2.3 Measures of the Location of the Data

Percentile

A measure of position, the *percentile*, p, is an integer $(1 \le p \le 99)$ such that the p^{th} percentile is the position of a data value where

> p% of the data values in the distribution are less than or equal to the value and
> $100 - p$% of the data values in the distribution are greater than or equal to the value.

We denote the 1st percentile, P_1, the 2nd percentile, P_2, …, the highest percentile, P_{99}.

To find the data value in the p^{th} percentile in an ordered list we use the formula

$$p = \frac{x + 0.5y}{n} \bullet 100$$

x = the number of data values less than the data value for which you wish to find the percentile
y = the number of data values equal to the data value for which you wish to find the percentile
n = the total number of data values

We can reverse the formula to find the position of the data value at a given percentile. For this we use the formula

$$i = \frac{k(n+1)}{100}$$

i is used to locate the position of the value in an ordered list
n is the number of data values in the set
k is the given percentile

- If i is a whole number, count to the data value in the i^{th} position starting with the smallest data value.
- If i is not a whole number, average the data values just above and just below the i^{th} position.

Example 1: Consider the weights of ten newborn babies born on a given day at St. Elizabeth Hospital:

5.9 6.2 6.3 7.0 7.2 7.7 7.8 7.9 8.8 10.5

(a) Let's determine the percentile of the baby weight **7.8** relative to the others in this list.

5.9	6.2	6.3	7.0	7.2	7.7	**7.8**	7.9	8.8	10.5
1	2	3	4	5	6	7	8	9	10

There are $x = 6$ values below 7.8, 7.8 occurs $y = 1$ time, and $n = 10$ total values.

Thus, we have $P = \frac{x + 0.5y}{n} \bullet 100 = \frac{6 + 0.5 \bullet 1}{10} \bullet 100 = \frac{6 + 0.5}{10} \bullet 100 = 65$,

that is, **7.8** is at the **65**th percentile of this list.

Notice that

65% of the data values are less than or equal to **7.8**

35% of the data values are greater than or equal to **7.8**

5.9	6.2	6.3	7.0	7.2	7.7	**7.8**	7.9	8.8	10.5
1	2	3	4	5	6	7	8	9	10

(b) Suppose we want to find the value that lies in the **45**$^{\text{th}}$ percentile.

With $n = 10$ and $k = 45$,

we have

$$i = \frac{k(n+1)}{100} = \frac{45(10+1)}{100} = 4.95$$

5.9	6.2	6.3	7.0	**7.2**	7.7	7.8	7.9	8.8	10.5
1	2	3	4	5	6	7	8	9	10

Find the average of data values in the

4$^{\text{th}}$ and 5$^{\text{th}}$ positions $\frac{7.0+7.2}{2} = 7.1$

Thus, **7.1** is at the 45$^{\text{th}}$ percentile in our list.

(c) Suppose our data is the number of siblings of each of 9 surveyed individuals and we want to find the response value that lies in the **60**$^{\text{th}}$ percentile.

Using $n = 9$ and $k = 60$, we have

$$i = \frac{k(n+1)}{100} = \frac{60(9+1)}{100} = 6$$

0	1	2	3	**3**	3	4	7	18
1	2	3	4	5	**6**	7	8	9

Since i is a whole number, we choose the data value in the **6**$^{\text{th}}$ position.

Thus, the data value, **3** siblings, is at the 60$^{\text{th}}$ percentile in our list.

Quartiles

When percentiles give too much detailed information for analysis purposes, we can measure the position of data elements by quartiles. A *quartile* is the value of the boundary at the 25th, 50th, or 75th percentiles of a frequency distribution, dividing it into four equal parts, denoted Q_1, Q_2, Q_3, and the maximum data value.
We also call Q_2, the median of the whole data set, Q_1, is the median of the set of data values less than the median, and Q_3, is the median of the set of data values greater than the median The formula $i = \dfrac{k(n+1)}{100}$ can be used to find the median. When the median, Q_2, is known, Q_1 and Q_3 can be found using the same formula on the smaller half-sets.

Example 2:

(a) Let's find Q_1, Q_2, and Q_3 in the list of weights of newborns.

5.9	6.2	$\boxed{6.3}$	7.0	7.2		7.7	7.8	$\boxed{7.9}$	8.8	10.5
					$\boxed{7.45}$					
1	2	$\boxed{3}$	4	5		6	7	$\boxed{8}$	9	10

First, we'll find the median, Q_2, using $k = 50$ and $n = 10$,

$$Q_2 = P_{50}:\ i = \frac{k(n+1)}{100} = \frac{50 \bullet 11}{100} = 5.11 \qquad Q_2 = \frac{7.2 + 7.7}{2} = 7.45$$

Then, we find the median of the lower half using $k = 50$ and $n = 5$,

$$Q_1:\ i = \frac{k(n+1)}{100} = \frac{50(5+1)}{100} = \frac{50 \bullet 6}{100} = 3 \qquad Q_1 = 6.3$$

And finally, we find the median of the upper half using $k = 50$ and $n = 5$,

$$Q_3:\ i = \frac{k(n+1)}{100} = \frac{50(5+1)}{100} = \frac{50 \bullet 6}{100} = 3 \ \text{(Use 8)} \qquad Q_3 = 7.9$$

The *interquartile range*, *IQR*, is the difference between the 3rd and 1st quartiles.

$$IQR = Q_3 - Q_1$$

(b) Find the interquartile range of the weights of the newborn babies.

5.9	6.2	6.3	7.0	7.2	7.7	7.8	7.9	8.8	10.5
1	2	3	4	5	6	7	8	9	10
		Q_1					Q_3		

$IQR = 7.9 - 6.3 = \mathbf{1.6}$

The interquartile range of this data set is **1.6**.

Outliers

An *outlier* is a data value that lies outside the overall pattern of a distribution.

An extremely large data value or extremely small data value can affect the measures of mean and standard deviation. For this reason, the mean and standard deviation are called *nonresistant* statistics. The median and interquartile range are less affected by outliers, so they are called *resistant* statistics.

If a data value is suspiciously large or small, we check to see if the value lies within

$$[Q_1 - 1.5(IQR),\ Q_3 + 1.5(IQR)]$$

If the data value lies outside this interval, it is considered an outlier.

Example 3:

Determine if the data value \$1.17 million is an outlier in this list of home values:

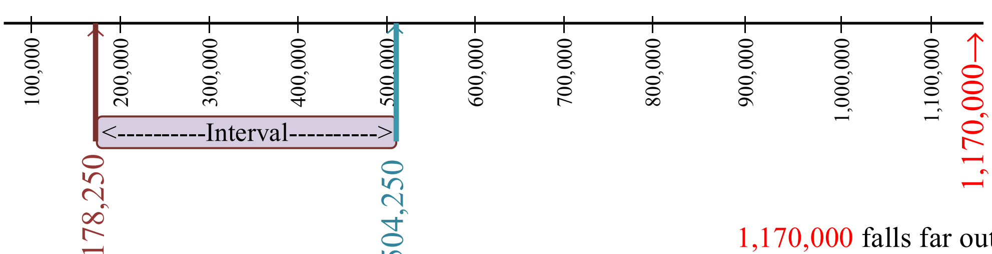
\$272,000	\$303,000	\$341,000	\$384,000
\$272,000	\$304,000	\$346,000	\$404,000
\$275,000	\$305,000	\$348,000	\$434,000
\$277,000	\$328,000	\$351,000	\$738,000
\$297,000	\$337,000	\$359,000	\$912,000
\$298,000	\$339,000	\$380,000	\$1,170,000

$$Q_1 = \frac{298{,}000 + 303{,}00}{2} = \frac{601{,}000}{2} = 300{,}500$$

$$Q_3 = \frac{380{,}000 + 384{,}00}{2} = \frac{764{,}000}{2} = 382{,}000$$

$$IQR = 382{,}000 - 300{,}500 = 81{,}500$$

$$\begin{aligned}
&[Q_1 - 1.5(IQR),\ Q_3 + 1.5(IQR)] \\
&= [300{,}500 - 1.5(81{,}500),\ 382{,}000 + 1.5(81{,}500)] \\
&= [300{,}500 - 122{,}250,\ 382{,}000 + 122{,}250] \\
&= [178{,}250,\ 504{,}250]
\end{aligned}$$



1,170,000 falls far outside the interval shown at left and is, thus, an outlier.

Interpreting Percentiles from a Frequency Distribution

In an ungrouped frequency distribution, we can find specific percentiles within the set of data by glancing at the cumulative frequency column.

Example 4:

Let's collect some data from this class: How many dogs do you have?

Number of dogs	Frequency	Relative frequency	Cumulative relative frequency
0			
1			
2			
3			
4			
5 or more			

Using the cumulative frequency column, answer the following questions.

1. Which number of dogs is in the 30^{th} percentile?

2. Which number of dogs is in the 80^{th} percentile?

3. An owner of 2 dogs is in which percentile in this distribution?