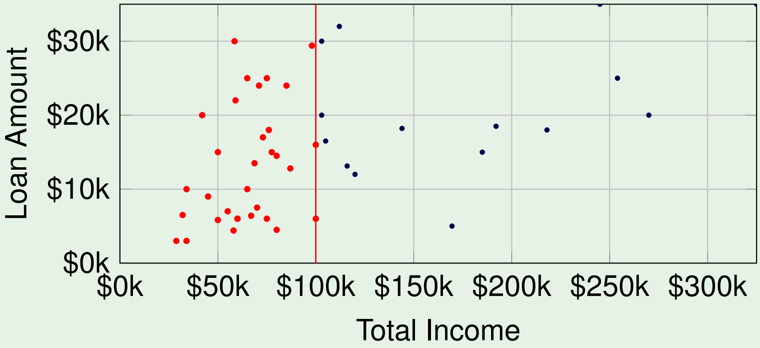
Examining Numerical Data

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.

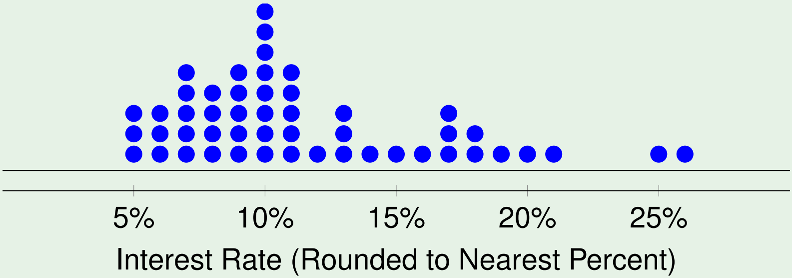
Definition

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.

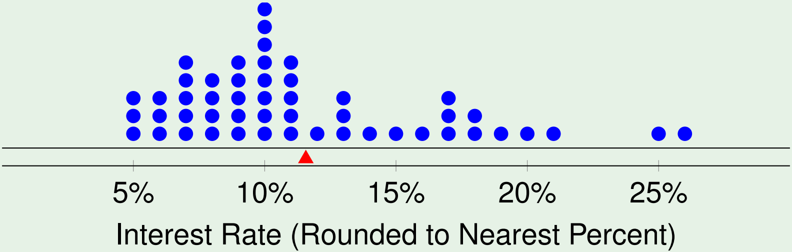
Note

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:

New drug: $200 \text{ attacks}/500 \text{ patients} = 0.4 \text{ attacks per patient}$

Standard drug: $300 \text{ attacks}/1000 \text{ patients} = 0.3 \text{ attacks per patient}$

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.

Note

Histograms are similar to dot plots, except that each bar represents a range of data values.

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: