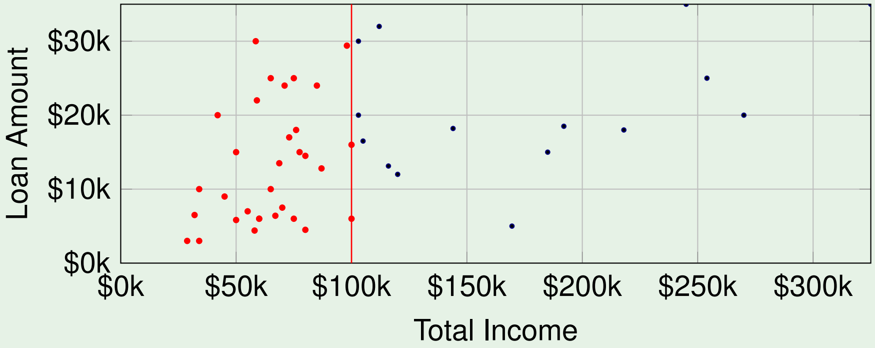
Examining Numerical Data

Colby Community College

Example 1

Let us consider a scatterplot of borrowers total income and the loan amount from the `loan50` data set.



We can see that the many of borrowers earn $100,000 a year or less.

Example 2

Let us consider a scatterplot of borrowers total income and the loan amount from the `loan50` data set.

It is clear there is a **nonlinear** association between the median household income and the poverty rate.

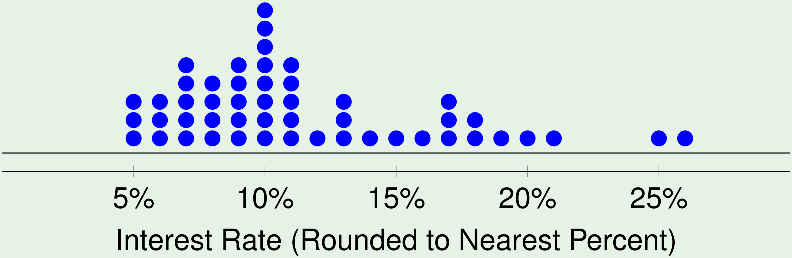
Definition

A **dot plot** is a one-variable scatterplot. Each data value is plotted as a point above a horizontal scale of values. Dots representing equal values are stacked.

Note

Dot plots work best with integer data. It is common to round decimals before building a dot plot.

Example 3



Definition

A **parameter** is a numerical measurement describing some characteristic of a population.

Definition

A **statistic** is a numerical measurement describing some characteristic of a sample.

Note

Parameter and population both start with a “P.”
Statistic and sample both start with a “S.”

Definition

A **measure of center** is a value at the center or middle of a data set.

Definition

The **mean** of a set of data is the measure of center found by adding all the data values and dividing by the total number of data values.

Note

The mean is also known as the **average**.

Properties of the Mean

- Sample means drawn from the same population tend to vary less than other measures of center.
- A disadvantage of the mean is that just one extreme value can change the value of the mean substantially.

Common Notation

Sample statistics are usually represented by English letters, such as $\bar{x}$, while population parameters are usually represented by Greek letters, such as μ.

$\sum$	denotes the sum of a set of data values.
x	is used as a placeholder for the variable of interest.
n	represents the number of data values in a sample.
N	represents the number of data values in a population.
$\bar{x} = \dfrac{\sum x}{n}$	is the mean of a set of sample values.
$\mu = \dfrac{\sum x}{N}$	is the mean of all values in a population.

Example 4

Suppose we measure the of data speeds of smartphones from the four major carriers. The table contains five data speeds, in megabits per second (Mbps), from this data set.

Carrier	Verizon	Verizon	Verizon	Verizon	Verizon
Mbps	38.5	55.6	22.4	14.1	23.1

The mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{38.5 + 55.6 + 22.4 + 14.1 + 23.1}{5} = \frac{153.7}{5} = 30.74 \text{ Mbps}$$

Note

Round statistics and parameters to one more decimal place than found in the data.

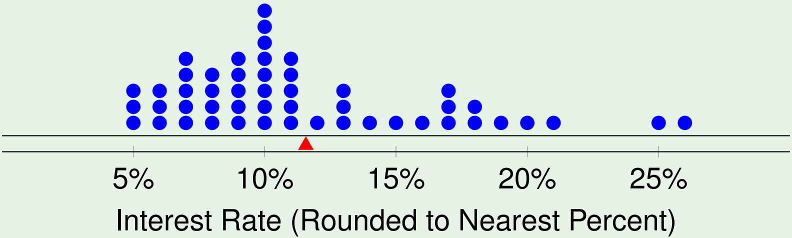
Note

It is common to mark the mean on a dot plot.

Example 5

The mean of `interest_rate` is: (Do not round the data values.)

$$\bar{x} = \frac{\left(\begin{array}{c} 5.31\% + 5.31\% + 5.32\% + 6.08\% + 6.08\% + 6.08\% + 6.71\% + 6.71\% + 7.34\% \\ +7.35\% + 7.35\% + 7.96\% + 7.96\% + 7.96\% + 7.97\% + 9.43\% + 9.43\% + 9.44\% \\ +9.44\% + 9.44\% + 9.92\% + 9.92\% + 9.92\% + 9.92\% + 9.93\% + 9.93\% + 10.42\% \\ +10.42\% + 10.9\% + 10.9\% + 10.91\% + 10.91\% + 10.91\% + 11.98\% + 12.62\% \\ +12.62\% + 12.62\% + 14.08\% + 15.04\% + 16.02\% + 17.09\% + 17.09\% + 17.09\% \\ +18.06\% + 18.45\% + 19.42\% + 20\% + 21.45\% + 24.85\% + 26.3\% \end{array}\right)}{50} = 11.567\%$$



Example 6

We saw in Example 5 that the average loan interest rate was 11.567%.

*What is the mean of **all** loans in the country?*

The best guess we can make is to use the sample mean of 11.567%.

Is this a good guess?

Just because the sample mean is the only educated guess we can make, doesn't mean it's anywhere close to the population mean.

Definition

A **point estimate** is a single value used to estimate a population parameter.

Note

We will discuss tools in Chapter 5 and beyond to determine how well a point estimate estimates a parameter.

Example 7

We would like to determine if a new drug is more effective at treating asthma attacks than the standard drug. A trial of 1500 adults is setup, giving the following data.

	New drug	Standard drug
Number of patients	500	1000
Total asthma attacks	200	300

Can we conclude that the new drug is more effective?

Raw numbers can be deceptive when group sizes are unbalanced.

Looking at the table, 200 is a smaller number than 300.

But, when we calculate the means we get:

$$\begin{aligned} &\text{New drug:} && 200 \text{ attacks}/500 \text{ patients} = 0.4 \text{ attacks per patient} \\ &\text{Standard drug:} && 300 \text{ attacks}/1000 \text{ patients} = 0.3 \text{ attacks per patient} \end{aligned}$$

The average number of asthma attacks per patient is higher with the new drug, so it's not more effective.

Example 8

Emilio opened a food truck last year, and his business has stabilized over the last three months. During this three month period he made $\$11,000$, while working 625 hours.

Is Emilo doing well with his new business?

If you haven't ran a food truck, it can be hard to tell if $\$11,000$ is a high amount or a low amount.

Calculating the mean gives:

$$\frac{\$11000}{625 \text{ hours}} = \$17.60 \text{ per hour}$$

Note

The mean gives a standardized a metric into something easier to interpret and compare.

Example 9

Suppose we want to find the average income per person across the entire United States. To do so, we take the mean of the `per_capita_income` variable from the `county` data set.

Is this the best approach?

No. Each county represents multiple people. If we computed the mean of `per_capita_income` we would be treating a county with $5,000$ residents and a county with $5,000,000$ residents the same.

To account to differences in the population of each county, we should:

1. Calculate the total income for each county.
   ($\texttt{pop2017} \times \texttt{per\_capita\_income}$)
2. Add up all the county income totals
3. Then divide by the total number of people in the country.

Using this method we would find the average income per person in the US is $\$30,861$. If we had used the simple mean of `per_capita_income` the result would have been $\$26,093$, which is much lower.

Definition

A **weighted mean** is a mean where some values contribute more than others.

$$\bar{x} = \frac{\sum w_x \cdot x}{\sum w_x}$$

The values w_x are called the **weights.**

Example 10

Your final grade in this class is a weighted mean of the following four values:

Value	% of Grade	Weight
Your average attendance score	10%	10
Your average assignment score	30%	30
Your average exam score	40%	40
Your final exam score	20%	20

So, your final grade is calculated using the formula:

$$\text{Grade} = \frac{10 \cdot \overline{\text{attendance}} + 30 \cdot \overline{\text{assignments}} + 40 \cdot \overline{\text{exams}} + 20 \cdot \text{final}}{10 + 30 + 40 + 20}$$

Note

We could also use the decimal versions of the percentages as the weights, instead of the whole numbers.

Definition

The **median** of a data set is the middle value when the original data values are arranged in order of smallest to largest.

Properties

- The median does not change by large amounts when we include an extreme value.

Notation

The median of a sample is denoted $\tilde{x}$.

Procedure

1. Sort the values.
2. - If the number of data values is odd, the median is the number located in the exact middle of the sorted list.
   - If the number of data values is even, the median is found by computing the mean of the two middle numbers in the sorted list.

Example 11

Let find the median data speed using the table from Example 4.

Carrier	Verizon	Verizon	Verizon	Verizon	Verizon
Mbps	38.5	55.6	22.4	14.1	23.1

First sort the data values.

14.1	22.4	23.1	38.5	55.6

We have 5 data values so the median is $\tilde{x} = 23.1$ Mbps.

Note

This different than the mean 30.74 Mbps.

Example 12

Let find the median data speed using the table from Example 4, but with an extreme value added in.

Carrier	Verizon	Verizon	Verizon	Verizon	Verizon	Verizon
Mbps	38.5	55.6	22.4	14.1	23.1	192.6

First sort the data values.

14.1	22.4	23.1	38.5	55.6	192.6

We have 6 data values so $\tilde{x} = \dfrac{23.1 + 38.5}{2} = 30.80$ Mbps.

Note

This is very different from the mean of this table.

$$\bar{x} = \frac{14.1 + 22.4 + 23.1 + 38.5 + 55.6 + 192.6}{6} = 173.15 \text{ Mbps}$$

Definition

A **histogram** is a graph consisting of bars of equal width drawn adjacent to each other. Each bar represents a “bin” of data values and the height of each bar is how many data values are in the “bin”.

Important Uses

- Visually displays the shape of the distribution of the data.
- Shows the location of the center of the data.
- Shows the spread of the data.
- Identifies extreme values.

Definition

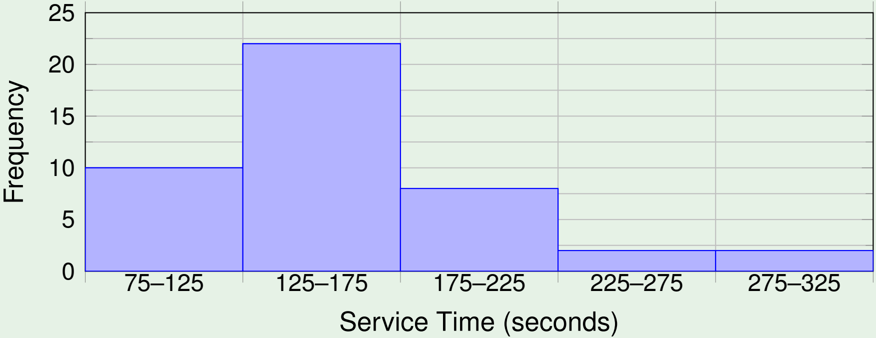
Histograms provide a view of the **data density**. Higher bars represent where the data is relatively more common.

Example 13

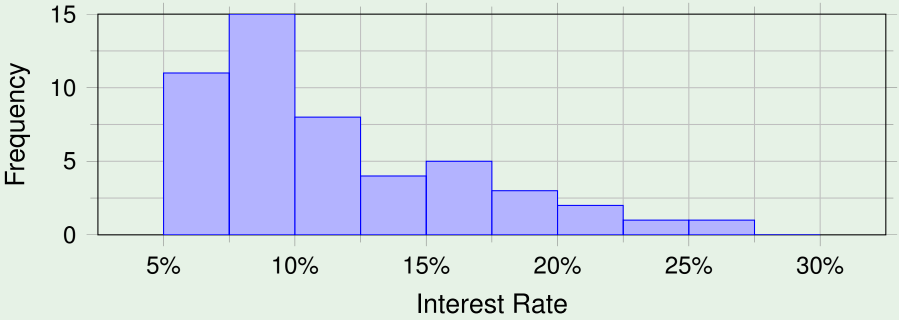
The table contains drive-through service times, in seconds.

107	139	197	209	281	254	163	150	127	308	206
169	83	127	133	140	143	130	144	91	113	153
252	200	117	167	148	184	123	153	155	154	100
101	138	186	196	146	90	144	119	135	151	197

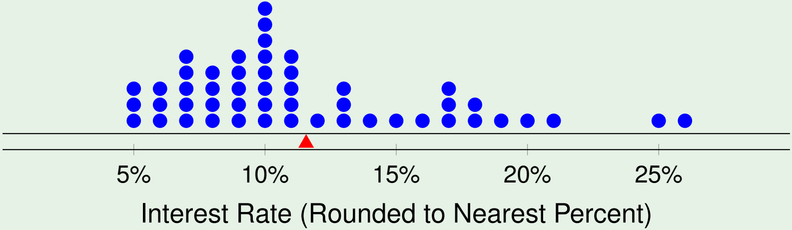
Let's build a histrogram:



Example 14



15
10
5
0
Frequency
5%
10%
15%
20%
25%
30%
Interest Rate

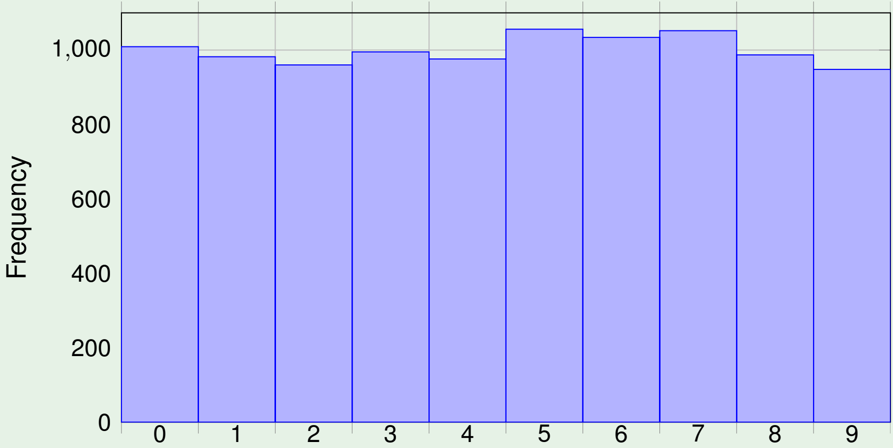


5%
10%
15%
20%
25%
Interest Rate (Rounded to Nearest Percent)

Definition

If all of the bars in a histogram are close to the same height, then the distribution is said to be **uniformly distributed**.

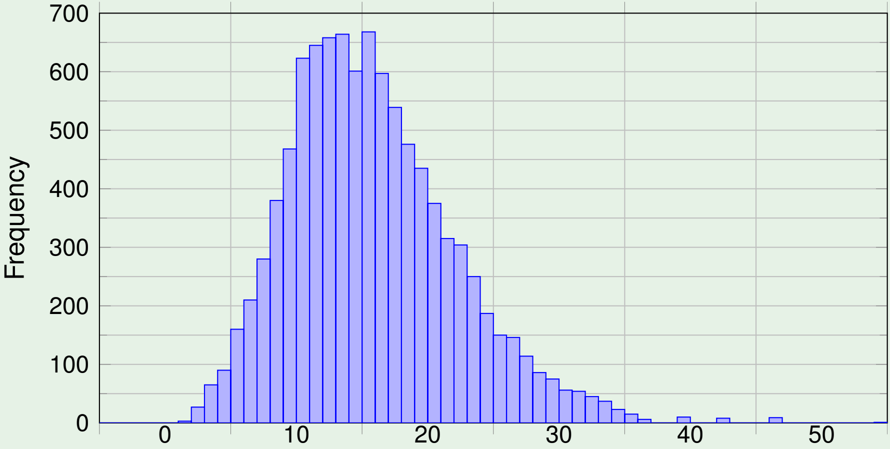
Example 15



Definition

When the data trails off to the right and has a longer right tail, the distribution is said to be **right skewed.**

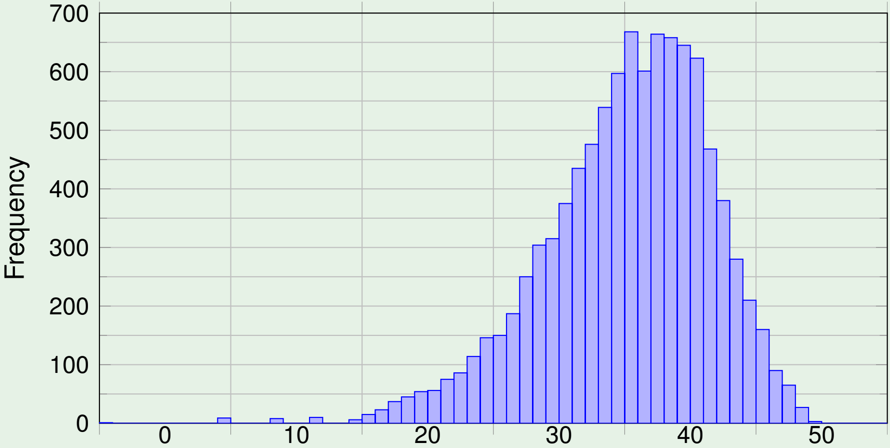
Example 16



Definition

When the data trails off to the left and has a longer left tail, the distribution is said to be **left skewed**.

Example 17



Note

Skewed to the left resembles the toes on your left foot.

Skewed to the right resembles the toes on your right foot.

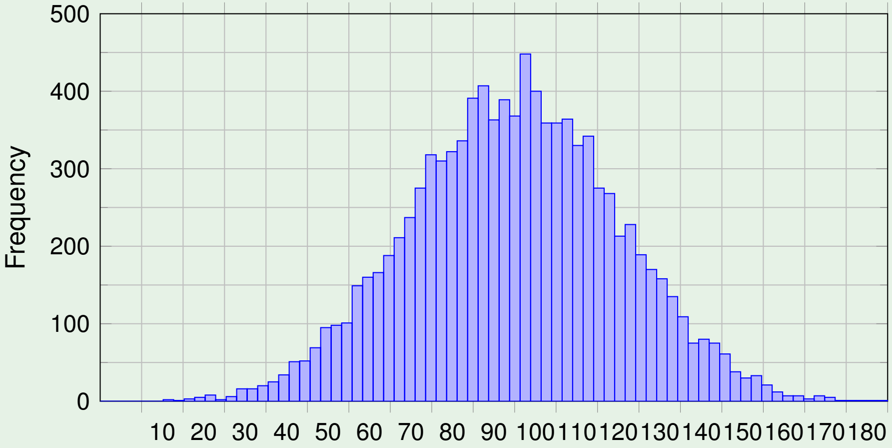
Definition

If the distribution of data is skewed to the left or skewed to the right, the distribution is called **skewed**.

Definition

Data sets that show roughly equal trailing off in both directions are called **symmetric.**

Example 18



Definition

A **mode** is represented by a prominent peak in the distribution.

Definition

If a distribution has exactly one mode, it is called **unimodal**.

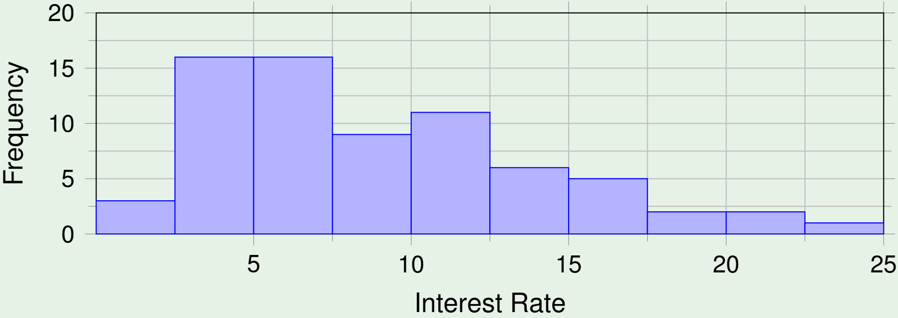
Definition

If the distribution has exactly two modes, it is called **bimodal**.

Definition

If the distribution has more than two modes, it is called **multimodal**.

Example 19



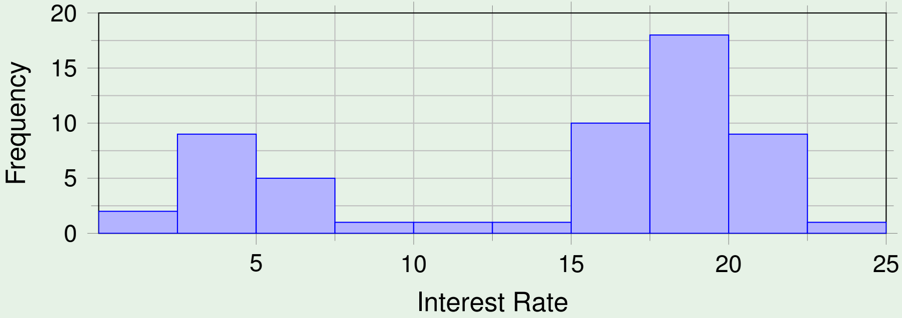
How many modes does this distribution have?

One

Is this distribution unimodal, bimodal, or multimodal?

Unimodal

Example 20



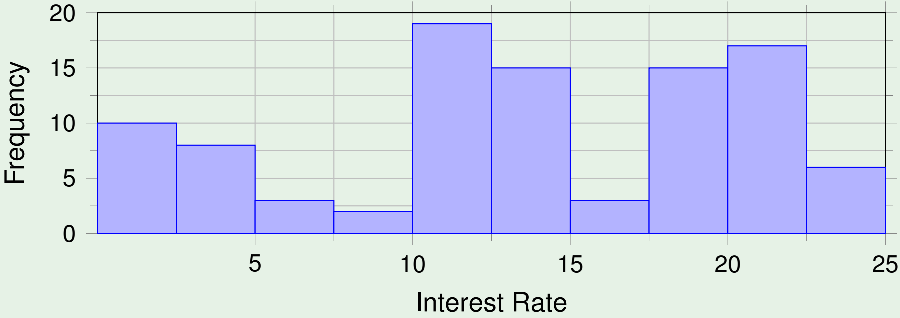
How many modes does this distribution have?

Two

Is this distribution unimodal, bimodal, or multimodal?

Bimodal

Example 21



How many modes does this distribution have?

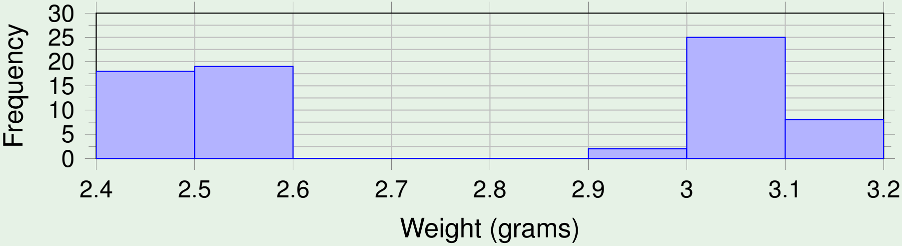
Three

Is this distribution unimodal, bimodal, or multimodal?

Multimodal

Example 22

Let us consider the weights of randomly selected pennies.



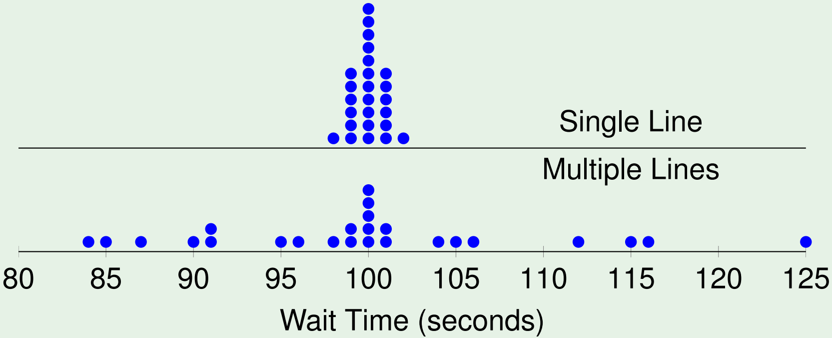
The modes describe two different types of pennies in circulation:

- Pennies made before 1983 are 95% copper and 5% zinc.
- Pennies made after 1983 are 2.5% copper and 97.5% zinc.

Since copper is more dense than zinc, pennies made after 1983 weigh less than those made before 1983.

Example 23

Consider the waiting times at a bank.



Both of these data sets have the same mean, but they are very different.

Definition

A **measure of variation** describes how spread out a distribution is.

Definition

The distance between an observation and it's mean is called it's **deviation**. You calculate the deviation as: $x - \bar{x}$.

Example 24

Recall that the mean of `interest_rate` is 11.57%.

data	deviation
10.90	$x_1 - \bar{x} = 10.90 - 11.57 = -0.67$
9.92	$x_2 - \bar{x} = 9.92 - 11.57 = -1.65$
26.30	$x_3 - \bar{x} = 26.30 - 11.57 = 14.73$
	$\vdots$
6.08	$x_{50} - \bar{x} = 6.08 - 11.57 = -5.49$

Note

A positive deviation means the data value is larger than the mean.
A negative deviation means the data value is smaller than the mean.

Definition

The **variance** of a sample, denoted as s^2, is

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

Note

When computing a variance of a population, you divide by N instead of $n - 1$ for fiddly mathematical reasons.

Definition

The **standard deviation**, denoted s, is the square root of the variance.

Note

The standard deviation of a population is denoted σ and variance σ^2.

Example 25

The variance of `interest_rate` is

$$
\begin{aligned}
s^2 &= \frac{\sum (x - \bar{x})^2}{n-1} \\
&= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\
&= \frac{0.45 + 2.72 + 216.97 + \cdots + 30.14}{49} \\
&= 25.524
\end{aligned}
$$

The standard deviation of `interest_rate` is

$$s = \sqrt{s^2} = \sqrt{25.524} = 5.052$$

Note

Computers are often used to compute variance and standard deviation.

Properties

- The standard deviation is a measure of how much most of the data values deviate from the mean.
- The value of the standard deviation is never negative.
- The value of the standard deviation is only zero when all of the data values are exactly the same.
- Larger values of s indicate greater amounts of variation.
- The standard deviation can increase dramatically with one or more extreme values.
- The units of the standard deviation are the same units as the original data values.

Note

For reason will explore in Chapter 4, we expect most of the data to fall within one standard deviation of the mean.

Definition

Percentiles are measures of location, denoted $P_1, P_2, \ldots, P_{99}$, which divide a set of data into 100 groups with about 1% of the values in each group.

Formula

The process of finding the percentile that corresponds to a particular data value x is given by the following:

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Note

Round percentiles to the nearest whole number.

Example 26

The table lists the 50 smartphone data speeds, in Mbps.

38.5	55.6	22.4	14.1	23.1	24.5	6.5	21.5	25.7	14.7
77.8	71.3	43.0	20.2	15.5	13.7	11.1	13.5	10.2	21.1
15.1	14.2	4.5	7.9	9.9	10.3	6.2	17.5	22.2	13.1
18.2	28.5	15.8	15.0	11.1	11.8	16.0	10.9	1.8	34.6
4.6	12.0	11.6	3.6	1.9	7.7	0.8	4.5	1.4	3.2

Let us find which percentile the data value 11.8 Mbps in.

There are 20 data values less than 11.8 Mbps.

$$\text{Percentile of } 11.8 = \frac{20}{50} \cdot 100 = 40$$

A data speed of 11.8 Mbps is in the 40th percentile.

Note

This can be interpreted loosely as 40% of the data speeds are slower than 11.8 Mbps and 60% of the data speeds are faster than 11.8 Mbps.

Converting a Percentile to a Data Value

Notation:

- n is the total number of values in the data set.
- k is the percentile being used.
- L is the locator that gives the position of a value.
- P_k is the kth percentile.

To find which data value is in the P_k percentile:

1. Sort the data from lowest to highest.
2. Compute $L = \left(\frac{k}{100}\right) n$
3.
   - If L is a whole number, the value of the kth percentile is midway between the Lth value and the next value in the sorted data. Add the Lth value and $(L+1)$th value, then divide by 2.
   - If L is not a whole number, round L up to the nearest whole number. $P_k$is the Lth data value.

Example 27

The table lists the 50 smartphone data speeds, in Mbps.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

Let us find which value is in the 25th percentile, P_{25}.

First, sort the data.

We next need to compute

$$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 50 = 12.5$$

Since $L = 12.5$ is not a whole number, we round up to 13.

So, P_{25} is the 13th data value, 7.9 Mbps.

Definition

Quartiles are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

Quartile Descriptions

First Quartile, Q_1: Same value as P_{25}. It separates the bottom 25% of the sorted values from the top 75%.

Second Quartile, Q_2: Same as the P_{50} and the median. It separates the bottom 50% of the sorted values from the top 50%

Third Quartile, Q_3: Same as P_{75}. It separates the bottom 75% of the sorted values from the top 25%.

Note

Use the same procedure for calculating percentiles to calculate quartiles.

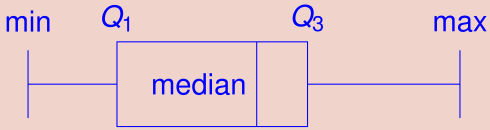
Definition

For a set of data, the **5-number summary** consists of the five values:

1. Minimum
2. Q_1
3. Median (Q_2)
4. Q_3
5. Maximum

Definition

A **boxplot** is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q_1, the median, and the third quartile Q_3.



Example 28

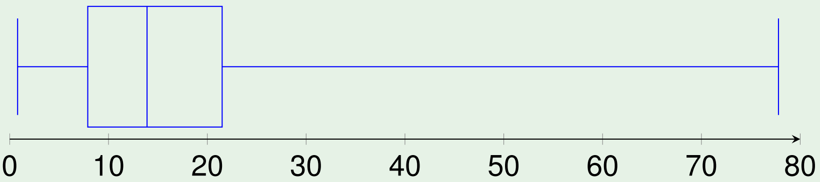
The table lists, in order from lowest to highest, the 50 smart phone speeds, in Mbps.

0.8	1.4	1.8	1.9	3.2	3.6	4.5	4.5	4.6	6.2
6.5	7.7	7.9	9.9	10.2	10.3	10.9	11.1	11.1	11.6
11.8	12.0	13.1	13.5	13.7	14.1	14.2	14.7	15.0	15.1
15.5	15.8	16.0	17.5	18.2	20.2	21.1	21.5	22.2	22.4
23.1	24.5	25.7	28.5	34.6	38.5	43.0	55.6	71.3	77.8

The five number summary is:

minimum=0.8, $Q_1 = 7.9$, Median=13.9, $Q_3 = 21.5$, maximum=77.8

The box plot is:



Definition

The **interquartile range** (**IQR**) is $Q_3 - Q_1$.

Definition

A data value is often considered an **outlier** if

- The data value is greater than $Q_3 + 1.5 \cdot$ **IQR**.
- The data value is less than $Q_1 - 1.5 \cdot$ **IQR**.

Why Care About Outliers?

Examining data for outliers can help identify:

- Strong skew in the distribution.
- Possible data collection or data entry errors.
- Insights into some interesting property of the data.

Example 29

The table lists 50 smartphone speeds, in Mbps.

38.5	55.6	22.4	14.1	23.1	24.5	6.5	21.5	25.7	14.7
77.8	71.3	43.0	20.2	15.5	13.7	11.1	13.5	10.2	21.1
15.1	14.2	4.5	7.9	9.9	10.3	6.2	17.5	22.2	13.1
18.2	28.5	15.8	15.0	11.1	11.8	16.0	10.9	1.8	34.6
4.6	12.0	11.6	3.6	1.9	7.7	0.8	4.5	1.4	3.2

Recall that the five number summary is:

minimum=0.8, $Q_1 = 7.9$, Median=13.9, $Q_3 = 21.5$, maximum=77.8

The IQR would then be:

$$\text{IQR} = Q_3 - Q_1 = 21.5 - 7.9 = 13.6$$

The limits for outliers would be:

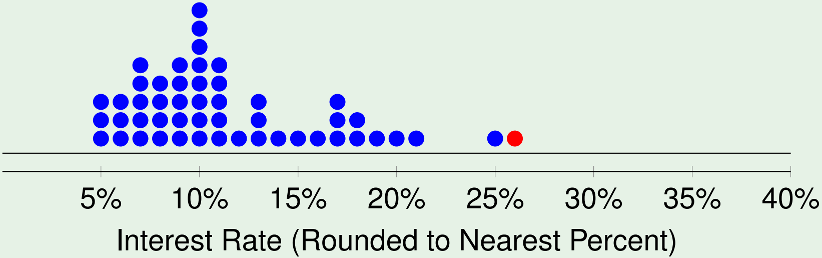
$$\text{Lower Limit} = Q_1 - 1.5 \cdot \text{IQR} = 7.9 - 1.5 \cdot 13.6 = -12.5$$

$$\text{Upper Limit} = Q_3 + 1.5 \cdot \text{IQR} = 21.5 + 1.5 \cdot 13.6 = 41.9$$

The outliers are marked in red.

Example 30

`intrest_rate` has an outlier at 26.3%.



What would happen if this loan had a different interest rate?

scenario	median	IQR	$\bar{x}$	s
original value of 26.3%	9.93%	5.76%	11.57%	5.05%
move 26.3% $\rightarrow$ 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% $\rightarrow$ 35%	9.93%	5.76%	11.74%	5.68%

Definition

A statistic is called **robust** if extreme observations have little effect on the value.

Robustness

The robust statistics are:

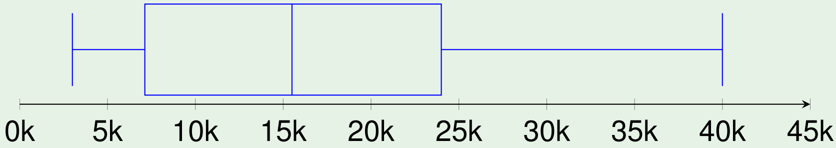
- The median.
- The IQR.

The non-robust statistics are:

- The mean.
- The standard deviation.

Example 31

The distribution of loan amounts in the `loan50` data set is skewed right, with a few large loans lingering out into the right tail.



If you wanted to understand a typical loan size, should you be more interested in the mean or the median?

It depends!

If we wish to know what a typical loan looks like, the median would probably be more useful.

But, if the goal is something that scales well, such as how much money a bank has to have on hand to cover 1,000 loans, the mean would be more useful.