact notation,

$$\bar{x} = \frac{1}{n}\sum x_1$$

The **median M** is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

To calculate the quartiles:
1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile Q1** is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile Q3** is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The **interquartile range (IQR)** is the distance between the first and third quartiles, IQR = Q3 - Q1.

68 - 60 = 8

Call an observation an **outlier** if it falls more than 1.5 IQR above the third quartile or below the first quartile.

IQR is a number -
Many students write things like "The IQR goes from 15 to 32".

Every AP grader knows exactly what you mean, namely, "The box in my boxplot goes from 15 to 32.", but this statement is not correct. The IQR is defined as Q3 - Q1 which gives a single value. Writing the statement above is like saying "17 goes from 15 to 32." It just doesn't make sense.

The **five number summary** of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum Q1 M Q3 Maximum

A **modified boxplot** is a graph of the five number summary, with outliers plotted individually.

* A central box spans the quartiles.
* A line in the box marks the median.
* Observations more than 1.5 IQR outside the central box are plotted individually.
* Lines extend from the box out to the smallest and largest observations that are not outliers.

Box plots do not display shape well.

Heights of 91 high school males at a suburban public school.

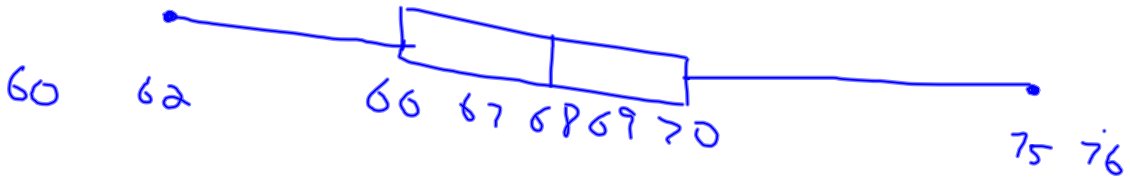
Height (inches)	Tally
62	|
63	|||
64	𝍸 |
65	𝍸 ||
66 Q_1	𝍸 0 |||
67	𝍸 𝍸 ||
68 M	𝍸 0 𝍸 |||
69	𝍸 𝍸
70 Q_3	𝍸 0 𝍸
71	𝍸 ||||
72	𝍸 |
73	|||
74	|
75	|

Answer

Box plots display well:
outliers
center
spread

but not
shape
clusters
gaps

IQR = 70 − 66
= 4



Min Q_1 Med Q_3 Max

$IQR=Q_3-Q_1=$

$1.5(IQR)=$

Outlier rule:

$Q_1-1.5(IQR)=$

$Q_3-1.5(IQR)=$

The variance s^2 of a set of observations is the average of the squares of the deviations from their mean. In symbols, the variance of n observations $x_1, x_2, x_3, ..., x_n$ is

$$s^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots \left(x_n - \bar{x}\right)^2}{n-1}$$

or, more compactly,

$$s^2 = \frac{\sum\left(x_1 - \bar{x}\right)^2}{n-1}$$

The standard deviation s is the square root of the variance s^2

$$s = \sqrt{\frac{\sum\left(x_1 - \bar{x}\right)^2}{n-1}}$$

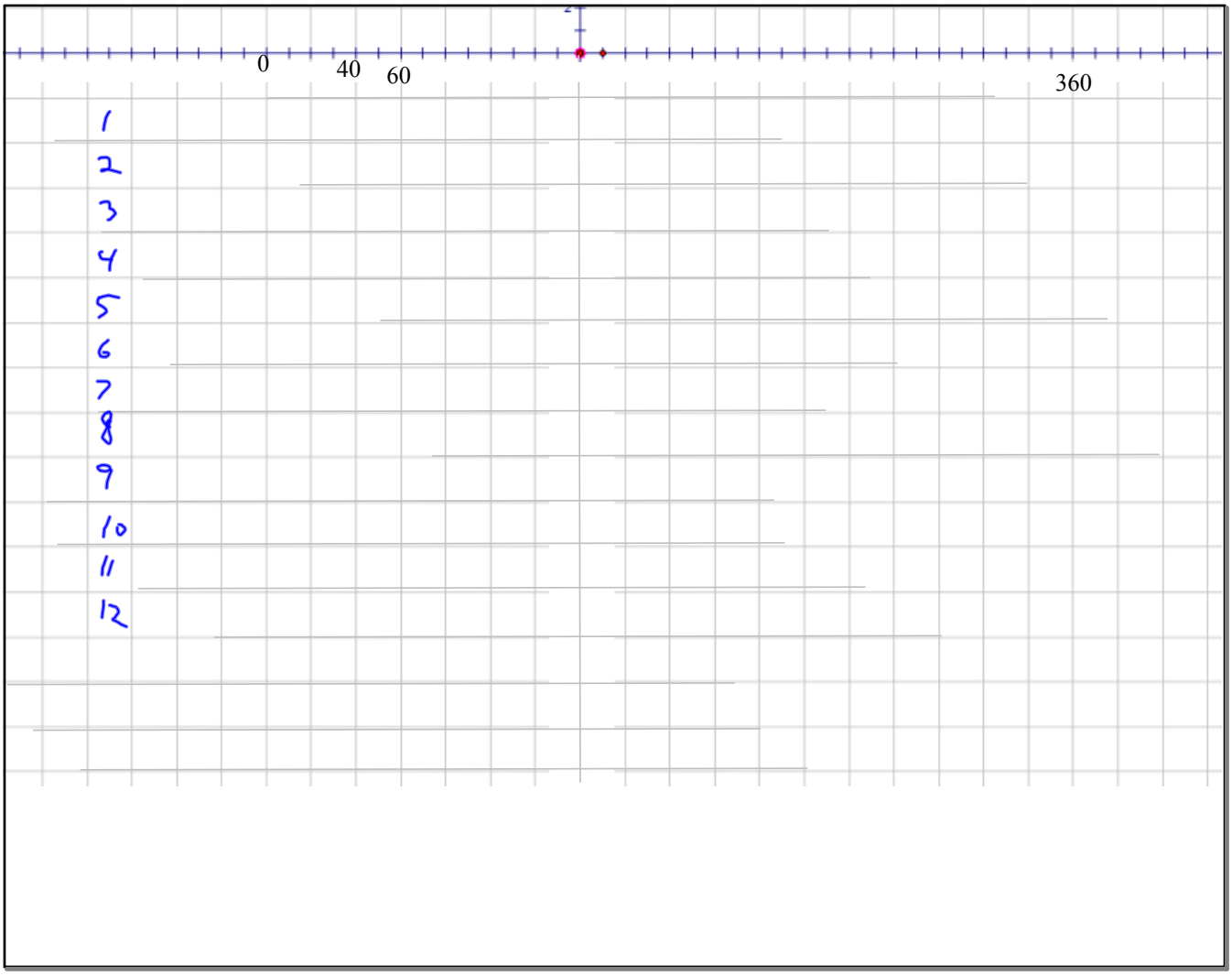
$$s^2 = \frac{\sum\left(x_1 - \bar{x}\right)^2}{n-1}$$

Properties of the standard deviation, s

* s measures spread about $\bar{x}$ and should be used only when $\bar{x}$ is chosen as the measure of center.
* s = 0 only when there is no spread. This happens only when all observations have the same value. Otherwise, s>0. As the observations become more spread out about $\bar{x}$, s gets larger.
* s, like $\bar{x}$, is not resistant. Strong skewness or a few outliers can make s very large.

	skewed or has outliers	nearly symmetric
center	median M	mean $\bar{x}$
spread	$IQR = Q_3 - Q_1$	standard deviation s

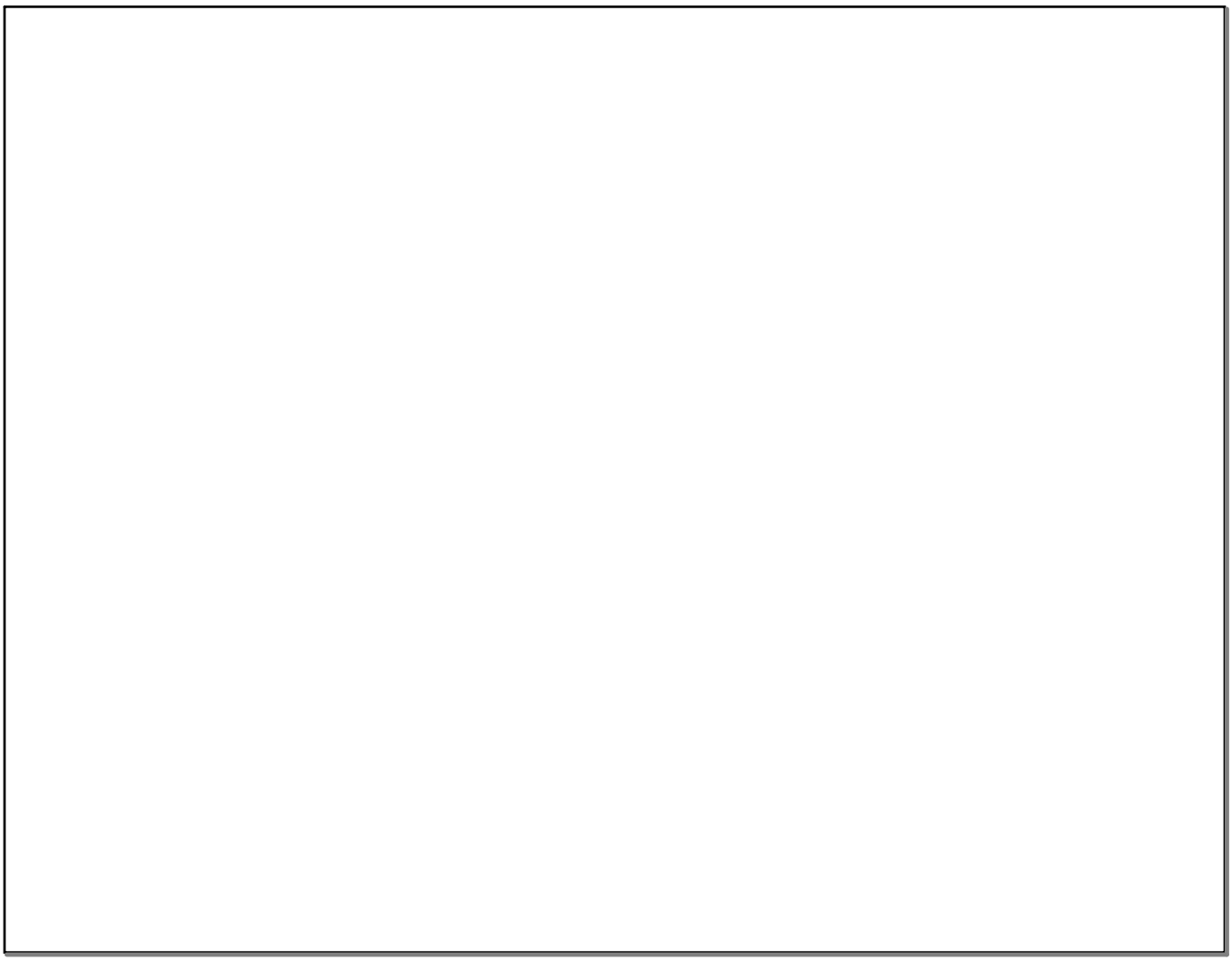
Choosing a summary
The five number summary is usually better than $\bar{x}$ and s for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.


0
40
60
360
1
2
3
4
5
6
7
8
9
10
11
12

Here's more info about the Vietnam Draft, in case you want to check out the pictures, data, or history involved.

http://www.sss.gov/lotter1.htm

http://www.niles-hs.k12.il.us/timmil/draft_project.html



A linear transformation changes the original variable x into a new variable x_{new} by an equation of the form

$$x_{new}=a+bx$$

Adding the constant a shifts all values of x upward or downward by the same amount.

Multiplying by the positive constant b changes the size of the unit of measurement.

To see the effect of a linear transformation on measures of center and spread, apply these rules:

Multiplying each observation by positive number b multiplies both measures of center (mean and median) and measures of spread (standard deviation and IQR) by b.

Adding the same number a (either positive or negative) to each observation adds a to the measures of center and to quartiles but does not change measures of spread.

The effect of changing units or using linear transformations

Change	Effect on summary statistic
add the same number c to each data value	add that same number c to each statistic (c + the mean, median, quartile, max, or min) However, s remains the same.
multiply the same number c by each data value	multiply that same number c by each statistic (c times the mean, median, quartile, max, or min) However, s is multiplied by the absolute value of of c.