4.3 Measures of Variation

- How much variation is there in the data?
- Look for the spread of the distribution.
- What do we mean by “spread”?

Example Data set:

- Weight of contents of regular cola (grams).

368, 367, 369, 370, 369, 370
366, 373, 365, 362, 378, 368

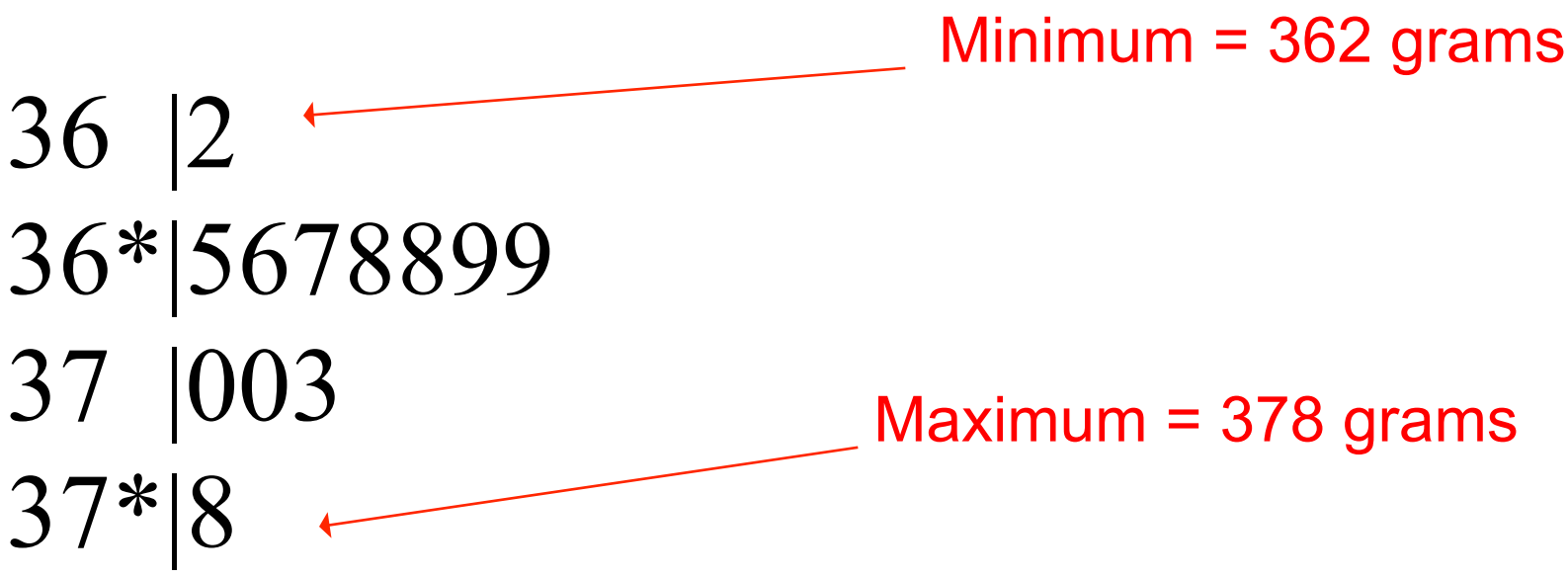
Example Data set:

- Weight of contents of regular cola (grams).
  - Put the data in ascending order:

362, 365, 366, 367, 368, 368,
369, 369, 370, 370, 373, 378

The stem and leaf plot

- Weight of contents of regular cola



Measures of Spread (or variation)

- **Range**
  - The distance from the minimum to the maximum.

    (378 – 362) = **16 grams**
  - The length of the interval that contains 100% of the data.
  - This measure of variation **is greatly affected by outliers.**

Range

- Definition
  - The **range** of a set of data values is the difference between its highest and lowest data values.
- range = highest value (max) - lowest value (min)
- Range is a **single value**! (not two values).

EXAMPLE 2 Misleading Range

Consider the following two sets of quiz scores for nine students.

Which set has the greater range?

Would you also say that this set has the greater variation?

Quiz 1:	1	10	10	10	10	10	10	10	10
Quiz 2:	2	3	4	5	6	7	8	9	10

Quiz 2 has greater variation in scores even though Quiz 1 has greater range.

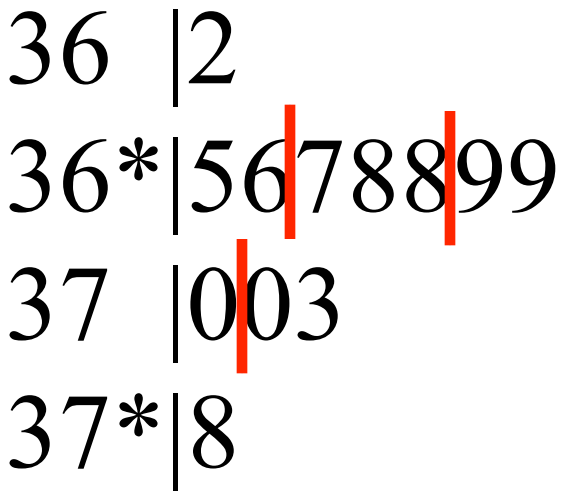
Measures of Spread

- **Quartiles** (notice the reference to 4 parts)
  - Consider a few intermediate values in the distribution to describe the spread.
  - The quartiles are values that divide the data into quarters.

1st quartile = 366.5

Median = 368.5

3rd quartile = 370



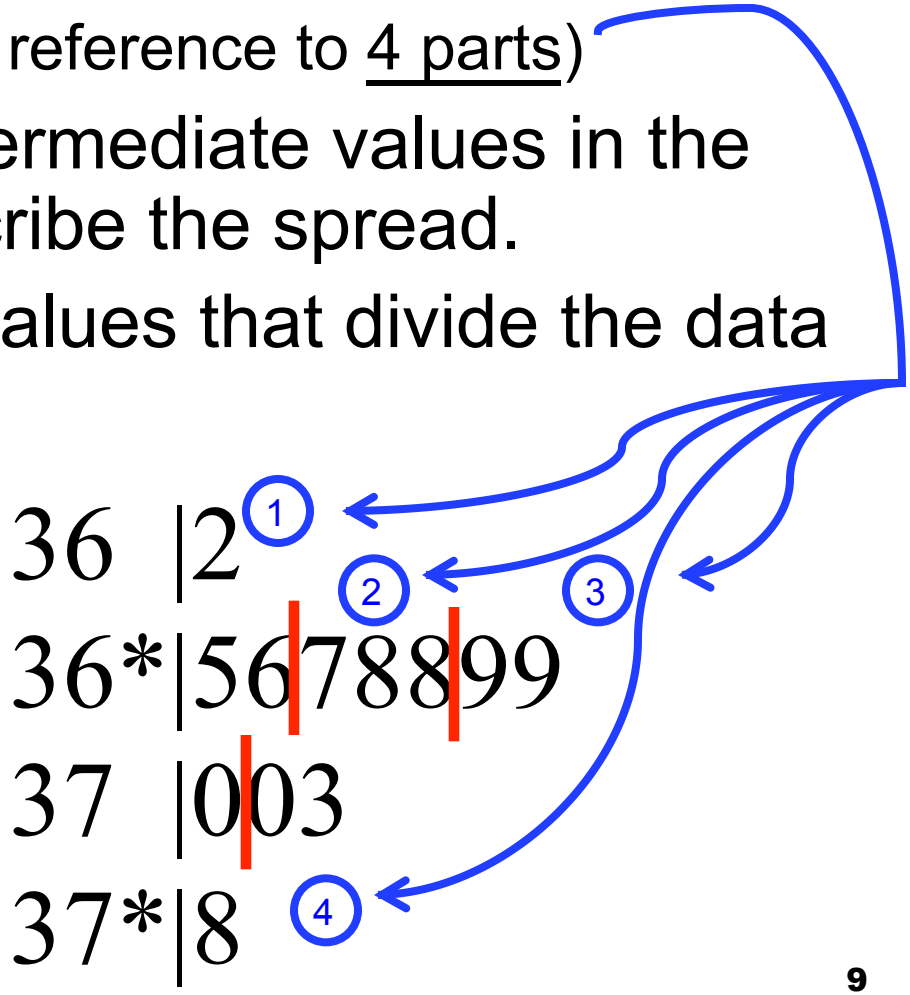
Measures of Spread

- **Quartiles** (notice the reference to 4 parts)
  - Consider a few intermediate values in the distribution to describe the spread.
  - The quartiles are values that divide the data into quarters.

1st quartile = 366.5

Median = 368.5

3rd quartile = 370

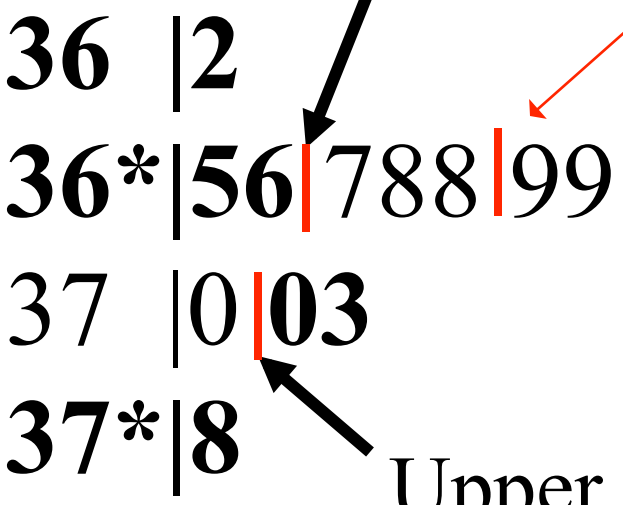


Quartiles

n = 12

Median = 368.5 grams

Lower quartile = (366+367)/2
= 366.5 grams



Upper quartile = (370+370)/2
= 370.0 grams

Measures of Spread

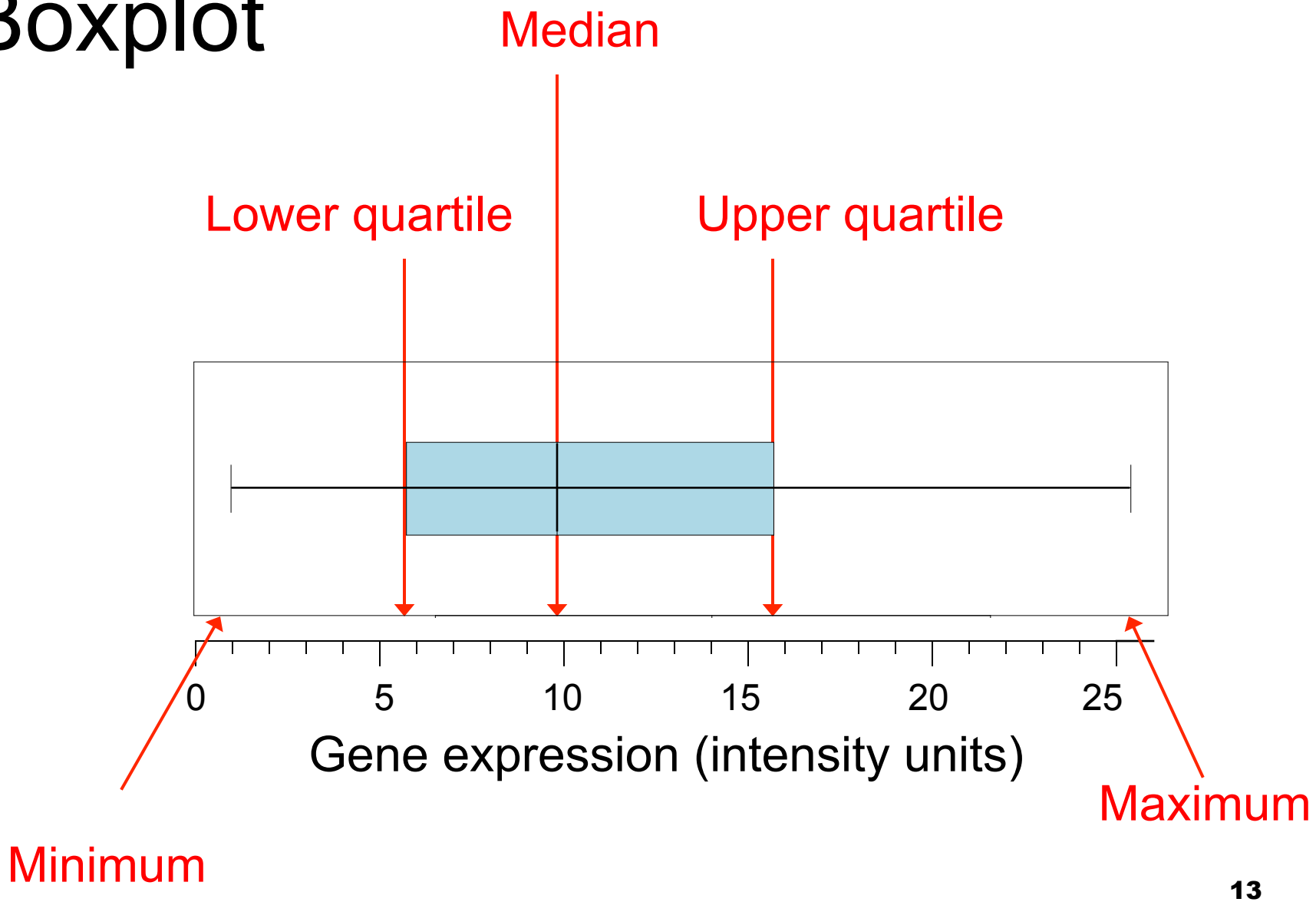
- **Quartiles** (notice the reference to 4 parts)
  - The first quartile (or lower quartile) is the **median of the lower half**.
  - The third quartile (upper quartile) is the **median of the upper half**.

Measures of Spread

- **Five Number Summary** (describes spread)
  - Minimum 362 grams
  - Lower Quartile, Q_1 366.5 grams
  - Median 368.5 grams
  - Upper Quartile, Q_3 370 grams
  - Maximum 378 grams

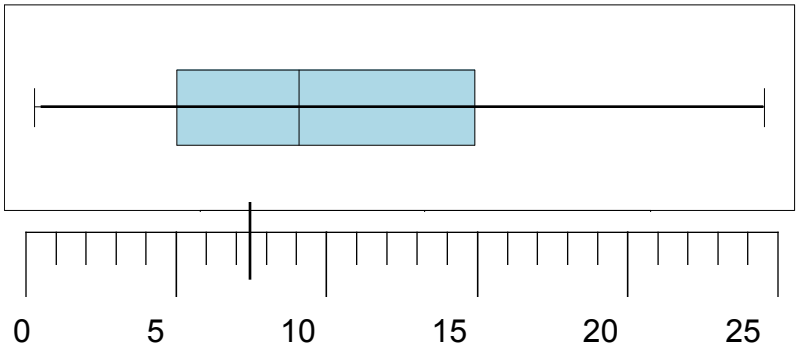
Leads to a boxplot…

Boxplot


Median
Lower quartile
Upper quartile
0
5
10
15
20
25
Gene expression (intensity units)
Minimum
Maximum

Drawing a Boxplot

a.k.a. box-n-whiskers



- Establish an axis with a scale.
- Draw a box that extends from the lower (Q_1) to the upper quartile (Q_3).
- Draw a line from the lower quartile (Q_1) to the minimum and another line from the upper quartile (Q_3) to the maximum.
- Draw a line at the median (Q_2) .

EXAMPLE 3 Wait times at banks

Wait times for 11 customers at two banks are shown below in minutes.

			Lower quartile (Q_1)			Median (Q_2)			Upper quartile (Q_3)		
Big Bank:	4.1	5.2	**5.6**	6.2	6.7	**7.2**	7.7	7.7	**8.5**	9.3	11.0
Best Bank:	6.6	6.7	**6.7**	6.9	7.1	**7.2**	7.3	7.4	**7.7**	7.8	7.8

EXAMPLE 3 Wait times at banks

Five Number Summary

Big Bank:		*Best Bank:*	
low	= 4.1	low	= 6.6
lower quartile	= 5.6	lower quartile	= 6.7
median	= 7.2	median	= 7.2
upper quartile	= 8.5	upper quartile	= 7.7
high	= 11.0	high	= 7.8

Drawing a Boxplot

Step 1. Draw a number line that spans all the values in the data set.

Step 2. Enclose the values from the lower to the upper quartile in a box. (The thickness of the box has no meaning.)

Step 3. Draw a line through the box at the median.

Step 4. Add “whiskers” extending to the low and high values.

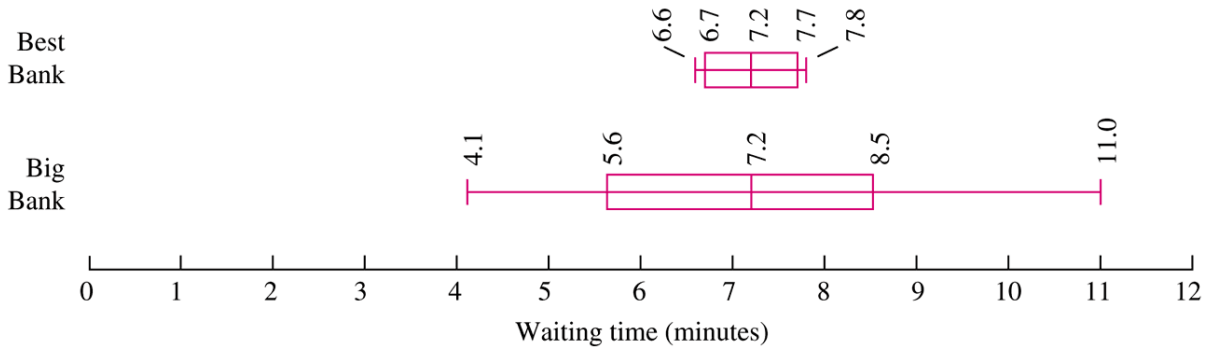
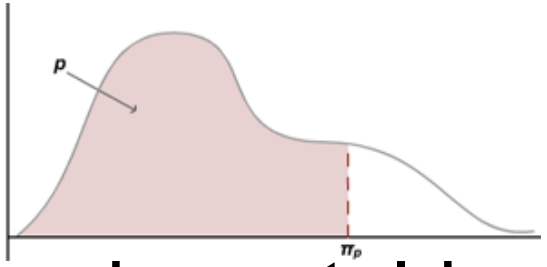


Figure 4.14 Boxplots show that the variation of the waiting times is greater at Big Bank than at Best Bank.

Percentiles

- Quartiles divide a data set into 4 segments, but it is possible to divide a data set even more.
- For example, we could consider the *deciles* which divide a data set into 10 segments.
- Or we could divide the data set into 100 segments using ***percentiles***.

Percentiles



- If you are tall, you might have been told that you are in the **95th percentile** in height, meaning that you are taller than 95% of the population.
- When you took the ACT Exams, you might have been told that you are in the **80th percentile** in math ability, meaning that you scored better than 80% of the population on the math portion of the SAT Exams.

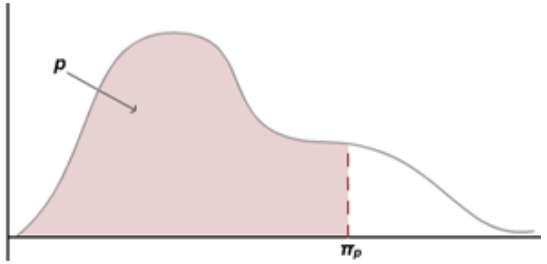
SAT® Percentile Ranks

2012 College-Bound Seniors — Critical Reading, Mathematics and Writing Percentile Ranks

This table allows you to compare a student's score with the performance of other test-takers. The percentile rank shows students what percentage of college-bound seniors earned a lower score than they did. A student with a writing score of 510, for example, can see that 58 percent of other test-takers scored lower.

	Critical Reading	Mathematics	Writing
Score	Percentile	Percentile	Percentile
800	99+	99	99+
790	99	99	99
780	99	98	99
770	99	98	99
760	99	97	99
750	98	97	98
740	98	96	98
730	97	96	98
720	97	95	97
710	96	94	96
700	95	93	96
690	94	92	95
680	93	90	94

Percentiles

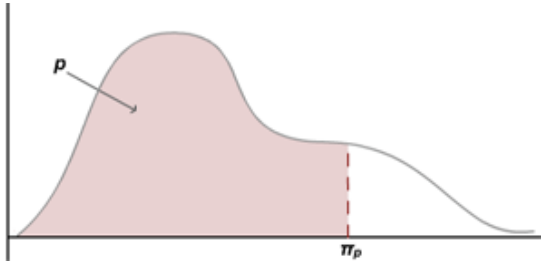


Definition

The ***n*th percentile** of a data set divides the bottom *n*% of data values from the top (100 - *n*)%. A data value that lies between two percentiles is often said to lie *in* the lower percentile. You can approximate the percentile of any data value with the following formula:

$$\text{percentile of data value} = \frac{\text{number of values less than this data value}}{\text{total number of values in data set}} \times 100$$

Percentiles



Example

A data set consists of the 85 ages of women at the time that they won an Oscar in the category of best actress.

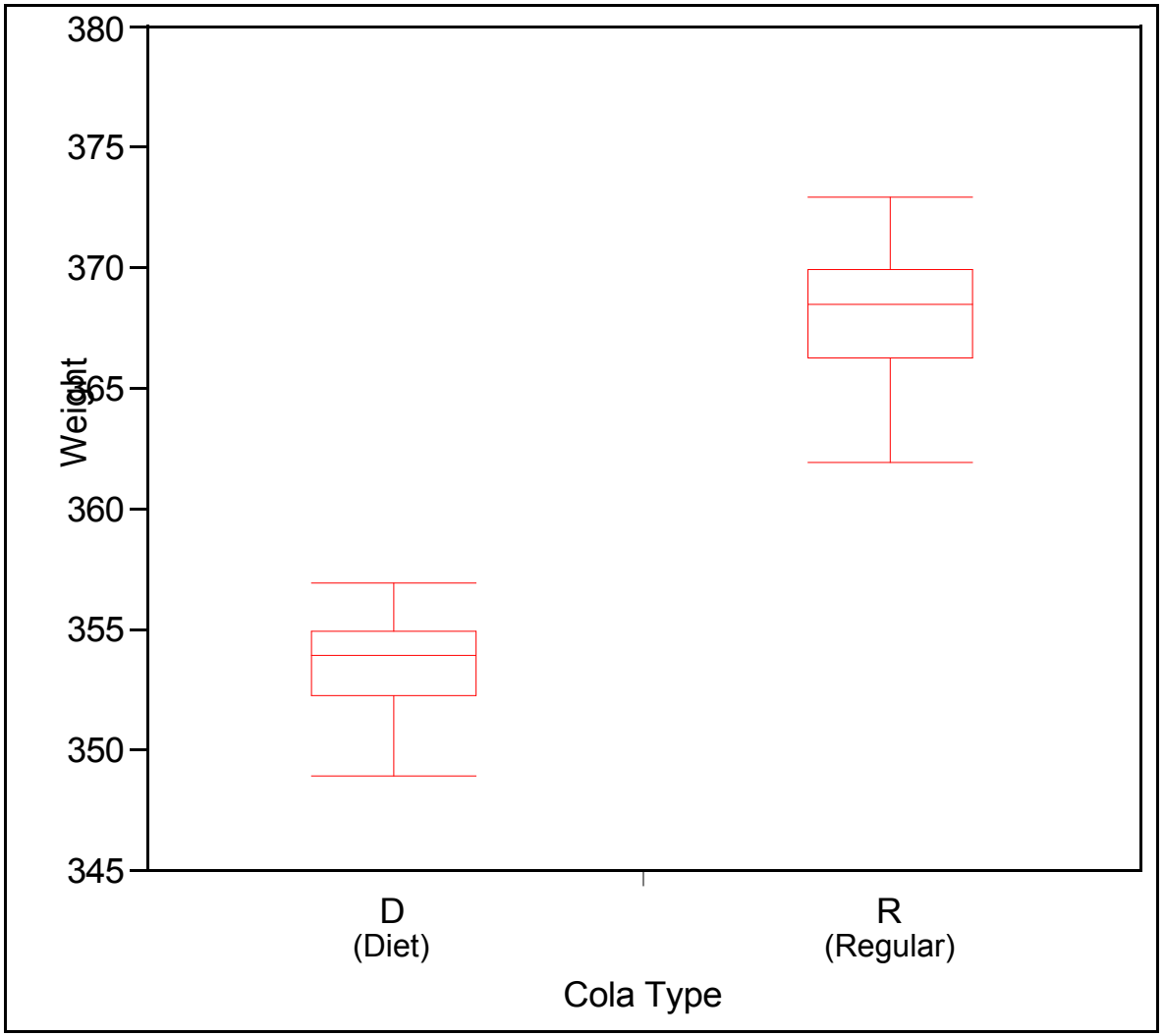
a) One of the actresses was 40 years of age, and she was older than 63 of the other actress at the time they they won Oscars. What is the percentile of the age of 40?

ANS: So, she was 64^{th} in the order out of 85 individuals.

$$\frac{63}{85} \times 100 = 74$$ which means 74^{th} percentile.

Comparing Distribution of Groups

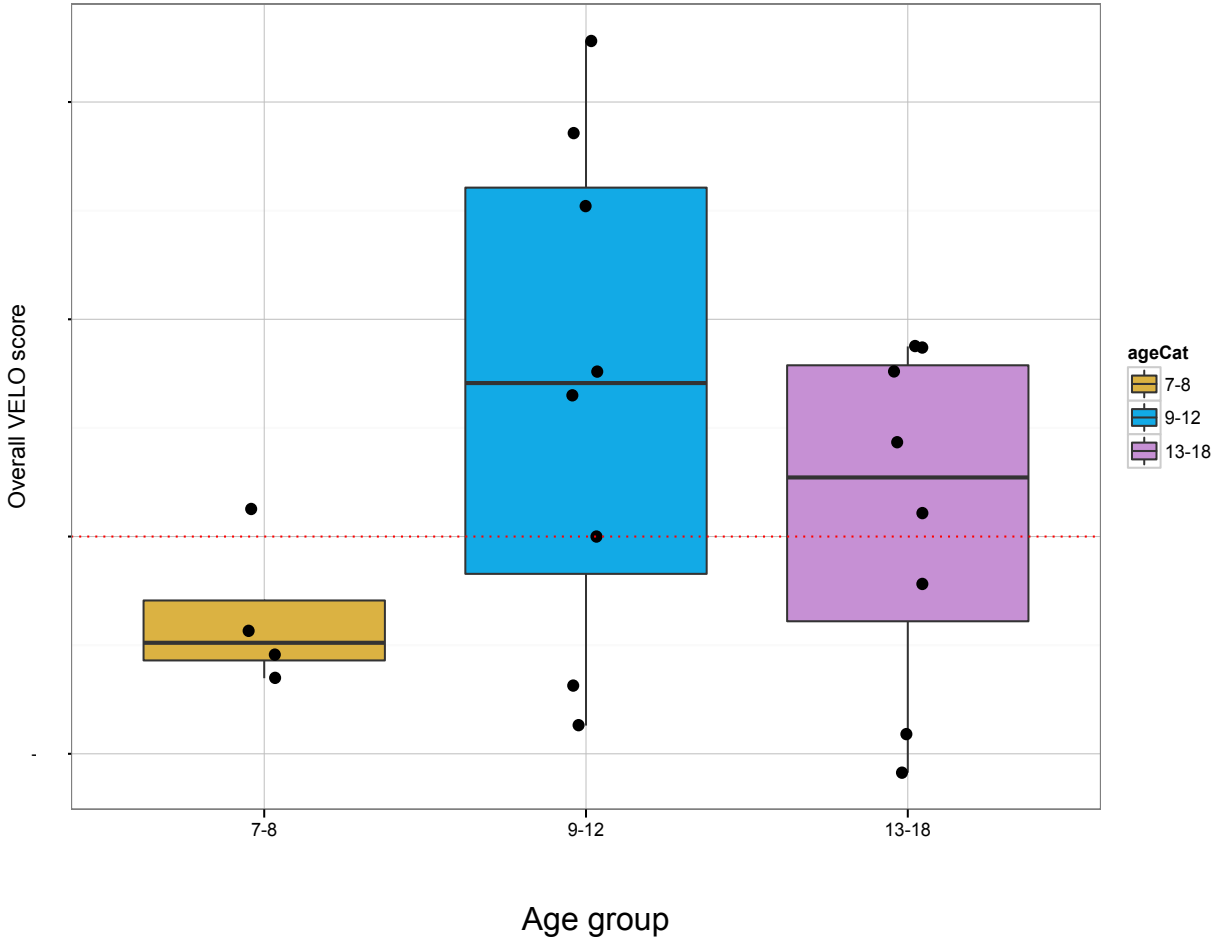
- Compare shape.
- Compare center.
- Compare spread.



Comparing Distribution of Groups

- Compare shape.
- Compare center.
- Compare spread.

VELO=Quality of Life Outcome instrument



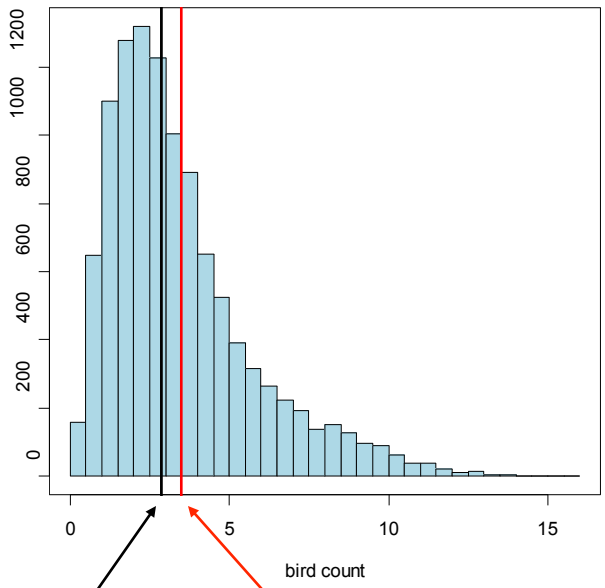
Comment (shapes and centers)

- For skewed distributions, the **median** is a better measure of center than the **mean**.
- For skewed distributions, the **mean** is ‘pulled’ toward the tail (or the direction of the skew), but the **median** is not.
- The **mean** can be greatly affected by outliers (extreme values) in the tail of a distribution.

Mean vs. Median

For right-skewed data,

Mean > Median

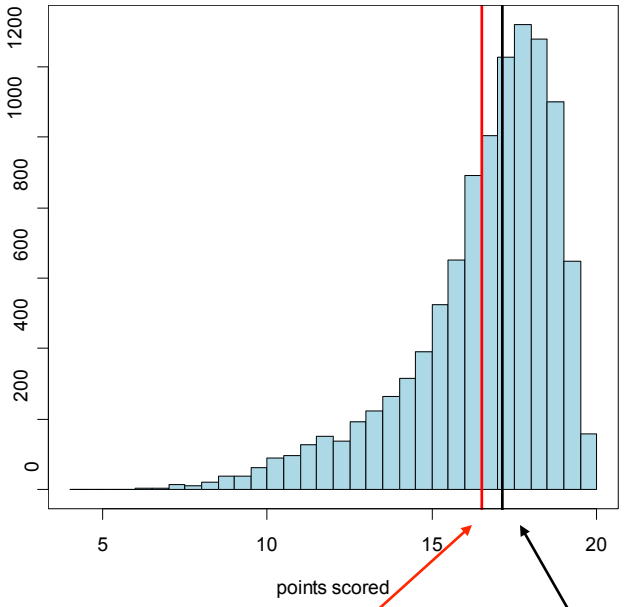


Mean = 3.48

Median = 2.88

For left-skewed data,

Mean < Median

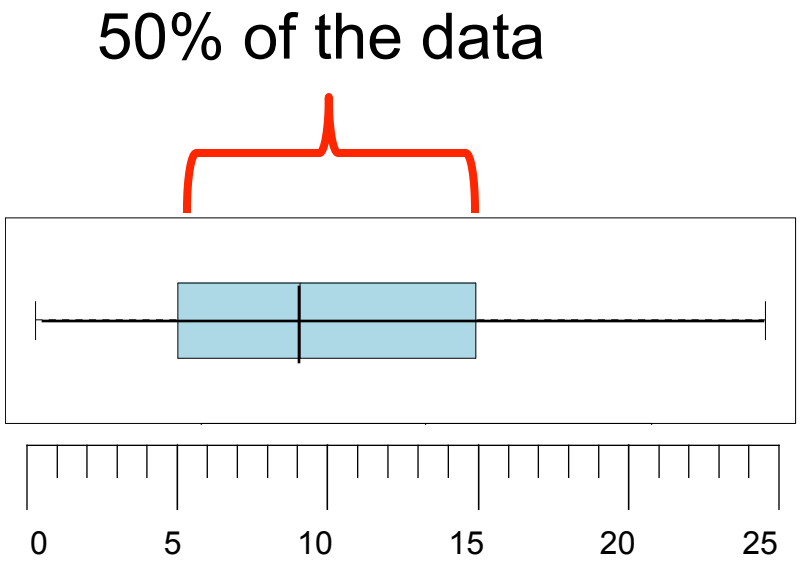


Median = 17.11

Mean = 16.52

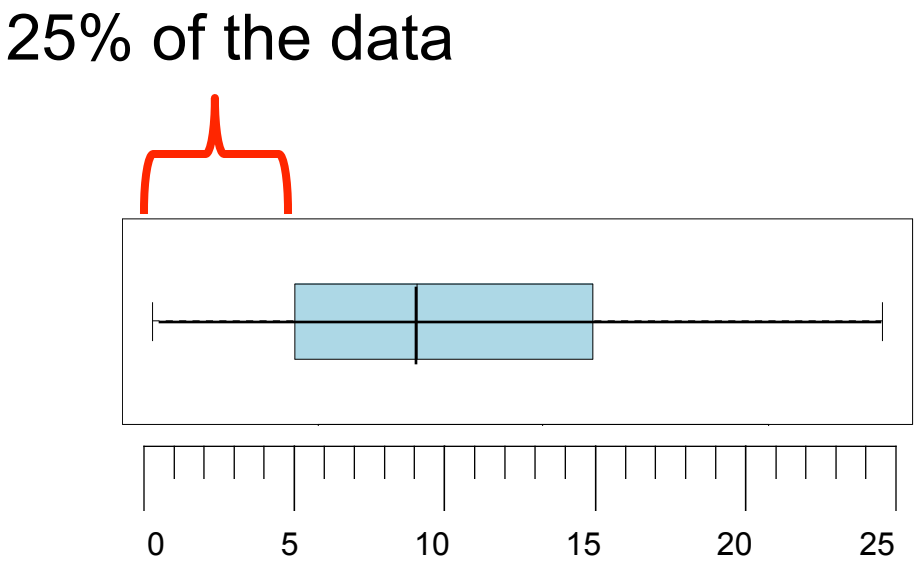
Comment (boxplots and the IQR)

- 50% of the observations **fall between** the lower quartile (Q_1) and the upper quartile (Q_3).



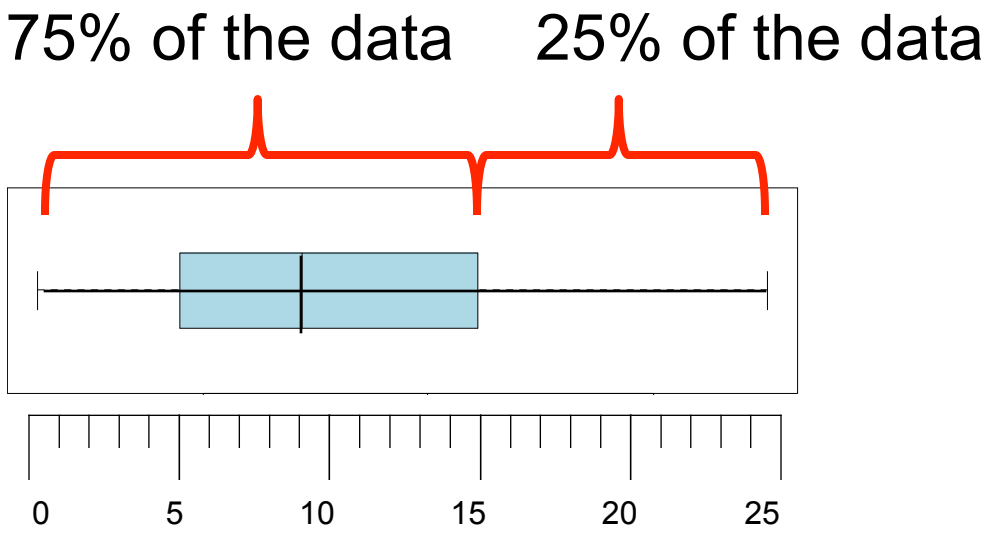
Comment (boxplots and the IQR)

- 25% of the observations fall **below** the lower quartile (Q_1).



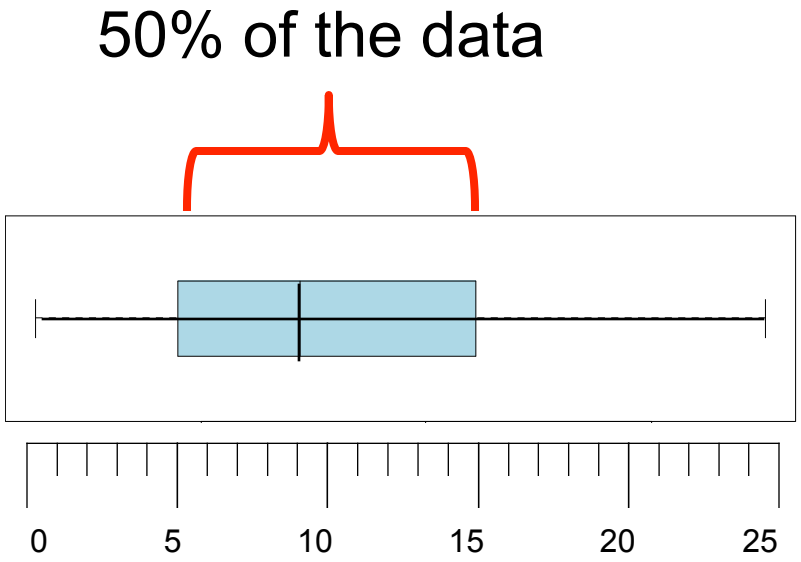
Comment (boxplots and the IQR)

- 25% of the observations fall **above** the upper quartile (Q_3).
- Or similarly, 75% of the observations fall **below** the upper quartile (Q_3).



Comment (boxplots and the IQR)

- The **range** of the middle 50% of the data is called the **Interquartile Range** or **IQR**.

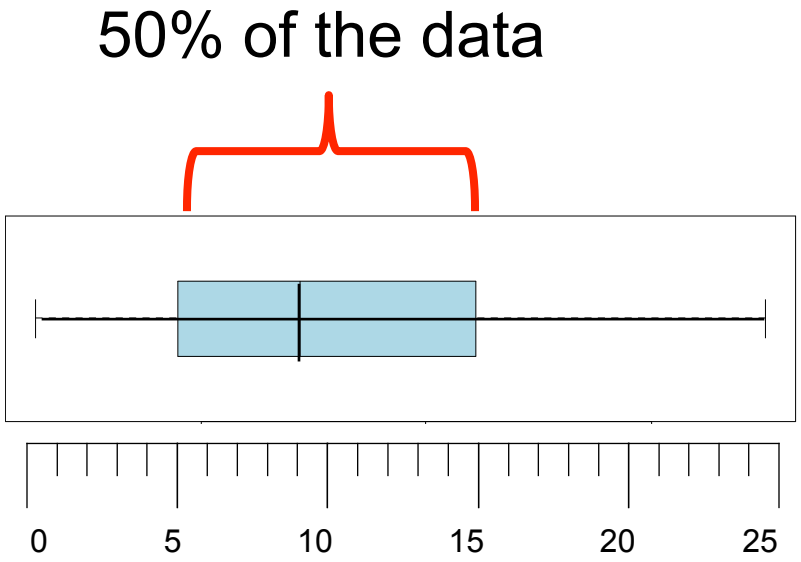


$$IQR = Q_3 - Q_1 = 15 - 5 = 10$$

NOTE: Range = max-min=25-1=24

Comment (boxplots and the IQR)

- The **IQR** quantifies the spread of the middle 50% of the data. It is a measure of variation.

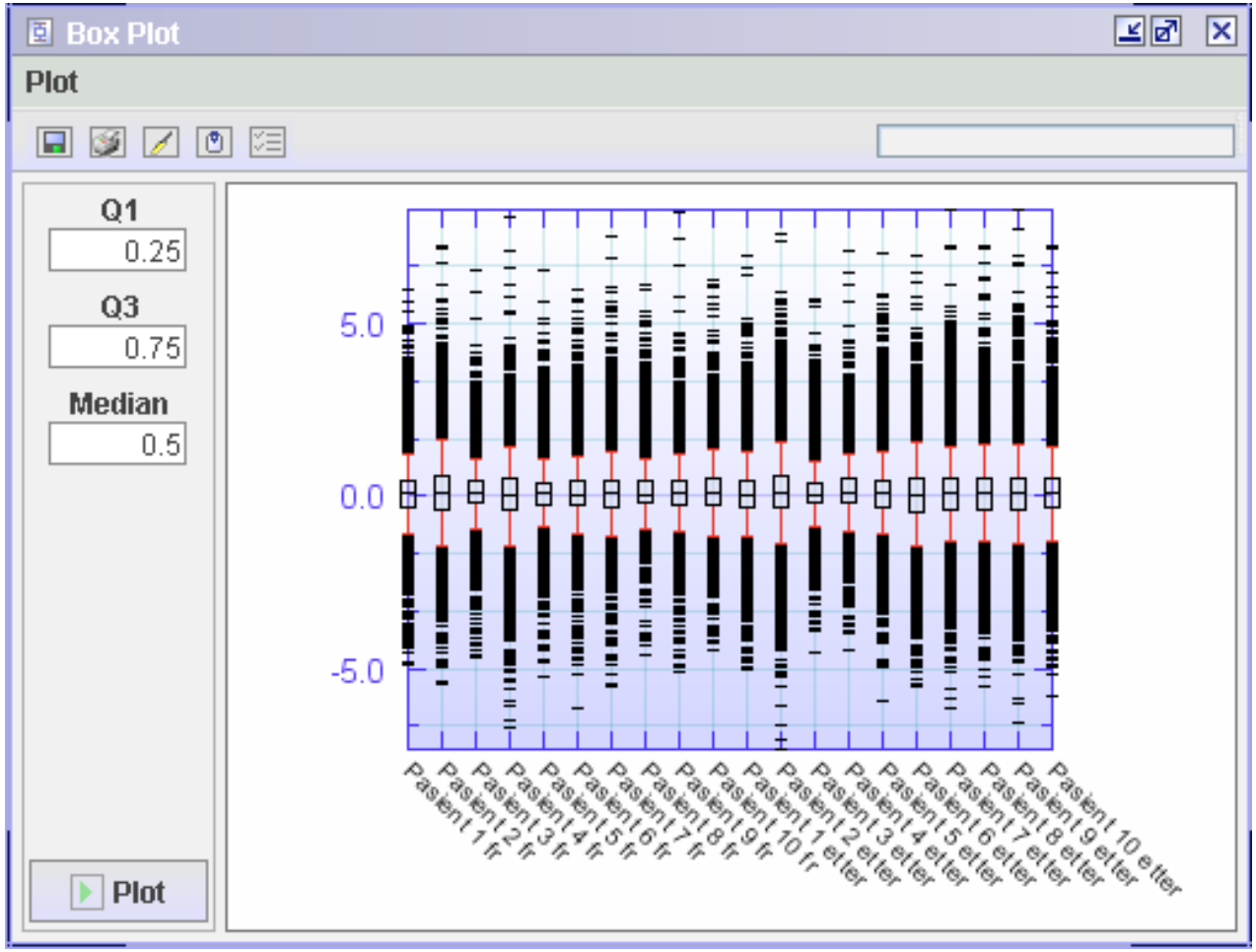


IQR =15 – 5 = 10

Boxplot comparisons

- Can be used to check for technical problems (e.g. gene expression values).
- Usually, you want to see similar characteristics (center, shape, spread) for all side-by-side boxplots.

Boxplot comparisons



20 different ‘slides’.

Centers all around 0.

Some have more spread than others.

The **IQR** is larger for some ‘slides’ (you can see this from the rectangles at the middle of the plot).