Measures of Dispersion

- Measures of central tendency alone cannot completely characterize a set of data. Two very different data sets may have similar measures of central tendency.

- Measures of dispersion are used to describe the spread, or variability, of a distribution

- Common measures of dispersion: range, variance, and standard deviation

Range

Range: The difference in value between the highest-valued (H) and the lowest-valued (L) pieces of data:

$$\text{range} = H - L$$

- Other measures of dispersion are based on the following quantity

Deviation from the Mean: A deviation from the mean, $x - \overline{x}$, is the difference between the value of x and the mean $\overline{x}$

Example

✓ **Example:** Consider the sample {12, 23, 17, 15, 18}.
Find 1) the range and 2) each deviation from the mean.

Solutions:

1) $\text{range} = H - L = 23 - 12 = 11$

2) $\bar{x} = \frac{1}{5}(12+23+17+15+18) = 17$

Data x	Deviation from Mean $x - \bar{x}$
12	-5
23	6
17	0
15	-2
18	1

Mean Absolute Deviation

Note: $\sum (x - \bar{x}) = 0$ (Always!)

Mean Absolute Deviation: The mean of the absolute values of the deviations from the mean:

$$\text{Mean absolute deviation} = \frac{1}{n}\sum |x - \bar{x}|$$

For the previous example:

$$\frac{1}{n}\sum |x - \bar{x}| = \frac{1}{5}(5+6+0+2+1) = \frac{14}{5} = 2.8$$

Sample Variance & Standard Deviation

Sample Variance: The sample variance, s^2, is the mean of the squared deviations, calculated using $n - 1$ as the divisor:

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 \quad \text{where } n \text{ is the sample size}$$

Note: The numerator for the sample variance is called the sum of squares for x, denoted SS(x):

$$s^2 = \frac{SS(x)}{n-1} \qquad \text{where} \qquad SS(x) = \sum(x-\bar{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

Standard Deviation: The standard deviation of a sample, s, is the positive square root of the variance:

$$s = \sqrt{s^2}$$

Example

✓ **Example:** Find the 1) variance and 2) standard deviation for the data {5, 7, 1, 3, 8}:

Solutions:

First: $\bar{x} = \frac{1}{5}(5+7+1+3+8) = 4.8$

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	5	0.2	0.04
	7	2.2	4.84
	1	-3.8	14.44
	3	-1.8	3.24
	8	3.2	10.24
Sum:	**24**	**0**	**32.08**

1) $s^2 = \frac{1}{4}(32.8) = 8.2$

2) $s = \sqrt{8.2} = 2.86$

Notes

- The shortcut formula for the sample variance:

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

- The unit of measure for the standard deviation is the same as the unit of measure for the data

Mean & Standard Deviation of Frequency Distribution

- If the data is given in the form of a frequency distribution, we need to make a few changes to the formulas for the mean, variance, and standard deviation

- Complete the extension table in order to find these summary statistics

To Calculate

- In order to calculate the mean, variance, and standard deviation for data:

1. In an *ungrouped* frequency distribution, use the frequency of occurrence, *f*, of each observation

2. In a *grouped* frequency distribution, we use the frequency of occurrence associated with each class midpoint:

$$\bar{x} = \frac{\sum xf}{\sum f} \qquad s^2 = \frac{\sum x^2 f - \frac{\left(\sum xf\right)^2}{\sum f}}{\sum f - 1}$$

Example

- ✓ **Example:** A survey of students in the first grade at a local school asked for the number of brothers and/or sisters for each child. The results are summarized in the table below. Find 1) the mean, 2) the variance, and 3) the standard deviation:

Solutions:

First:

	x	f	xf	$x^2 f$
	0	15	0	0
	1	17	17	17
	2	23	46	92
	4	5	20	80
	5	2	10	50
Sum:		**62**	**93**	**239**

1) $\bar{x} = 93/62 = 1.5$

2) $s^2 = \dfrac{239 - \dfrac{(93)^2}{62}}{62-1} = 1.63$

3) $s = \sqrt{1.63} = 1.28$

Measures of Position

- Measures of position are used to describe the relative location of an observation
- Quartiles and percentiles are two of the most popular measures of position
- An additional measure of central tendency, the midquartile, is defined using quartiles
- Quartiles are part of the 5-number summary

Measures of Relative Position

- **A measure of relative position tells where data values fall within the ordered set.**

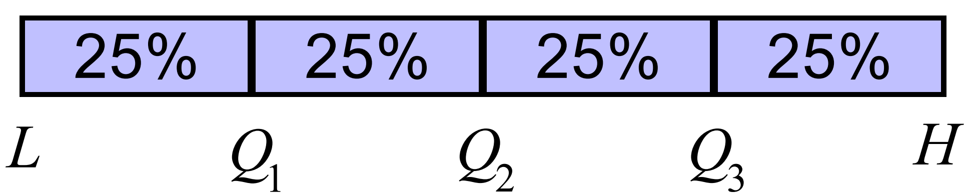
- **The measures of relative position we will calculate are the quartiles, percentiles, and standard score.**

Quartiles

Quartiles: Values of the variable that divide the ranked data into quarters; each set of data has three quartiles

1. The first quartile, Q_1, is a number such that at most 25% of the data are smaller in value than Q_1 and at most 75% are larger
2. The second quartile, Q_2, is the median
3. The third quartile, Q_3, is a number such that at most 75% of the data are smaller in value than Q_3 and at most 25% are larger

Ranked data, increasing order



Quartiles:

- **Quartiles divide a data set into four equal parts.**
- **To find the quartiles of a data set:**

1. **Find the median, Q_2.**
2. **Use the median to divide the data into two groups.**
   a. **For an odd number of data points, include the median in both the upper and lower halves.**
   b. **For an even number of data points, do not include the median in either half.**
3. **The median of the lower group is Q_1 and the median of the upper group is Q_3.**

Find the quartiles for the following data set:

2 3 5 7 8 9 10 12 15

Solution:

First find the median.

$Q_2 = 8$.

Now, find the median of the first half of data.

$Q_1 = 5$.

Finally, find the median of the second half of data.

$Q_3 = 10$.

Find the quartiles for the following data set:

10 12 14 15 14 16 17 18 10 19 17 17

Solution:

First order the data.

10 10 12 14 14 15 16 17 17 17 18 19

Q_2 = 15.5

Q_1 = 13

Q_3 = 17

Find the quartiles for the following data set:

11	**11**	**14**	**15**	**16**
16	**17**	**19**	**22**	**25**
26	**27**	**31**	**34**	**36**

Solution:

First order the data.

11 11 14 15 16 16 17 19 22 25 26 27 31 34 36

$Q_2 = 19$

$Q_1 = 15.5$

$Q_3 = 26.5$

Midquartile

Midquartile: The numerical value midway between the first and third quartile:

$$\text{midquartile} = \frac{Q_1 + Q_3}{2}$$

✓ **Example:** Find the midquartile for the 20 pH values in the previous example:

$$\text{midquartile} = \frac{Q_1 + Q_3}{2} = \frac{6 + 6.95}{2} = \frac{12.95}{2} = 6.475$$

Note: The mean, median, midrange, and midquartile are all measures of central tendency. They are *not* necessarily equal. Can you think of an example when they would be the same value?

5-Number Summary

5-Number Summary: The 5-number summary is composed of:

1. L, the smallest value in the data set
2. Q_1, the first quartile (also P_{25})
3. $\tilde{x}$, the median (also P_{50} and 2nd quartile)
4. Q_3, the third quartile (also P_{75})
5. H, the largest value in the data set

Notes:

- The 5-number summary indicates how much the data is spread out in each quarter
- The **interquartile range** is the difference between the first and third quartiles. It is the range of the middle 50% of the data

Box-and-Whisker Display

Box-and-Whisker Display: A graphic representation of the 5-number summary:

- The five numerical values (smallest, first quartile, median, third quartile, and largest) are located on a scale, either vertical or horizontal
- The box is used to depict the middle half of the data that lies between the two quartiles
- The whiskers are line segments used to depict the other half of the data
- One line segment represents the quarter of the data that is smaller in value than the first quartile
- The second line segment represents the quarter of the data that is larger in value that the third quartile

Box Plot:

- **A box plot is a graphical representation of a five-number summary.**

Steps for creating a box plot:

1. **Begin with a horizontal (or vertical) number line.**
2. **Draw a small line segment above (or next to) the number line to represent each of the numbers in the five-number summary.**
3. **Connect the line segment that represents the first quartile to the line segment representing the third quartile, forming a box with the median's line segment in the middle.**
4. **Connect the "box" to the line segments representing the minimum and maximum values to form the "whiskers".**

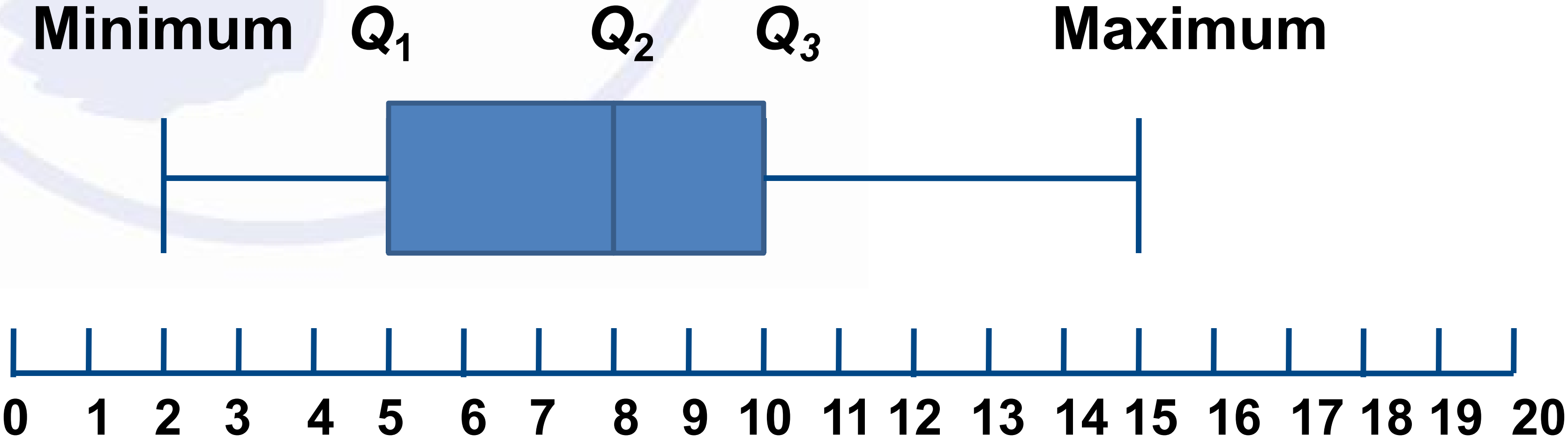
Draw a box plot for the given sample data:

8 9 10 2 5 3 7 12 15

Solution:

First order the data.

2 3 5 7 8 9 10 12 15



Example

✓ **Example:** A random sample of students in a sixth grade class was selected. Their weights are given in the table below. Find the 5-number summary for this data and construct a boxplot:

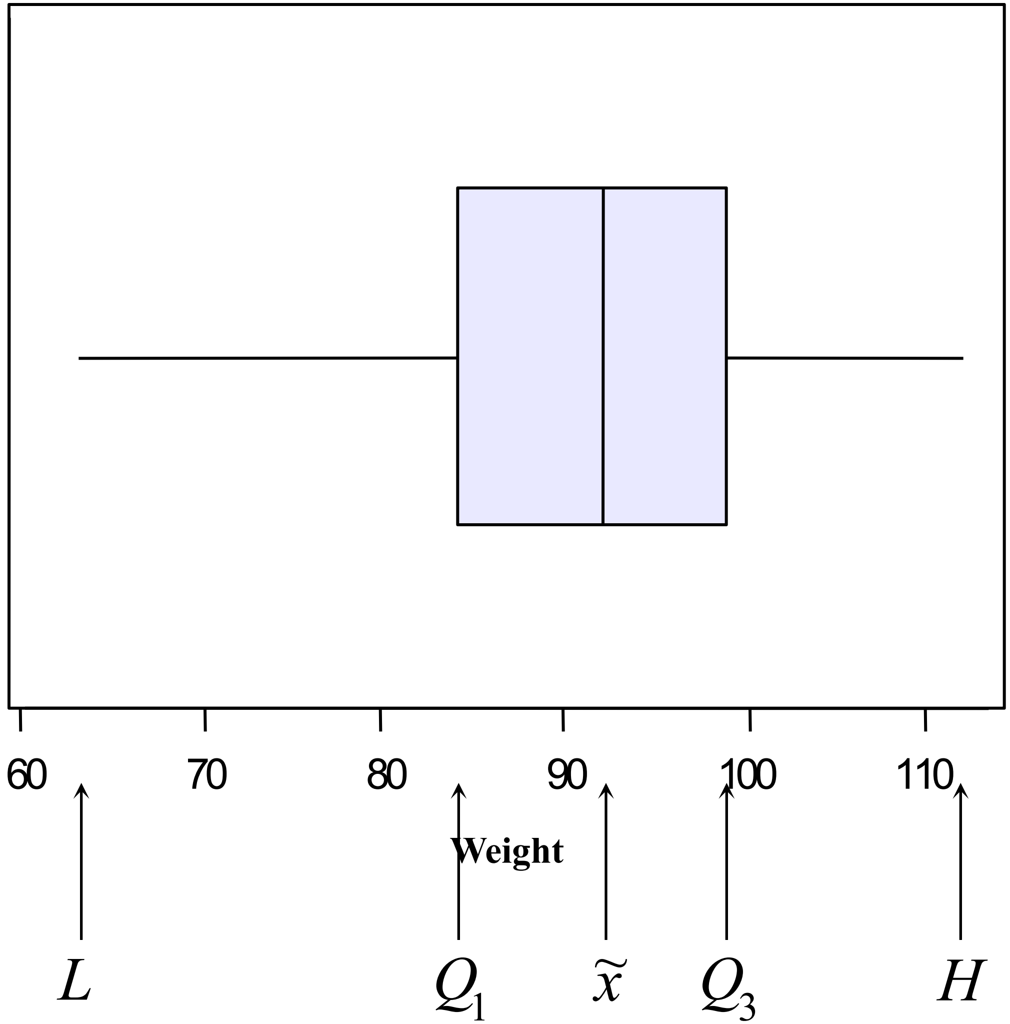
63	64	76	76	81	83
85	86	88	89	90	91
92	93	93	93	94	97
99	99	99	101	108	109
112					

Solution:

63	85	92	99	112
L	Q_1	$\tilde{x}$	Q_3	H

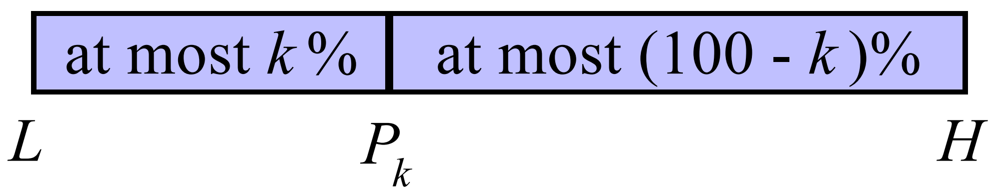
Boxplot for Weight Data

Weights from Sixth Grade Class



Percentiles

Percentiles: Values of the variable that divide a set of ranked data into 100 equal subsets; each set of data has 99 percentiles. The kth percentile, P_k, is a value such that at most k% of the data is smaller in value than P_k and at most $(100 - k)$% of the data is larger.



Notes:

- The 1st quartile and the 25th percentile are the same: $Q_1 = P_{25}$
- The median, the 2nd quartile, and the 50th percentile are all the same: $\tilde{x} = Q_2 = P_{50}$

Percentiles:

- **Percentiles divide the data into 100 equal parts.**
- **At the n^{th} percentile, n% of the data lies at or below a given value.**
- **Formula:**

$$l = n\frac{p}{100}$$

where l = location of the data value

p = percentile as a whole number

n = sample size

Percentiles (continued):

- **When using this formula to find the location of the percentile's value in the data set you must make sure to follow these two rules:**
  1. **If the formula results in a decimal value for *l*, the location is the next largest integer.**
  2. **If the formula results in a whole number, the percentile's value is the average of the value in that location and the one in the next largest location.**

When calculating the percentile, always round up to the next integer.

What data value lies at the 30th percentile?

11	**11**	**14**	**15**	**16**
16	**17**	**19**	**22**	**25**
26	**27**	**31**	**34**	**36**

Solution:

First order the data.

11 11 14 15 16 16 17 19 22 25 26 27 31 34 36

The sample size is $n = 15$.

The 30th percentile means $p = 30$.

$$l = 15\frac{30}{100} = 4.5$$

Since l = 4.5 we will round up to 5 and the value in the 5th position is 16.

z-Score

z-Score: The position a particular value of x has relative to the mean, measured in standard deviations. The z-score is found by the formula:

$$z = \frac{\text{value} - \text{mean}}{\text{st.dev.}} = \frac{x - \bar{x}}{s}$$

Notes:

- Typically, the calculated value of z is rounded to the nearest hundredth
- The z-score measures the number of standard deviations above/below, or away from, the mean
- z-scores typically range from -3.00 to +3.00
- z-scores may be used to make comparisons of raw scores

Standard Scores:

- **Standard scores, or *z*-scores, tell a data value's position in relation to the mean of the set.**
- **Formula:**

$$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad z = \frac{x - \bar{x}}{s}$$

$\mu =$ population mean
$\bar{x} =$ sample mean
$\sigma =$ population standard deviation
$s =$ sample standard deviation

Find the Standard Score:

Suppose that the mean on test 1 was 80.1 with a standard deviation of 6.3 points. If a student made a 92.5, what is the student's standard score?

Solution:

$$\mu = 80.1$$

$$\sigma = 6.3$$

$$x = 92.5$$

$$z = \frac{92.5 - 80.1}{6.3}$$

$$\approx 1.97$$

When calculating the standard score, always round to two decimal places.

Who did better on their exam with respect to their class?

Student A scored an 87 **Student B scored an 82**

$\mu = 80$ $\mu = 73$

$\sigma = 5$ $\sigma = 6$

Solution:

$$z = \frac{87 - 80}{5}$$
$$= 1.40$$

$$z = \frac{82 - 73}{6}$$
$$= 1.50$$

Since Student B’s score was more standard deviations above the mean, Student B did better with respect to their class.

Why *z*-Scores?

- **Transforming raw scores to *z*-scores facilitates making comparisons, especially when using different scales.**
- **A *z*-score provides information about the relative position of a score in relation to other scores in a sample or population.**
  - **A raw score provides no information regarding the relative standing of the score relative to other scores.**
  - **A *z*-score tells one how many standard deviations the score is from the mean. It also provides the approximate percentile rank of the score relative to other scores. For example, a *z*-score of 1 is 1 standard deviation above the mean and equals the 84.1 percentile rank (50% of occurrences fall below the mean and 34.1% of the occurrences fall between 0 and 1; 50% + 34.1% = 84.1%).**

Example

- ✓ **Example:** A certain data set has mean 35.6 and standard deviation 7.1. Find the *z*-scores for 46 and 33:

Solutions:

$$z=\frac{x-\overline{x}}{s}=\frac{46-35.6}{7.1}=1.46$$

46 is 1.46 standard deviations *above* the mean

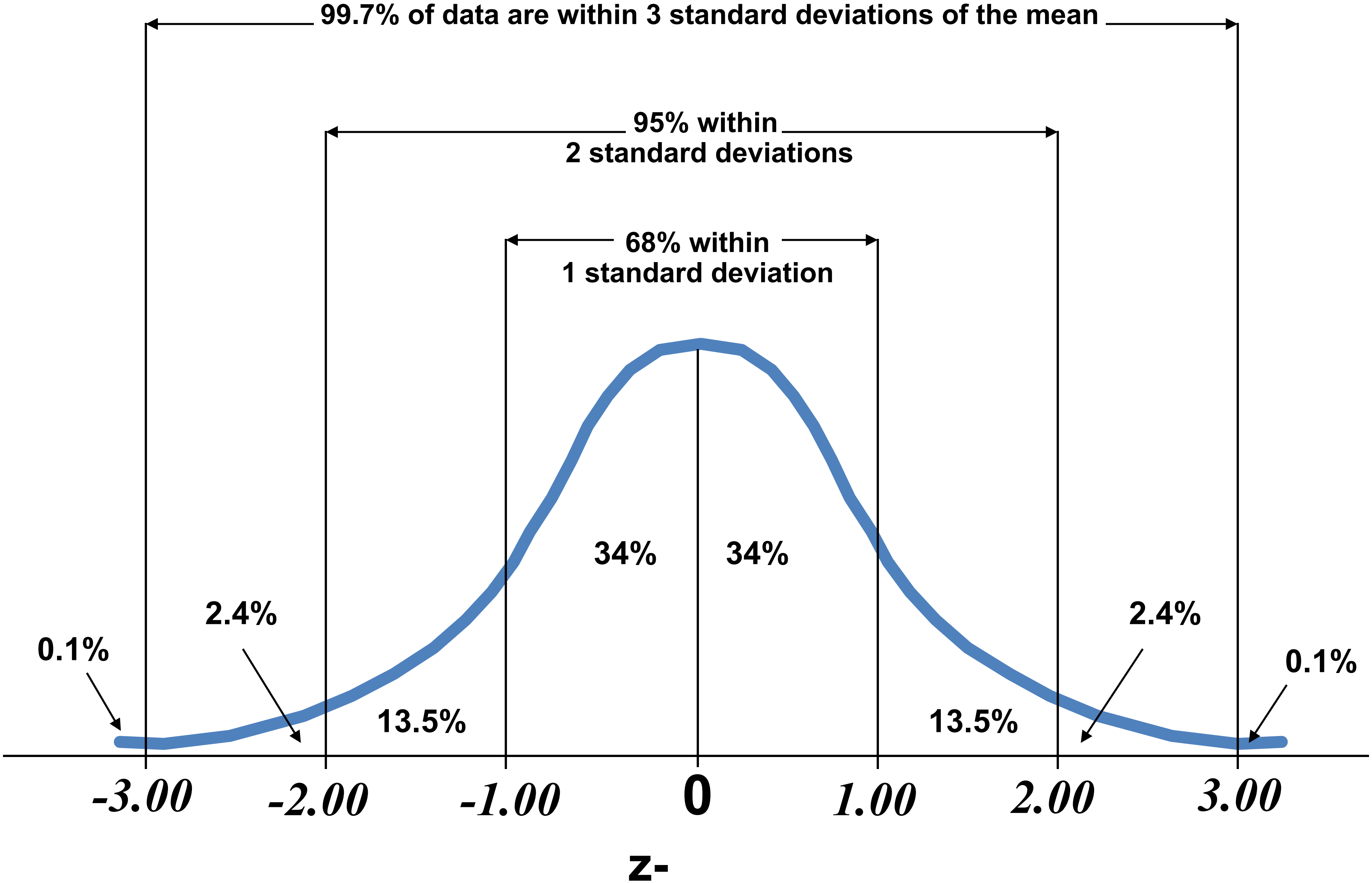
$$z=\frac{x-\overline{x}}{s}=\frac{33-35.6}{7.1}= \quad 0.37$$

33 is 0.37 standard deviations *below* the mean.

Interpretation of z-Scores

- If $z = 0$ an observation is at the mean.
- If $z > 0$ the observation is above the mean in value, e.g. if $z = 2.00$ the observation is 2 SDs above the mean.
- If $z < 0$ the observation is below the mean in value, e.g. if $z = -1.00$ the observation is 1 SD below the mean.

The Empirical Rule (z-scores)



The Empirical Rule (z-scores)

Therefore for normally distributed data:

- 68% of observations have z-scores between

  -1.00 and 1.00

- 95% of observations have z-scores between

  -2.00 and 2.00

- 99.7 of observations have z-scores between

  -3.00 and 3.00

Outliers based on z-scores

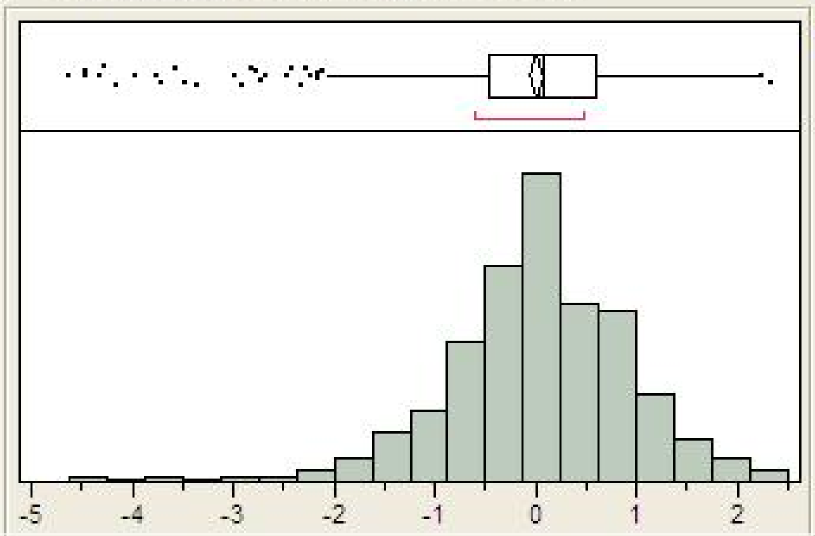
- When we consider the empirical rule an observation with a
  z-score < -2.00 or z-score > 2.00
  might be characterized as a mild outlier.

- Any observation with a
  z-score < - 3.00 or z-score > 3.00
  might be characterized as an extreme outlier.

Standardized Variables

We can convert each observed value of a numeric variable to its associated z-score. This process is called **standardization** and the resulting variable is called the **standardized variable**.

Standardized Birthweight (z-scores)



Quantiles

100.0%	maximum	2.335
99.5%		2.202
97.5%		1.847
90.0%		1.181
75.0%	quartile	0.604
50.0%	median	0.072
25.0%	quartile	-0.460
10.0%		-1.082
2.5%		-2.410
0.5%		-4.320
0.0%	minimum	-4.631

Moments

Mean	7.417e-16
Std Dev	1
Std Err Mean	0.0353553
upper 95% Mean	0.0694003
lower 95% Mean	-0.0694
N	800

Note: When standardized the mean is 0 and standard deviation is 1!

Interpreting & Understanding Standard Deviation

- Standard deviation is a measure of variability, or spread

- Two rules for describing data rely on the standard deviation:

  - ***Empirical rule***: **applies to a variable that is normally distributed**

  - ***Chebyshev's theorem***: **applies to any distribution**

Empirical Rule

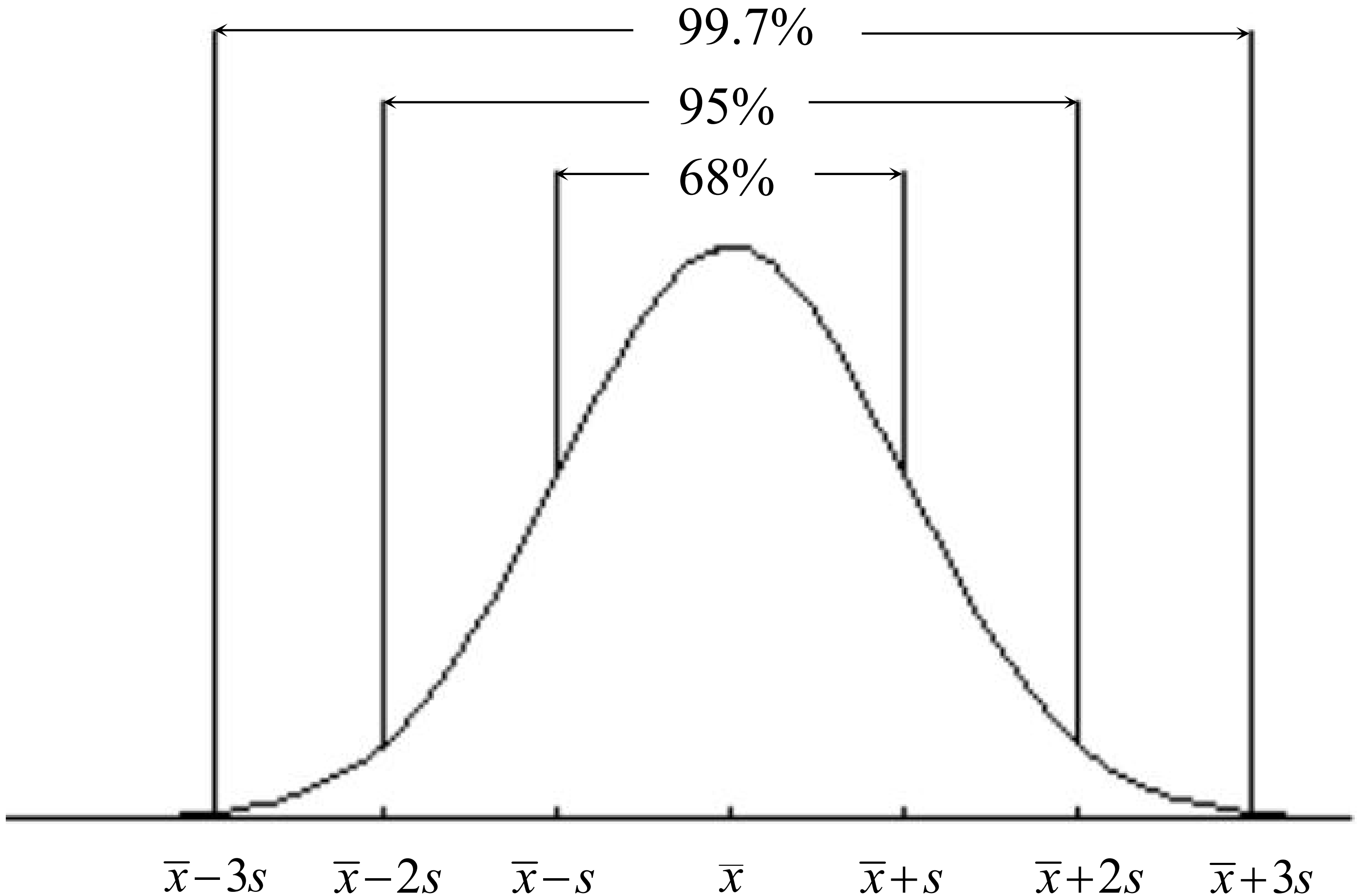
Empirical Rule: If a variable is normally distributed, then:

1. Approximately 68% of the observations lie within 1 standard deviation of the mean
2. Approximately 95% of the observations lie within 2 standard deviations of the mean
3. Approximately 99.7% of the observations lie within 3 standard deviations of the mean

Notes:

- The empirical rule is more informative than Chebyshev's theorem since we know more about the distribution (normally distributed)
- Also applies to populations
- Can be used to determine if a distribution is normally distributed

Illustration of the Empirical Rule



Example

- ✓ **Example:** A random sample of plum tomatoes was selected from a local grocery store and their weights recorded. The mean weight was 6.5 ounces with a standard deviation of 0.4 ounces. If the weights are normally distributed:

1) What percentage of weights fall between 5.7 and 7.3?
2) What percentage of weights fall above 7.7?

Solutions:

1) $(\bar{x}-2s,\ \bar{x}+2s)=(6.5-2(0.4),\ 6.5+2(0.4))=(5.7,\ 7.3)$
   Approximately 95% of the weights fall between 5.7 and 7.3

2) $(\bar{x}-3s,\ \bar{x}+3s)=(6.5-3(0.4),\ 6.5+3(0.4))=(5.3,\ 7.7)$
   Approximately 99.7% of the weights fall between 5.3 and 7.7
   Approximately 0.3% of the weights fall outside (5.3, 7.7)
   Approximately (0.3/2)=0.15% of the weights fall above 7.7

A Note about the Empirical Rule

Note: The empirical rule may be used to determine whether or not a set of data is approximately normally distributed

1. Find the mean and standard deviation for the data

2. Compute the actual proportion of data within 1, 2, and 3 standard deviations from the mean

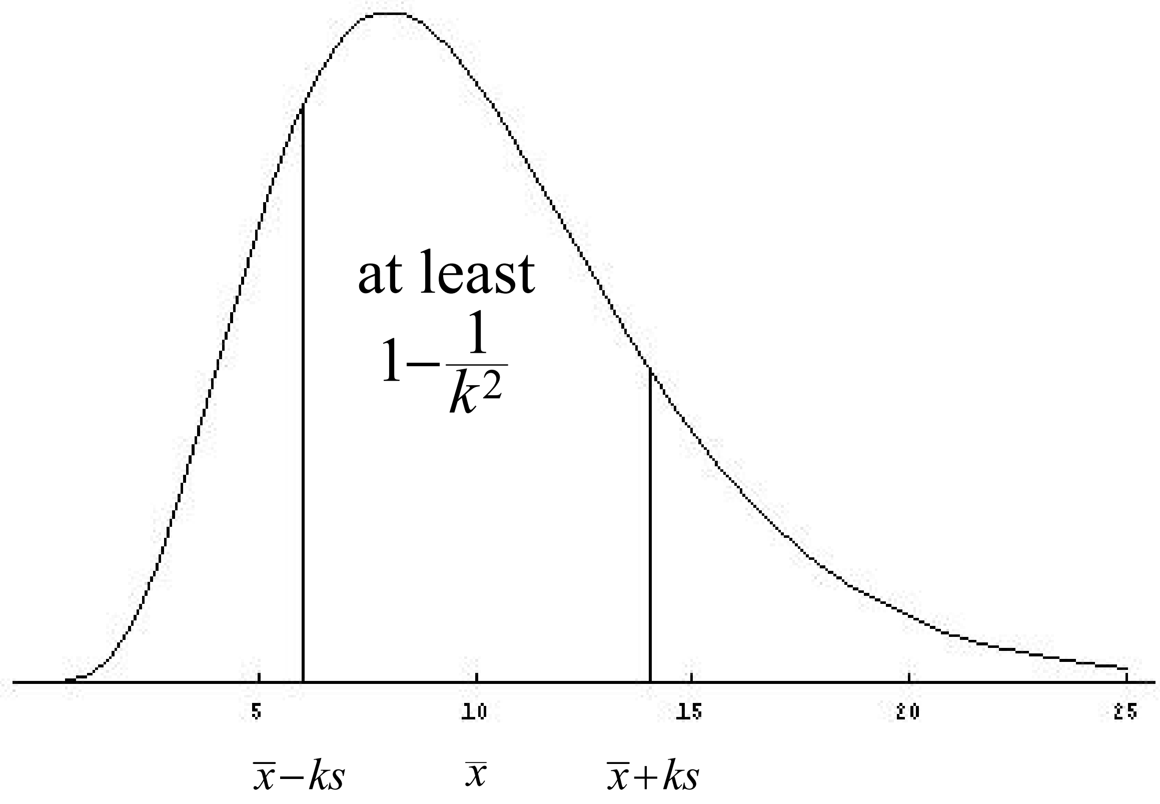
3. Compare these actual proportions with those given by the empirical rule

4. If the proportions found are reasonably close to those of the empirical rule, then the data is approximately normally distributed

Chebyshev's Theorem

Chebyshev's Theorem: The proportion of any distribution that lies within k standard deviations of the mean is at least $1 - (1/k^2)$, where k is any positive number larger than 1. This theorem applies to all distributions of data.

Illustration:



Important Reminders!

- ➔ Chebyshev's theorem is very conservative and holds for any distribution of data
- ➔ Chebyshev's theorem also applies to any population
- ➔ The two most common values used to describe a distribution of data are $k = 2, 3$
- ➔ The table below lists some values for k and $1 - (1/k^2)$:

k	1.7	2	2.5	3
$1-(1/k^2)$	0.65	0.75	0.84	0.89

Example

- ✓ **Example:** At the close of trading, a random sample of 35 technology stocks was selected. The mean selling price was 67.75 and the standard deviation was 12.3. Use Chebyshev's theorem (with $k = 2, 3$) to describe the distribution.

Solutions:

Using k=2: At least 75% of the observations lie within 2 standard deviations of the mean:

$$(\bar{x} - 2s, \bar{x} + 2s) = (67.75 - 2(12.3), 67.75 + 2(12.3) = (43.15, 92.35)$$

Using k=3: At least 89% of the observations lie within 3 standard deviations of the mean:

$$(\bar{x} - 3s, \bar{x} + 3s) = (67.75 - 3(12.3), 67.75 + 3(12.3) = (30.85, 104.65)$$

The Art of Statistical Deception

Good Arithmetic, Bad Statistics

Misleading Graphs

Insufficient Information

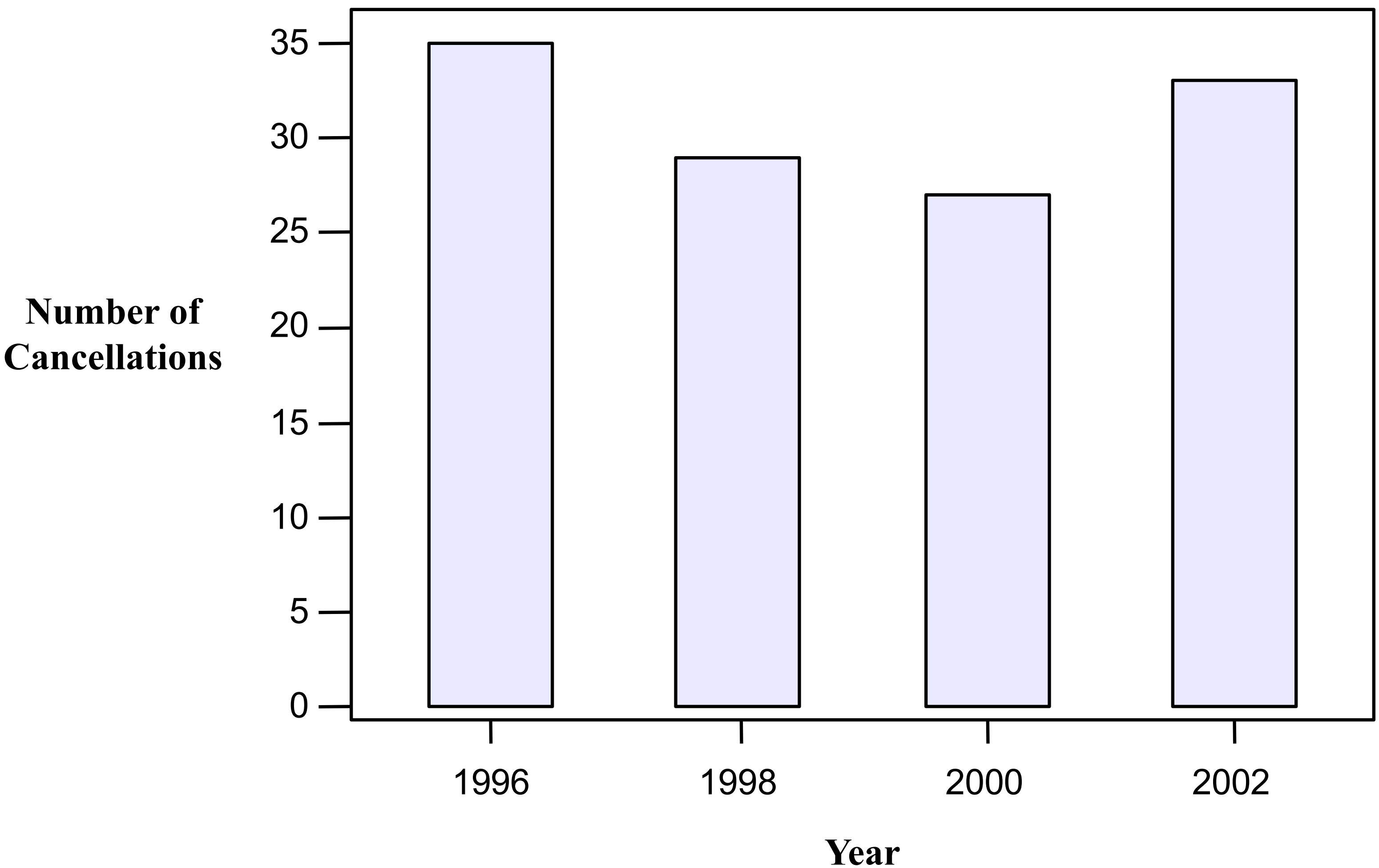
Good Arithmetic, Bad Statistics

- The mean can be greatly influenced by outliers
  - *Example*: The mean salary for all NBA players is $15.5 million

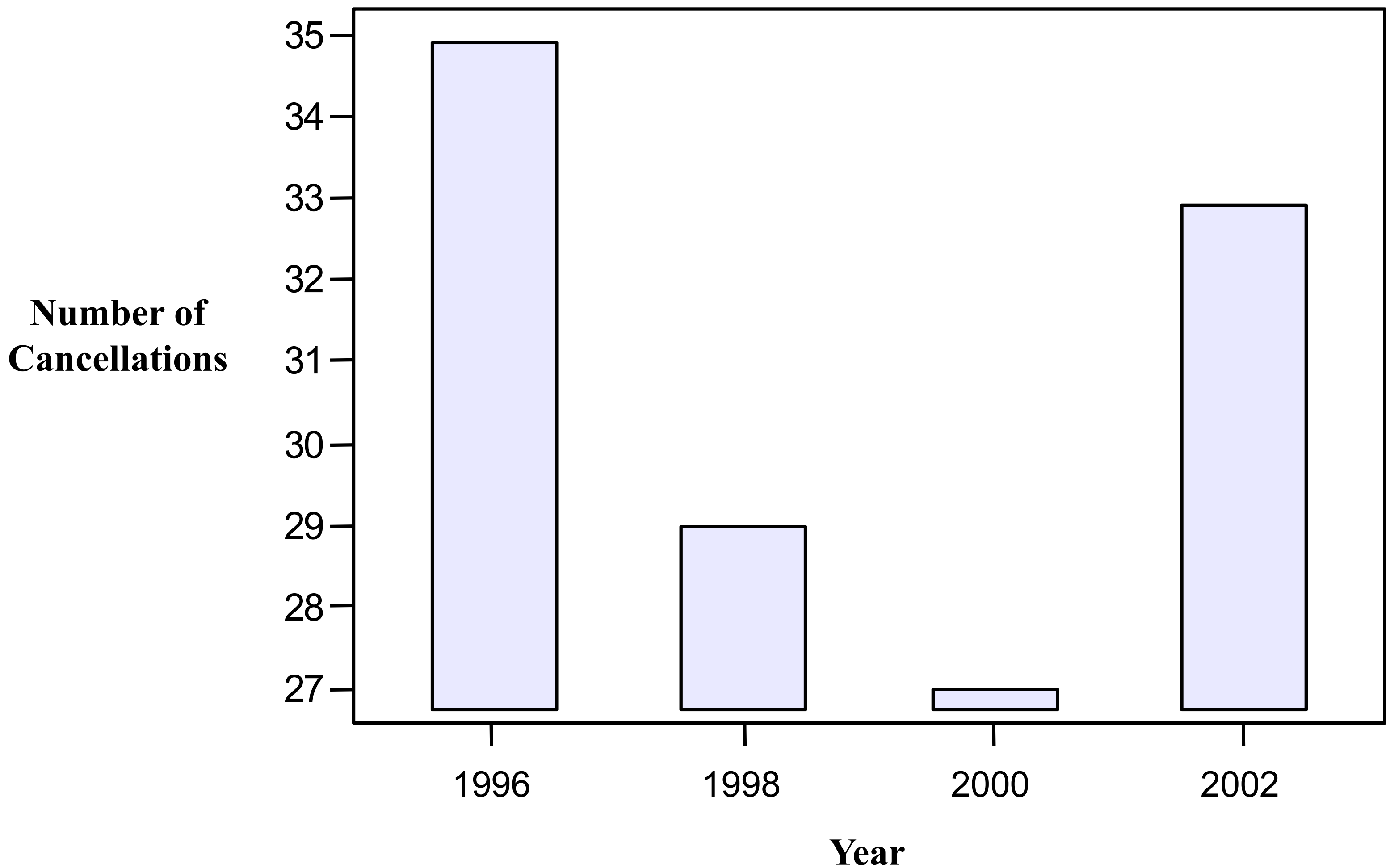
Misleading graphs:

1. **The frequency scale should start at zero to present a complete picture. Graphs that do not start at zero are used to save space.**
2. **Graphs that start at zero emphasize the size of the numbers involved**
3. **Graphs that are chopped off emphasize variation**

Flight Cancellations



Flight Cancellations



Insufficient Information

- *Example*: An admissions officer from a state school explains that the average tuition at a nearby private university is $13,000 and only $4500 at his school. This makes the state school look more attractive.
  - **If most students pay the full tuition, then the state school appears to be a better choice**
  - **However, if most students at the private university receive substantial financial aid, then the actual tuition cost could be much lower!**