Frequency Distribution

A **frequency distribution** is a summary technique that organizes data into classes and provides in tabular form a list of the classes along with the number of observations in each class.

DEFINITION

Constructing a Frequency Distribution

1. Choose the classifications.
2. Count the number in each class.

PROCEDURE

The first question in the survey was: *In general, how would you rate the quality of American public schools?* The frequency of each response category is shown in the table on the left. The frequency distribution for the second question involving *a lack of parental involvement with a child's education* is given in the table on the right. The summary tables are much more informative than looking at 1250 observations for each question.

Frequency Distribution of Responses	
In general, how would you rate the quality of American public schools?	
Excellent	462
Pretty good	288
Only fair	225
Poor	225
Not sure	50

Frequency Distribution of Responses	
How serious is a lack of parental involvement with a child's education?	
Very serious	700
Somewhat serious	325
Not very serious	112
Not a problem	75
Not sure	38

Relative Frequency Distribution of Responses	
In general, how would you rate the quality of American public schools?	
Excellent	37%
Pretty good	23%
Only fair	18%
Poor	18%
Not sure	4%

Relative Frequency Distribution of Responses	
How serious is a lack of parental involvement with a child's education?	
Very serious	56%
Somewhat serious	26%
Not very serious	9%
Not a problem	6%
Not sure	3%

Bar Chart

The **bar chart** is a simple graphical display in which the length of each bar corresponds to the number of observations in a category.

DEFINITION

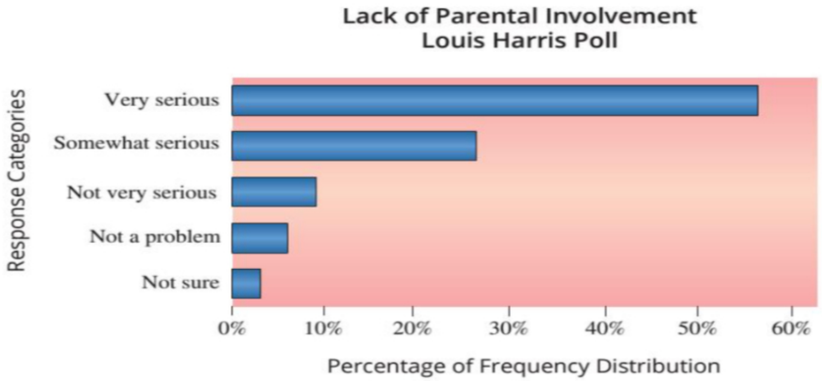


Table 3.2.1 - Sales Performance

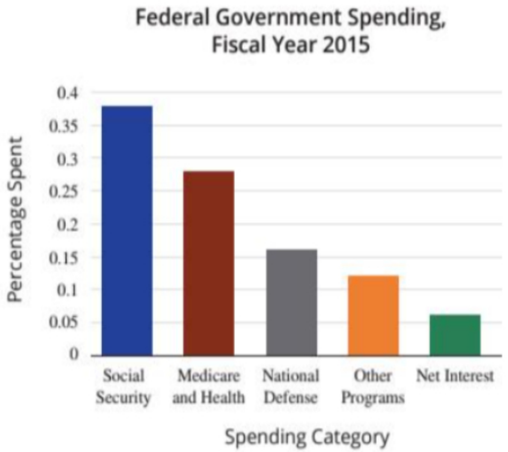
Salesperson	Total Sales (in thousands of dollars)
Susan	187
William	201
Beth	207
Rob	193

of a circle. The size of each slice is proportional to the amount or frequency in each category. The proportion of the total that each slice represents is often displayed as a percentage on the chart. These percentages should total 100%.

DEFINITION

Table 3.2.2 - Percentage Spent by the Federal Government in 2015

Category	Percentage Spent
Social Security	38%
Medicare and Health	
National Defense	
Other Programs	
Net Interest	



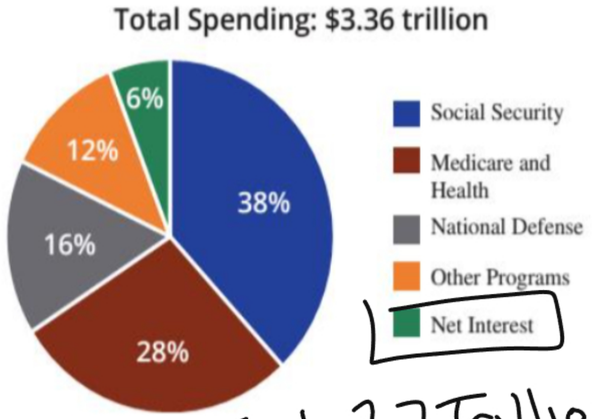
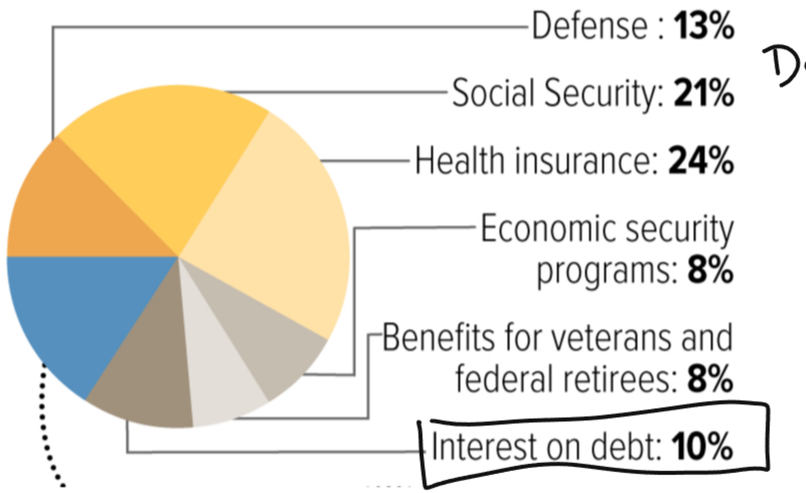


Figure 3.2.9

SS 1.27 Trillion

Defence .53 Trillion

Most of the Budget Goes Toward Defense, Social Security, and Major Health Programs

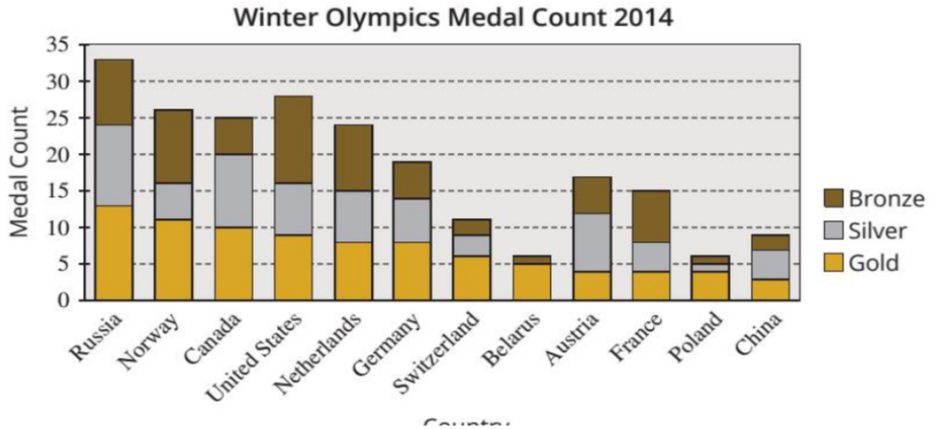


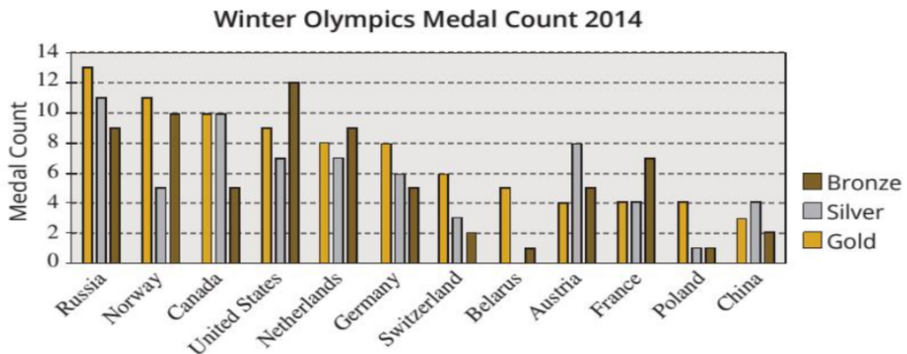
S.S. → 1.3 Trillion

Defence → .82 Trillon

Stacked Bar Charts

Stacked bar charts are an interesting variation on the standard bar chart. The number of medals (gold, silver, and bronze) won during the 2014 Winter Olympics for selected countries is given in Figure 3.2.7.





Heart Rates (per min.) of 50 Students

77	84	79	90	67	84	82	74	88	75
69	81	94	68	65	86	78	79	79	70
83	83	84	82	93	80	81	80	87	80
62	98	77	83	82	80	82	73	85	77
77	79	81	70	72	85	84	80	74	83

Table 3.3.1 - Frequency Distribution of Heart Rates

Heart Rate	Number of Students
57–66	2
67–76	10
77–86	32
87–96	5
97–106	1

2. **Determine the class width.** In some cases, the data set easily lends itself to natural divisions, such as decades or years. At other times, we must choose divisions for ourselves. You will want to choose a width so that the classes formed present a clear representation of the data and include all values in the data set. The width of each class should be the same whenever possible; exceptions may occur for the beginning and ending intervals. There is really no perfect formula for class width that will work for every data set. However, a good starting point for class width is to divide the difference between the maximum observation and minimum observation by the number of classes.

$$\text{Class Width} = \frac{\text{Maximum Value} - \text{Minimum Value}}{\text{Number of Classes}}$$

Class endpoints with fractional values will make the graph harder to understand. If possible, try to keep the width to an integer value by rounding the class width up to the next largest integer or choosing an integer value close to the calculated class width that makes sense.

Continued...

Constructing a Frequency Distribution (continued)

3. **Find the class limits.** The **lower class limit** is the smallest number th
can belong to a particular class, and the **upper class limit** is the larg
number that can belong to a class. Using the minimum data value, o
smaller number, as the lower limit of the first class is a good place to sta
However, judgment is required. You should choose the first lower limit
that reasonable classes will be produced. After choosing the lower li

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1 +79° night +55°
2 +81° night +55°	3 +86° night +64°	4 +75° night +57°	5 +79° night +59°	6 +72° night +68°	7 +81° night +68°	8 +75° night +61°
9 +63° night +59°	10 +73° night +55°	11 +72° night +55°	12 +77° night +57°	13 +81° night +57°	14 +79° night +66°	15 +77° night +66°
16 +75° night +57°	17 +79° night +54°	18 +86° night +66°	19 +88° night +68°	20 +90° night +68°	21 +93° night +72°	22 +82° night +70°
23 +82° night +68°	24 +79° night +72°	25 +86° night +66°	26 +84° night +70°	27 +79° night +68°	28 +73° night +57°	29 +72° night +57°
30 +79° night +72°						

63 - 68
69 - 74
75 - 80
81 - 86
87 - 92

$93 - 63 = \frac{30}{5} = 6$

62 - 68

Relative Frequency

The **relative frequency** of any class is the number of observations in the class divided by the total number of observations:

$$\text{Relative Frequency} = \frac{\text{Number in Class}}{\text{Total Number of Observations}}$$

DEFINITION

Table 3.3.2 - Heart Rate Relative Frequency Distribution

Heart Rate	Relative Frequency
57–66	2/50 = 0.04
67–76	10/50 = 0.20
77–86	32/50 = 0.64
87–96	5/50 = 0.10
97–106	1/50 = 0.02

Cumulative Frequency

The **cumulative frequency** is the sum of the frequency of a particular class and all preceding classes.

DEFINITION

Table 3.3.3 - Heart Rate Cumulative Frequency Distribution

Heart Rate	Frequency	Cumulative Frequency
57–66	2	2
67–76	10	12
77–86	32	44
87–96	5	49
97–106	1	50

In this example, the reader can easily see in Table 3.3.3 that 49 out of 50 heart rates are less than or equal to 96 beats per minute.

Cumulative Relative Frequency

The **cumulative relative frequency** is the proportion of observations in a particular class and all preceding classes.

DEFINITION

Table 3.3.4 - Heart Rate Cumulative Relative Frequency

Heart Rate	Relative Frequency	Cumulative Relative Frequency
57–66	0.04	0.04
67–76	0.20	0.24
77–86	0.64	0.88
87–96	0.10	0.98
97–106	0.02	1.00

Histogram

A **histogram** is a graphical representation of a frequency or relative frequency distribution. The horizontal scale corresponds to classes of quantitative data values and the vertical scale corresponds to the frequency or relative frequency of each class.

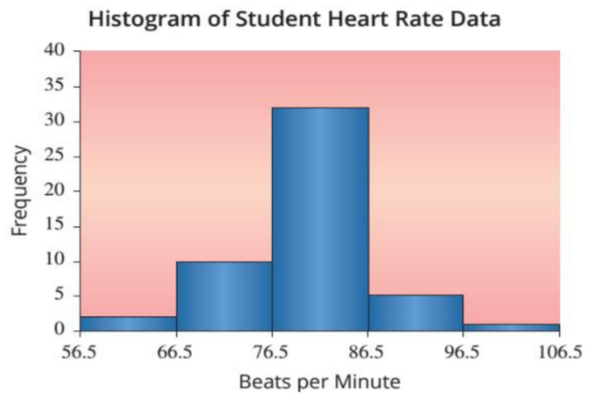


Table 3.3.1 - Frequency Distribution of Heart Rates

Heart Rate	Number of Students
57–66	2
67–76	10
77–86	32
87–96	5
97–106	1

Symmetric vs. Skewed

If you split the histogram of a distribution down the center, and the left and right sides of the histogram are approximately mirror images of one another, the distribution is said to be **symmetric**.

A **skewed distribution** is a nonsymmetric (or asymmetric) distribution that extends more to one side than the other. The distribution is said to be **skewed to the right** (or positively skewed) if the tail to the right of the peak of the distribution is longer than the tail to the left of the peak. The distribution is said to be **skewed to the left** (or negatively skewed) if the tail to the left of the peak of the distribution is longer than the tail to the right of the peak.

DEFINITION

When we look at a histogram, what features are important?

1. Is the distribution of the data symmetric or skewed?

Skewed Right

Skewed Left

Stem-and-Leaf Plot

The **stem-and-leaf plot** is a graph representing quantitative data that separates each data value into two parts: the stem and the leaf.

DEFINITION

Table 3.4.1 - Data, Stems, and Leaves

Data	Stem	Leaf
97	09	7
99	09	9
108	10	8
110	11	0
111	11	1

Stem-and-Leaf Plot

Stem	Leaf
09	7 9
10	8
11	0 1
Key: 10	8 = 108

Barry Bonds						
Year	1986	1987	1988	1989	1990	1991
HR	16	25	24	19	33	25
Year	1992	1993	1994	1995	1996	1997
HR	34	46	37	33	42	40
Year	1998	1999	2000	2001	2002	2003
HR	37	34	49	73	46	39
Year	2004	2005	2006	2007		
HR	45	5	26	28		

Babe Ruth						
Year	1914	1915	1916	1917	1918	1919
HR	0	4	3	2	11	29
Year	1920	1921	1922	1923	1924	1925
HR	54	59	35	41	46	25
Year	1926	1927	1928	1929	1930	1931
HR	47	60	54	46	49	46
Year	1932	1933	1934	1935		
HR	41	34	22	6		

Home runs Hit per Season: Babe Ruth vs Barry Bonds

Ruth		Bonds
10432	0	5
1	1	69
952	2	54568
54	3	3473749
1676961	4	620965
494	5	
0	6	
	7	3
Key: 0	6	1 = 60 HR for Ruth, 61 HR for Bonds

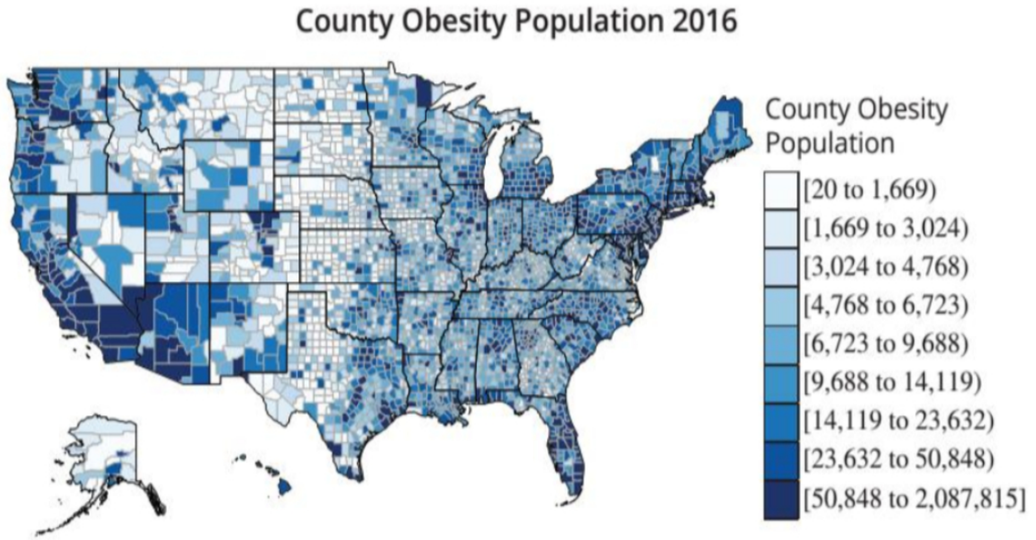


Figure 3.4.9

Example 4.1.1

Calculate the sample mean of the following sample data values: 4, 10, 7, 15. = 9

Average 4, 10, 7, 15, 123

31.8

Example 4.1.2

Meghan is a freshman in college and she received the following grades for her first semester.

Meghan's Grades		
Course	Grade	Credit Hours
Psychology 101	B 3	3 → 9
Probability and Statistics	A 4	4 → 16
Anatomy I	C 2	5 → 10
English 101	A 4	3 → 12
		15 → 47

$\frac{47}{15} = 3.13$

Example 4.1.5

Consider the following goal tallies from eleven games played by the Charleston Battery soccer team.

2, 3, 5, 4, 1, 7, 3, 3, 1, 2, 6

1, 1, 2, 2, 3, (3), 3, 4, 5, 6, 7

$\frac{11}{2} = 5.5$

Example 4.1.6

Consider the following ten test scores from a student taking a high school calculus class.

$\frac{10}{2} = 5$

65, 98, 76, 83, 94, 79, 88, 72, 90, 85 = 84

Find the median. 65, 72, 76, 79, 83, 85, 88, 90, 94, 98

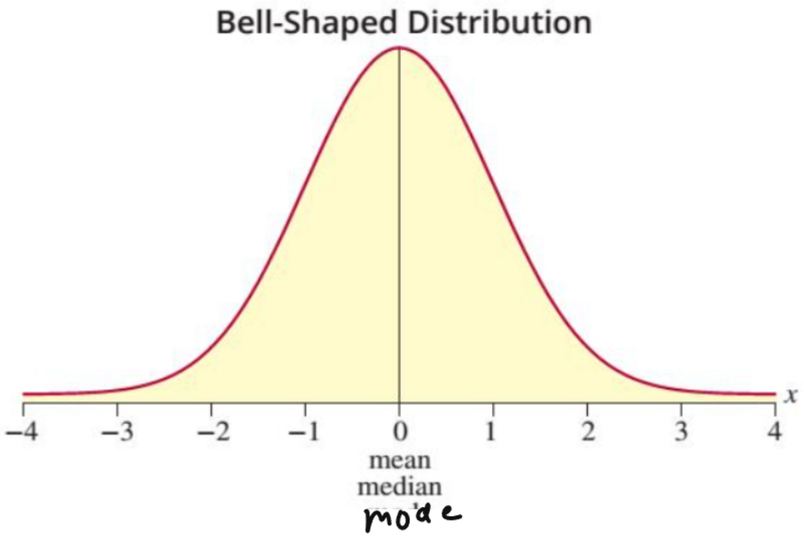
Example 4.1.7

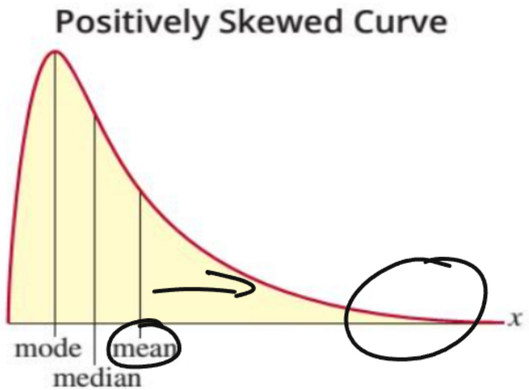
Find the mode of the following data regarding the number of power outages reported over a period of eleven days.

0, 1, 4, 3, 9, 8, 10, 0, 1, 3, 0 , 1, 3

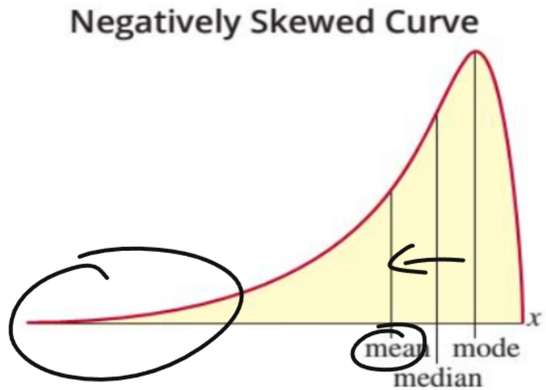
Table 4.1.3 – Applicable Level of Measurement

	Qualitative		Quantitative	
	Nominal	**Ordinal**	**Interval**	**Ratio**
Mean			✔	✔
Median		✔	✔	✔
Mode	✔	✔	✔	✔
Trimmed Mean			✔	✔





Right



Left

Daily Production										
Day	1	2	3	4	5	6	7	8	9	10
Units	100	104	117	20	20	111	105	106	115	101
Day	11	12	13	14	15	16	17	18	19	20
Units	101	102	115	116	113	103	104	119	118	108

mean = 99.9

median = 105.5

mode = 20, 101, 104, 115

A	B
10	2
15	5
16	16
19	27
20	30

mean =	16	16
median =	16	16
Range =	10	28
Stand Dev =	3.52	11.26

Range

The **range** is the difference between the largest and smallest data values.

DEFINITION

Variance

The **variance** of a data set containing the complete set of *population* data is given by

Population

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N},$$

μ = mean

σ = Standard Deviation

and is called the **population variance**.

The **variance** of a data set containing *sample* data is given by

Sample

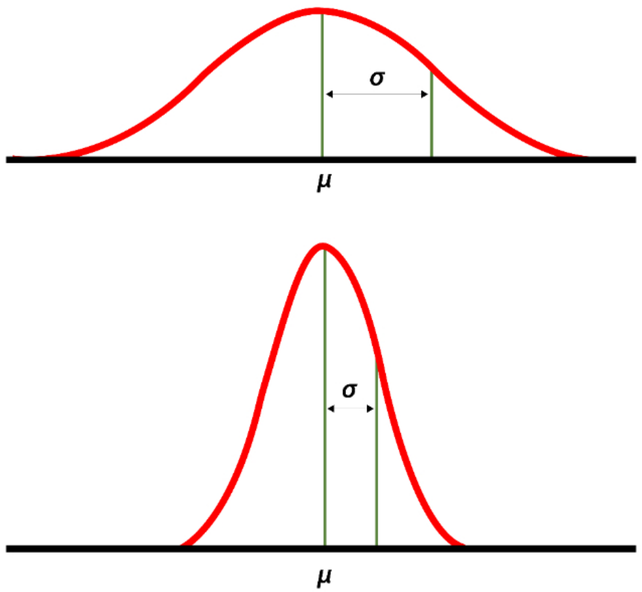
$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1},$$

$\bar{x}$ = mean

s = Standard Deviation

and is called the **sample variance**.

FORMULA



Properties of the Standard Deviation

- The standard deviation is always nonnegative. It is zero only if all the data values are exactly the same.
- The standard deviation can increase dramatically if there are one or more outliers in the data.
- The standard deviation is expressed in the same units as the original data values.

PROPERTIES

Empirical Rule

The empirical rule is so named because the given percentages are actually observed in practice. This rule doesn't apply to all distributions, only those that are symmetrical and bell-shaped.

Empirical Rule

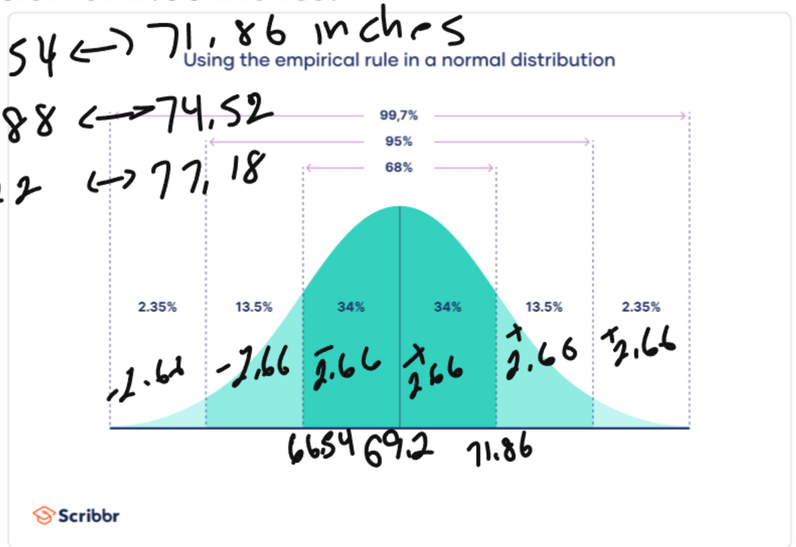
If the distribution of the data is bell-shaped, then

- About 68% of the data should lie within 1 standard deviation of the mean.
- About 95% of the data should lie within 2 standard deviations of the mean.
- About 99.7% of the data should lie within 3 standard deviations of the mean.

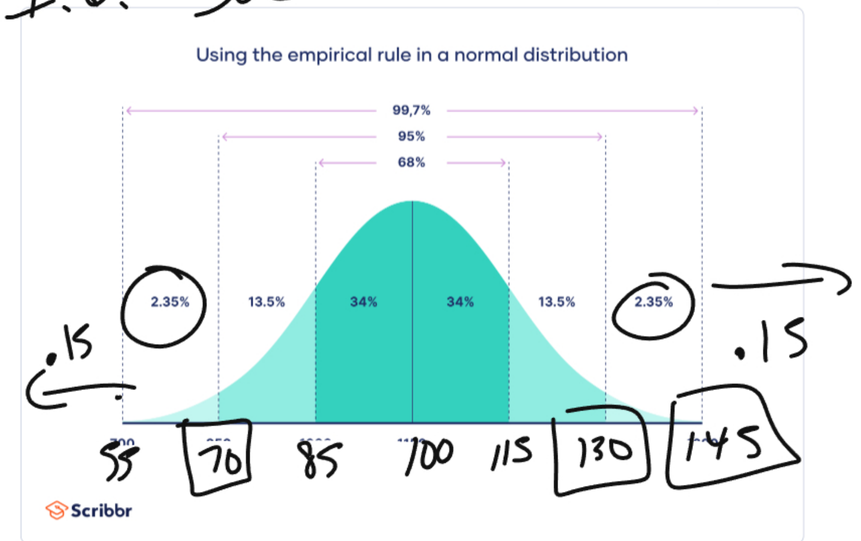
PROPERTIES

The heights of adult men in America are normally distributed, with a mean of 69.2 inches and a standard deviation of 2.66 inches.

68% → 66.54 ↔ 71.86 inches
95% → 63.88 ↔ 74.52
99.7 → 61.22 ↔ 77.18



I.Q. $\mu = 100$ $\sigma = 15$



The manager of a local diner has calculated his average daily sales to be $4500 with a standard deviation of $750.

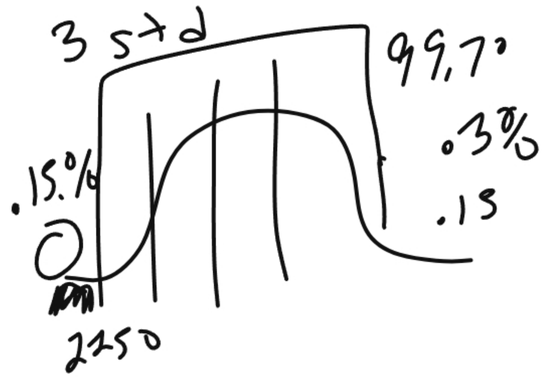
how many days bellow 2250

.15%

.0015

.55 days

4500
- 750
3750
- 750
3000
- 750
2250



To find the location of the P^{th} percentile in the ordered data, calculate,

$$\ell = n\left(\frac{P}{100}\right)$$

where n is the number of observations in the ordered data.

Is ℓ an integer?

No

Round ℓ up to next greatest integer.

Find the data value in the ℓ^{th} location.

Yes

Average the data value in the ℓ^{th} location with the data value in the $(\ell + 1)^{th}$ location

$500\left(\frac{30}{100}\right) = L$

$150 = L$

$L = 150$

$\frac{(150^{th} + 151^{st})}{2}$

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

$n = 40$

10^{th} percentile

$L = 40\left(\frac{10}{100}\right) = 4$

$\frac{27+29}{2} = 28$

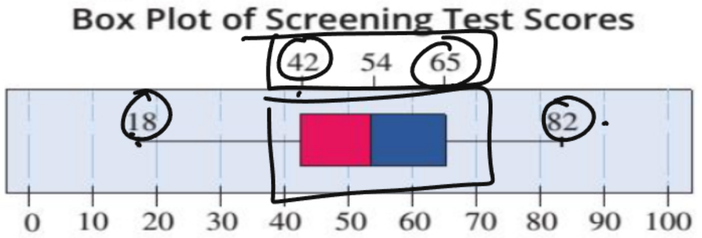
88^{th} percentile

$L = 40\left(\frac{88}{100}\right) = 35.2 \Rightarrow 36$

73

42 - 34.5

7.5



65 + 34.5

99.5

5 number summary

min

Q1 → 25%

median → 50%

Q3 → 75%

max

$L = 40\left(\frac{25}{100}\right) = 10$

$IQR = 65 - 42 = 23$

$\text{Outlier} = 1.5(IQR)$

$= 1.5(23)$

$= 34.5$