Chapter 13

Descriptive statistics

Only computers can make sense of row upon row of raw data. To human eyes, there are just too many numbers contained in any given dataset to grasp it as a whole. Thankfully, we can use descriptive statistics to summarize data in a way that's easy for our simple, flesh-based brains to interpret.

We can obtain a condensed summary of data by calculating certain representative values called *summary statistics*. A summary statistic is a number computed from the data, which quantitatively describes some data characteristic, such as the smallest value, the largest value, or the average. The best part about summary statistics is that you don't need to be a stats expert to calculate them—all the computations that we'll study in this chapter consist of basic counting, summations, and maybe a square root or two.

We can also get an overall impression of our dataset by making a *visual summary*: a plot or chart that shows the dataset. You may already be familiar with certain common visual summaries like bar charts and box plots. By mapping characteristics of data onto visual elements, we can get a quick overview of the dataset. The human visual cortex is surprisingly efficient at spotting trends and abnormalities in data presented graphically, so it's worth learning how to create visual summaries to take advantage of your innate pattern-spotting abilities.

Exploring your data is not only a fundamental skill that will be used throughout the rest of the book, but it's a mandatory, non-negotiable, do-this-or-else-the-stats-gods-will-curse-you, preliminary step to data analysis. Summarizing your dataset will help you to: compare your data to other datasets; identify problems in the data; notice patterns you might otherwise have missed; determine whether your dataset is suitable for answering the research question you want

to study, and if it is, help you interpret the eventual answers.

The goal of descriptive stats is to characterize data in a manner that makes it quick and easy to interpret. In this chapter we'll learn to create basic numerical and visual summaries for tabular data.

13.1 Definitions and formulas

This section we'll define the new concepts we'll use in this chapter and lists all the formulas used to calculate summary statistics.

13.1.1 Datasets

In the context of statistics, a *dataset* is a collection of *values* obtained from an experiment or an observational study. In this chapter we'll focus on *tabular data*, which consists of numerical and categorical variables that can be stored in a excel-like table:

index	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
3	x_3	y_3	z_3
$\vdots$	$\vdots$	$\vdots$	
n	x_n	y_n	z_n

Table 13.2: A dataset that consists of n observations of the variable `x`, `y`, and `z`. Each row in this table corresponds to one observation. The `index` column can be used to refer to specific observations. The size of this dataset is n, which means it contains n observations.

Each row in the dataset shown in Table 13.2 corresponds to one observation, and each observation contains three *variables* called x, y, and z. The three columns of values are denoted as $\mathbf{x}$ (the first column of values), $\mathbf{y}$ (the second column of values), and $\mathbf{z}$ (the last column). We can analyze each of the variables individually, or look for relationships between variables.

The variables in the dataset can be numerical (numbers like -0.3, 2.2, and 120), or categorical (discrete quantities like grades A, B, C, D, and F). We use different summary statistics to characterize variables depending on whether they are numerical or categorical.

13.1.2 Summary statistics for numerical variables

Numerical variables describe continuous quantities like weight, length, temperature, and time. The values of numerical variables can be compared, sorted, added together, subtracted, and manipulated though other math operations.

Measures of central tendency

One way to summarize a dataset of values $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ is to find a single number that represents the "centre" of the distribution of all the values. We can define the following different types of "centres" for a list of numerical values:

- *Mean*: The *arithmetic mean* represents the average value of the dataset and is computed by taking the sum of all values divided by the number of values:

$$\textbf{Mean}_\mathbf{x} = \frac{1}{n}\sum_i^n x_i.$$

  We'll also use the notation **Mean** to denote the arithmetic mean, when there is only one variable under consideration. When working with multiple variables like $\mathbf{x}$ and $\mathbf{y}$, we specify which variable we're computing the mean for using a subscript $\textbf{Mean}_\mathbf{x} = \frac{1}{n}\sum_i^n x_i$, $\textbf{Mean}_\mathbf{y} = \frac{1}{n}\sum_i^n y_i$, and $\textbf{Mean}_\mathbf{z} = \frac{1}{n}\sum_i^n z_i$.
- *Median*: Another way to describe the centre of the data is to find the middle value in the dataset, which is called the *median* and denoted **Med**. Half the values x_i in the dataset are smaller than the median **Med**, and half the values are larger than **Med**. To find the median we can sort the values in ascending order, and report the value that appears in the middle of the sorted list, as shown in Figure 13.1.

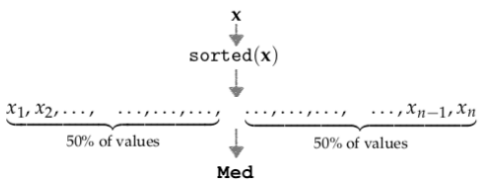


Figure 13.1: Illustration of the median value, **Med**, which split the dataset into two equal parts. Half the values x_i in the dataset are smaller than the median and the other half are larger than the median.

The following example code shows how to calculate the median:

```
def median(x):
    xs = sorted(x)
    midpoint = len(xs) // 2  # == math.floor(len(xs)/2)
    if len(xs) % 2 == 1:
        return xs[midpoint]
    else:
        return (xs[midpoint-1] + xs[midpoint])/2
```

If the number of values in the dataset **x** is an odd number, we return the middle value from the sorted list. If the number of values is odd, we use linear interpolation between the values straddling the midpoint.

- *Mode*: The most frequently observed value is denoted **Mode**. A variable can have no mode when no single value appears more often than any other, or it can have more than one mode when there is a "tie" for the most common value.

Measures of position

In addition to the centre, it's also useful to report the following summary statistics that tell us the position of the values in a dataset.

- *Quartiles*: The three quartiles, $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$, are values that the divide the dataset into four equal parts, as illustrated in Figure 13.2. This is similar to how the median **Med** divides the dataset into two equal parts.

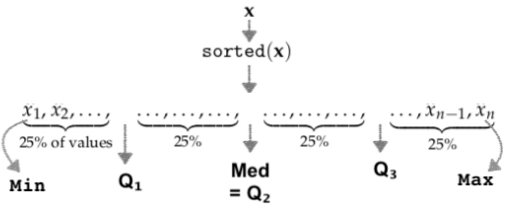


Figure 13.2: Illustration of the four quartiles $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ which split the dataset into four equal parts.

The following code snippet shows how to find the k^{th} quartile for the dataset **x**:

```
def quartile(x, k):
    xs = sorted(x)
    ifloat = k*(len(xs)-1)/4
    ifloor = int(ifloat)  # == math.floor(ifloat)
    ifrac = ifloat - ifloor
    return xs[ifloor] + ifrac*(xs[ifloor+1] - xs[ifloor])
```

We first sort the values x_i in ascending order, then compute the index $i = \frac{k}{4}(n-1)$ of the k^{th} quartile, which is a floating point number. We use the linear interpolation between the values that surround the index

$$\mathbf{Q}_{\text{k}} = x_{\lfloor i \rfloor} + (i - \lfloor i \rfloor)(x_{\lfloor i \rfloor + 1} - x_{\lfloor i \rfloor}),$$

where $\lfloor i \rfloor$ = `ifloor` denotes the greatest integer that's less than i. The quantity $(i - \lfloor i \rfloor)$ = `ifrac` denotes the fractional part of the index, and tells us the proportion between $x_{\lfloor i \rfloor}$ and $x_{\lfloor i \rfloor + 1}$ we must mix to compute the answer.

- *Minimum*: We denote **Min** the smallest value in the dataset.
- *Maximum*: We denote **Max** the largest value in the dataset.

Taken together, the five numbers **Min**, $\mathbf{Q}_1$, $\mathbf{Q}_2$, $\mathbf{Q}_3$, and **Max** are called the *five number summary* of the dataset, which tell us the boundary values that contain each 25% chunk of the dataset when it appears in sorted order. Note the median value **Med** is equal to the second quartile.

Measures of dispersion

Another important characteristic of any dataset is how spread out it is.

- *Range* **Range**: the difference between the largest and the smallest value in the dataset:

$$\mathbf{Range} = \mathbf{Max} - \mathbf{Min}.$$

- *Inter-quartile range* **IQR**: the span of the the middle fifty percent of a variable

$$\mathbf{IQR} = \mathbf{Q}_3 - \mathbf{Q}_1.$$

- *Variance* **Var**: the average of the squared differences from the mean:

$$\mathbf{Var} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbf{Mean})^2.$$

The *standard deviation* **Std** is the square root of the variance $\mathbf{Std} = \sqrt{\mathbf{Var}}$.

Measures of association

We're often interested in learning about the relationship between two variables in a dataset. A *positive linear association* between the variables x and y in the dataset means that large values of x_i tend to be associated with large values of y_i, and small values of x_i are associated with small values of y_i. A *negative linear association* describes the opposite phenomenon, where large values of x_i are associated with small values of y_i, and vice versa.

- *Covariance* **cov**(x,y) is a measure of the joint variability of two variables x and y:

$$\mathbf{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mathbf{Mean_x})(y_i - \mathbf{Mean_y}).$$

  Note the formula is similar to the formula for the variance **Var**, but is computed from the joint difference of x and y from their mean values.

  The values of the covariance between two random variables can take on values from from $-\infty$ to $+\infty$.
- *Correlation* **corr**(x,y) is a *normalized* measure of association between the variables x and y. The value of the **corr**(x,y) is also called the *Pearson correlation coefficient*, and gives the linear relatedness between the variables x and y. Correlation is the normalized version covariance, which involves division by the standard deviation of individual variables:

$$\mathbf{corr}(x, y) = \frac{\mathbf{cov}(x, y)}{\mathbf{Std}_x\,\mathbf{Std}_y}.$$

  The correlation coefficient is always a dimensionless quantity and ranges from -1 to $+1$. We'll learn how to interpret the values of the correlation coefficient in Section 13.5.

Note that covariance and the correlation coefficient are symmetric quantities, meaning $\mathbf{cov}(x, y) = \mathbf{cov}(y, x)$ and $\mathbf{corr}(x, y) = \mathbf{corr}(y, x)$, so the order of the variable is x and y appear in the formula doesn't matter.

13.1.3 Frequencies of categorical variables

Categorical variables are used to describe one discrete set of possible values, like the answers of a multiple choice question in a survey. Most of the summary statistics we might want to calculate for

categorical variables involve counting the number of occurrences of given value within the dataset, which we call the *frequency* of the value.

Consider a dataset of size n with two categorical variables x and y. In other words, we have n observations of the form (x_i, y_i). We'll use the following notation to describe the different counts of occurrences of specific values:

- $\mathbf{Freq_x}(v)$: The *frequency* of the value v denotes the count of the values x_i in the dataset $\mathbf{x}$ for which $x_i = v$:

  $$\mathbf{Freq_x}(v) \quad = \quad \text{count of } x_i = v \text{ in the dataset.}$$

  Similarly, we use the notation $\mathbf{Freq_y}(w)$ to denote the count of the observations in the dataset $\mathbf{y}$ for which $y_i = w$.
- $\mathbf{JointFreq_{x,y}}(v, w)$: denotes the *joint frequency* of observations that have $x_i = v$ and $y_i = w$ in the dataset.

  $$\mathbf{JointFreq_{x,y}}(v, w) \quad = \quad \text{count of } x_i = v \text{ and } y_i = w \text{ in the dataset.}$$

  For example, the relative frequency of $\mathbf{JointFreq_{x,y}}(v, w) = 7$ tells us there are 7 observations in the dataset that have $x_i = v$ and $y_i = w$. The term "joint" tells us we're counting the joint occurence of two variables in observations, rather than studying the two variables separately.

In addition to the frequency of occurrences of particular values, we're often interested in knowing the proportion of observations with a given value out of the total number of observations, which is called a *relative frequency*. Relative frequencies are computed by dividing frequency values by the total number of observations within some subset of interest. The following three types of relative frequencies will be used in this chapter:

- $\mathbf{RelFreq_x}(v)$: denotes the *relative frequency* of the value v in the dataset $\mathbf{x}$. The relative frequency is computed by dividing the number of times v occurs in the dataset by the total size of the dataset

  $$\mathbf{RelFreq_x}(v) \quad = \quad \frac{\mathbf{Freq_x}(v)}{n} = \frac{\text{count of } x_i = v \text{ in the dataset}}{n}.$$

  For example, the relative frequency of $\mathbf{RelFreq_x}(v) = 0.7$ tells us that 70% of the values x_i in the dataset $\mathbf{x}$ have the value v.

- $\textbf{JointRelFreq}_{\textbf{x,y}}(v, w)$: denotes the *joint relative frequency* of observations $x_i = v$ and $y_i = w$ in the dataset.

  $$\textbf{JointRelFreq}_{\textbf{x,y}}(v, w) \quad = \quad \frac{\text{count of } x_i = v \text{ and } y_i = w \text{ in the dataset}}{n}.$$

  For example, the relative frequency of $\textbf{JointRelFreq}_{\textbf{x,y}}(v, w) = 0.2$ tells us that 20% of the values have $x_i = v$ and $y_i = w$.
- $\textbf{CondRelFreq}_{\textbf{x|y}}(v|w)$: denotes the *conditional relative frequency* of the value v within the subset of observations with $y_i = w$. Conditional relative frequencies are computed by dividing the observations that have $x_i = v$ and $y_i = w$ by the total number of observations that have $y_i = w$:

  $$\textbf{CondRelFreq}_{\textbf{x|y}}(v|w) \quad = \quad \frac{\text{count of } x_i = v \text{ and } y_i = w \text{ in the dataset}}{\text{count of } y_i = w \text{ in the dataset}}.$$

  Instead of dividing by joint number of observations by the total number of observations, we divide by the number of observations that satisfy the condition $y_i = w$.

Don't worry about memorizing all the formulas that were presented above. The purpose of this section is to provide the main definitions in a single place to make it easier for you to refer to them. The rest of the chapter is organized around a worked example that illustrates how to compute all these summary statistics and interpret their values.

13.2 Introducing the example dataset

Throughout the rest of this chapter, we'll use an example dataset that consists of student activity obtained from an online learning platform. Imagine that a teacher has collected some data to compare the effectiveness of educational video materials delivered in one of two formats: a new model for teaching the material in the form of a debate and discussions vs. a more traditional lecture format where the teacher states facts and provides explanations. In order to compare the effectiveness of the two teaching methods, the teacher prepares two variants of the same course:

- In the first variant of the course, the video lessons are presented in the usual "`lecture`" format, consisting of recorded video lectures that explain the material, the same way a teacher would present the material in front of a class.

- In the second variant, the same material was covered through video lessons that presented the material in the form of a "`debate`" in which student actors present multiple points of view and discuss them.

Except for the different video lesson curriculum used, the two variants of the course were otherwise identical, covering the same topics, using the same total lecture time, and using the same assessment items (quizzes whose scores are recorded).

Table 13.4 shows the student data from was obtained from the trial course.

student_ID	background	curriculum	effort	score
1	arts	debate	10.96	75.0
2	science	lecture	8.69	75.0
3	arts	debate	8.60	67.0
4	arts	lecture	7.92	70.3
5	science	debate	9.90	76.1
6	business	debate	10.80	79.8
7	science	lecture	7.81	72.7
8	business	lecture	9.13	75.4
9	business	lecture	5.21	57.0
10	science	lecture	7.71	69.0
11	business	debate	9.82	70.4
12	arts	debate	11.53	96.2
13	science	debate	7.10	62.9
14	science	lecture	6.39	57.6
15	arts	debate	12.00	84.3

Table 13.4: The dataset that will be used for all the examples in this chapter.

Each row represents an observation about one student and each column represents one variable. The number of observations, n, is the most basic summary statistic. In this case, $n = 15$, since we have observations for 15 students. Besides student ID (the first column), four measurements were recorded for each observation:

- `background`: describes the student's academic background. This variable can take on one of three possible values: `science`, `business`, or `arts`, depending on the student's academic background (which faculty they are enrolled in).
- `curriculum`: describes the two different options for the educational videos. Students were free to select which of the two variants of the course they want to enrol in, `lecture` or `debate`.

- `effort`: contains the total time in hours spent on the platform, which includes watching videos, solving exercises, and independent exploration.
- `score`: is a grade out of 100 calculated from all the assessment items (quizzes) that the student completed on the platform. We take this `score` to be indicative of the students' proficiency on the material at the end of the course.

The variables `effort` and `score` are numeric, while `background` and `curriculum` are categorical variables.

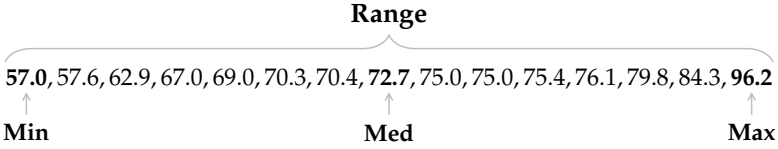
Our goal in this chapter is to make use of numerical and visual summaries to explore this dataset, and try to answer various questions about the effectiveness of the two teaching methods based on the data obtained. We'll start by analyzing variables individually in Section 13.3 (numerical data) and in Section 13.4 (categorical data). Then we'll look for relationships between pairs of variables in sections 13.5, 13.6, and 13.7.

13.3 Describing numeric data

Let's look at the data for the `score` variable, which corresponds to the last column in Table 13.4:

75.0, 75.0, 67.0, 70.3, 76.1, 79.8, 72.7, 75.4, 57.0, 69.0, 70.4, 96.2, 62.9, 57.6, 84.3.

The first thing we'll do is sort the `score` values in increasing order:



Looking at the sorted list of values allows us to immediately identify some important summary statistics. The smallest value is **Min** $=$ 57.0 (the minimum), the largest value is **Max** $=$ 96.2 (the maximum), and the difference between the two values is **Range** $= 96.2 - 57.0 =$ 39.2. The median value is **Med** $=$ 72.7 and it splits the `scores` into two equally-sized chunks: there is an equal number of `score` values smaller than **Med** as there are greater than **Med**.

The simplest way to plot this data is to use a strip chart. Strip charts have no y-axis, just an x-axis in the units of the variable, and each observation is represented by a dot placed at its respective value.

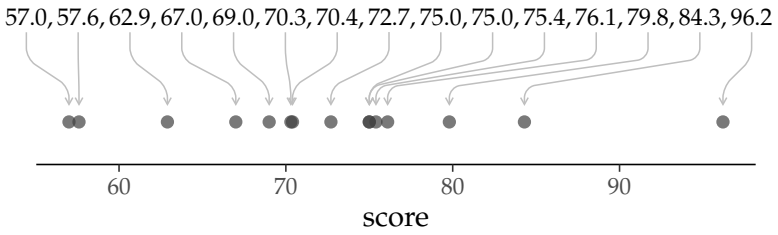


Figure 13.3: a strip chart is a one-dimensional plot where each observation is mapped to a point at the location that corresponds its value.

The visual display of the data allows us to see some patterns in the data that might not be visible when we're looking at the list of numbers.

13.3.1 Mean, variance, and standard deviation

The *mean* (**Mean**) is the sum of all values divided by the number of values:

$$\begin{aligned}\textbf{Mean} &= \tfrac{1}{n}\textstyle\sum_i^n x_i \\ &= \tfrac{1}{15}(75+75+67+70.3+\cdots+62.9+57.6+84.3) \\ &= 72.6.\end{aligned}$$

The mean approximates the value of a "typical" observation in the dataset. It tells us that an average student in this class would score around 73 on their assessment.

To judge the variability of the values in the dataset, we calculate *variance* and *standard deviation*. The variance (**Var**) is the average of the squared differences from the mean. To calculate the `score` variance, we find the difference between each student's `score` and the mean, square the differences, then average the result:

$$\begin{aligned}\textbf{Var} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i - \textbf{Mean})^2 \\ &= \tfrac{1}{15}\left((75-72.6)^2+(75-72.6)^2+\cdots+(57.6-72.6)^2+(84.3-72.6)^2\right) \\ &= 92.9.\end{aligned}$$

The *standard deviation* (**Std**) is the square root of the variance:

$$\textbf{Std} = \sqrt{\textbf{Var}} = \sqrt{92.9} = 9.6.$$

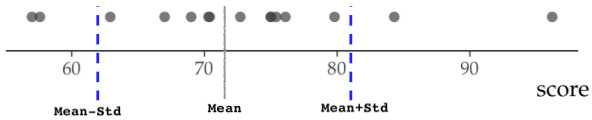


Figure 13.4: A strip chart of the `score` data with additional annotations for the mean and the standard deviation. The mean is shown as a solid line **Mean** $= 72.6$. The two dashed lines are drawn at values **Mean** $-$ **Std** $= 72.6 - 9.6 = 63$ and **Mean** $+$ **Std** $= 72.6 + 9.6 = 82.2$. Note many of the scores are contained between the two dashed lines.

Although variance is used more often in formulas and calculations (we'll see variance come up repeatedly in the probability and statistics chapters), we usually show standard deviation in plots, tables, and reports rather than variance. One of the reasons that standard deviation is the preferred statistic for reporting, is that it's measured in the original units of the variable, while variance is measured in the squared units of the variable. Note many of the values in the dataset are contained in the interval [**Mean** $-$ **Std**, **Mean** $+$ **Std**], which is indicated by the the two dashed lines.

Example The following figure shows the strip chart of three datasets with additional labels for the range of one standard deviation around the mean.

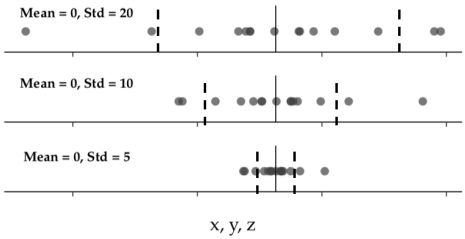


Figure 13.5: Strip charts showing three different datasets that all have the same mean (as indicated by the solid line), but with different standard deviations. The dashed lines are drawn at **Mean** $-$ **Std** and **Mean** $+$ **Std**. You can think of **Std**as the average amount by which the data differ from the mean.

The smaller the standard deviation, the more tightly clustered the values around the mean value. The larger the standard deviation, the more spread out the values. As you can see in Figure 13.5, three datasets can have the same mean **Mean**but very different datasets because they have different variance. That's why we include it's best to include a measure of variability like the standard deviation along

bin	values	frequency
>50 to 60	57.0, 57.6	2
>60 to 70	67.0, 69.0, 62.9	3
>70 to 80	75.0, 75.0, 70.3, 76.1, 79.8, 72.7, 75.4, 70.4	8
>80 to 90	84.3	1
>90 to 100	96.2	1

Table 13.6: One way table for student `score` data for bins of width 10.

with the mean.

13.3.2 Histograms

Strip charts are excellent for displaying the values of observations in a variable, but it can be difficult to see *how many* observations occur at each value, especially when the data has many observations (large n) or when many dots overlap. We'll now learn about the *histogram*, a plot that shows the number of observations that fall within different ranges of possible values.

To make a histogram, we first divide the entire range of values into a series consecutive, non-overlapping intervals called *bins*. For the student `score` data, if we choose bins that are 10 units wide, we'll need 5 bins to cover the entire range of values. We then count the number of observations that fall within each bin. Any time we count how many observations there are in an interval or a category, we call it a *frequency* and denote it as **Freq**. We can display the frequencies within bins in a table called a *one-way table* or a *frequency table*.

Figure 13.6 shows the process of grouping the data points into bins, then counting the total number of observations in each bin.

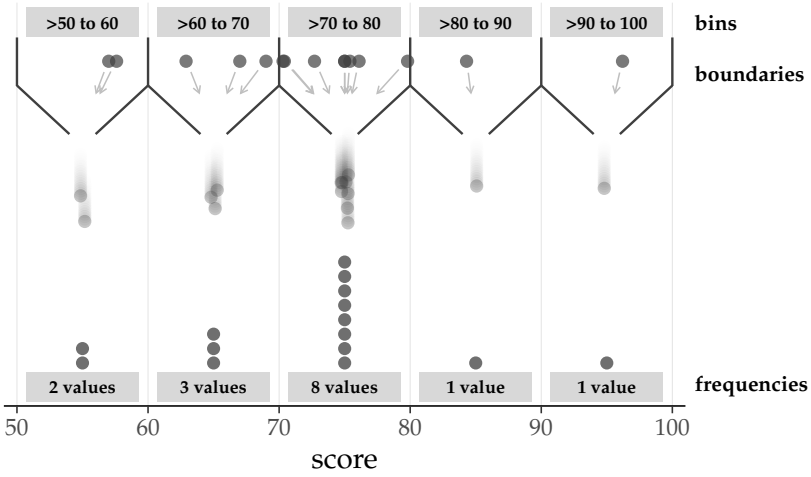


Figure 13.6: Visual representation of the binning process.

In the final step of creating a histogram, we draw a rectangle that spans the width of each bin and set the height proportional to frequency as shown in Figure 13.7.

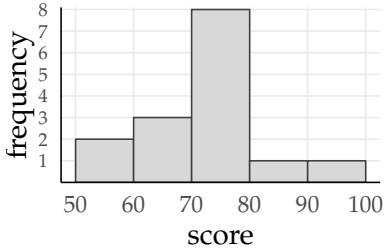


Figure 13.7: A histogram of the `score` data. In a histogram, each bar spreads horizontally across an interval of values called a *bin*. The height of the bars are proportional to the number of observations within each interval. Bins are usually (but not always) of equal size, with no gaps in between.

The histogram in Figure 13.7 gives us a convenient summary for all the data. We can quickly see how much data points fall within each bucket.

The bin with highest frequency is called the *mode*. For continuous data like the `score` variable, the mode is the most frequently observed range of values. In other words, the mode is the "peak" of the histogram. Because there is only one peak in the histogram, we say that the `score` variable is *unimodal*, meaning it as a single mode.

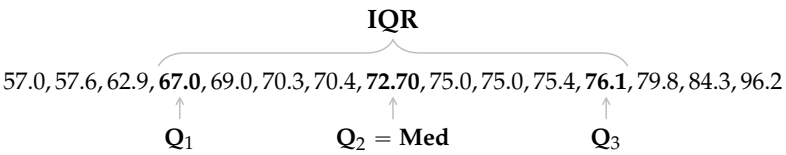
See Figure 13.12 for examples of *bimodal* (two peaks) and *multimodal* (multiple peaks) data distributions.

Plotting a histogram allows to see the data is *skewed*. We see from the histogram that lower `score` values tend to be more common than higher values, so we say that the data is slightly *right-skewed*. See Figure 13.11 (page 23) for illustrations of histograms that of left-skewed, right-skewed, and symmetric shapes.

13.3.3 Quartiles and boxplots

To draw the histogram, we divided data into five bins. Each bin was an interval with an equal width, and each interval could contain any number of observations. We'll now learn about another type of summary plot that divides the data into intervals of varying width, each containing the same number of observations.

The quartiles, denoted as $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$, are the three inner "fenceposts" that demarcate the data into four intervals with an equal number of observations in each:



The second quartile is the same concept as the median of the dataset, $\mathbf{Q}_1 = \mathbf{Med}$. Both concepts refer to the "middle" of the dataset when it appears in sorted order.

The *interquartile range* **IQR** is defined in terms of the first and third quartiles and contains the middle fifty percent of the observations. The interquartile range is the difference between $\mathbf{Q}_3$ and $\mathbf{Q}_1$, and in the case of the `score` data it is $\mathbf{IQR} = \mathbf{Q}_3 - \mathbf{Q}_1 = 76.1 - 67.0 = 9.1$.

We can visualize the quartiles graphically using a boxplot as shown in Figure 13.8. In a boxplot, the width of the rectangular "box" spreads across the **IQR** from $\mathbf{Q}_1$ to $\mathbf{Q}_3$. A vertical line is placed in between at $\mathbf{Q}_2$ (the median).

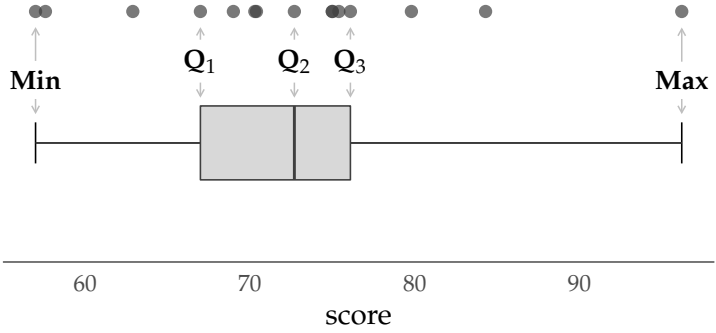


Figure 13.8: Boxplot for the `score` variable. The box boundaries represent the first and third quartiles, so the with of the box spans the interquartile range. The line in the middle of the box indicates the median. The long "whiskers" on either side of the box reach to the **Min** and **Max** values.

Lines extending outward from the box edges are called *whiskers*. The ends of the whiskers are marked with perpendicular lines which can be placed at different values depending the the type of box plot, in Figure 13.8, they reach to the maximum and minimum values.

The box plot is one of the most common visual summary plots that are used to plot data. We don't see the individual data points anymore, but the position the **Min**, the quartiles $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$, and the **Max** value provide an excellent overview.

Showing outliers using Tukey-style boxplots

Another type of box plot, shown in Figure 13.9, characterizes certain observations that we call *outliers*. Outliers are observations that are extremely high or low compared to the other data points. A common measurement is that x is an outlier if $x < \mathbf{Q}_1 - 1.5 \cdot \mathbf{IQR}$ or $x > \mathbf{Q}_3 + 1.5 \cdot \mathbf{IQR}$. Note there are several ways to define an outlier, but we'll use the "more than 1.5 times the interquartile-range away from last quartile" definition for the plots in this section.

Outliers are important because they can have disproportionate influence on some summary statistics and statistical analyses. In the `score` data, one student's `score` (the value 96.2) is an outlier because it is much higher than the other `scores`. This student performed unexpectedly better than the rest of the students.

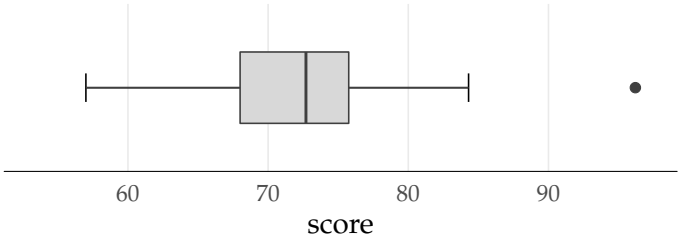


Figure 13.9: A Tukey boxplot for the `score` variable. In a Tukey boxplot the whiskers span from the the smallest observation that's still above $\mathbf{Q}_1 - 1.5 \cdot \mathbf{IQR}$, and the largest observation that's still below $\mathbf{Q}_3 + 1.5 \cdot \mathbf{IQR}$. Any points beyond the whiskers are drawn with a dot and considered an outlier.

In the boxplot shown in Figure 13.9, the whiskers reach from the highest and lowest values within the range $[\mathbf{Q}_1 - 1.5 \cdot \mathbf{IQR}, \mathbf{Q}_3 + 1.5 \cdot \mathbf{IQR}]$. Any observations that fall outside the whiskers are considered outliers and are presented with a dot. This is called a *Tukey boxplot* or *Spear-Tukey boxplot* in reference to Mary Eleanor Spear and John Tukey, who were the first to introduce this type of visual summary.

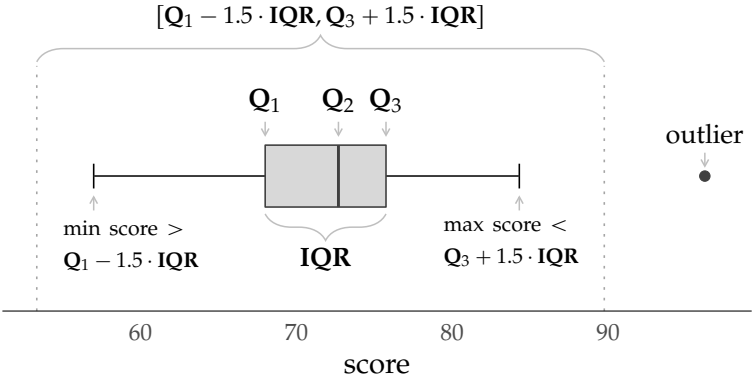


Figure 13.10: Boxplot for the `score` variable with additional labels for all the important quantities represented in the plot.

The Tukey boxplot (Figure 13.9) provides us with a better visual summary for the `score` data compared to the basic box plot (Figure 13.8), since it more accurately represents the where the values are clustered. Because there are different methods of drawing box plot whiskers, it's important to describe the convention used in the boxplot's caption as we have done in figures in this section.

13.3.4 Characteristics of numerical data

Let's review the data summarization techniques we've seen thus far, and categorize them according to the four different data characteristics that they capture: central tendency, dispersion, position, and shape.

Table 13.8 contains a summary of all the numerical summary statistics we computed for the `score` data. This compact table of values is a succinct way to report the most important characteristics of this variable.

statistic	value	measurement of
n	15	Number of observations
Mean	72.6	Central tendency
Med	72.7	Central tendency
Mode	70 - 80	Central tendency
Var	92.9	Dispersion
Std	9.6	Dispersion
Range	39.2	Dispersion
IQR	9.1	Dispersion
Min	57.0	Position
Max	96.2	Position
$\mathbf{Q}_1$	67.0	Position
$\mathbf{Q}_3$	76.1	Position

Table 13.8: Table of numerical summary statistics for the `score` variable.

Central tendency The median, mean, and mode are all measures that describe where the centre of the data is. Central tendency tells us about the "typical" value of a variable. It describes the approximate middle of a variable, where values tend to clump together. The **Mean** describes the average value, the **Med** describes the middle value (in a sorted list), while the **Mode** describes the most common value.

In the case of the `score` data, all the measures of central tendency have a similar value: mean is **Mean** $= 72.6$, the media was almost the same **Med** $= 72.7$, and **Mode** is in the 70–80 range. If you were a student in this class and scored 73 on your assessment, you'd know from these statistics that your performance was fair and and around the middle compared to your peers.

Dispersion The characteristic that describes how far values tend to diverge from the central point is called *dispersion*. The range (**Range** $=$

39.2) quantifies dispersion because it tells us how wide of a interval the data cover. The variance (**Var**) and standard deviation (**Std**) are both measures of dispersion relative to the mean **Mean**. The standard deviation was **Std** $= 9.6$ in this class, which indicates that the `scores` vary from one another by about 9.6 points on average.

Position The **Min**, **Max**, quartiles ($\mathbf{Q}_1$, $\mathbf{Q}_2$, $\mathbf{Q}_3$), and outliers we identified are examples of *position*. These statistics give us the location of specific values in the dataset compared to other values. If you were the student that scored 76.1, the value of $\mathbf{Q}_3$, you'd know that you scored better than about three quarters of your class. If you scored 96.2, then you'd know you did the best.

Shape A fourth characteristic of any numerical variable we can look at is its *shape*. The main tools for describing the shape of data are *skew* and the *modality*, which we can observe from a histogram.

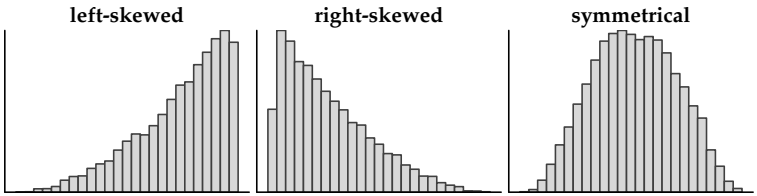


Figure 13.11: Histograms displaying distributions with three different *skews*. When higher values are more common, the data is *left-skewed*. If lower values are more common, the data is *right-skewed*. When neither higher nor lower values are more common, the data is *symmetrical*.

Figure 13.11 shows examples of histograms with different skews. Figure 13.12 shows examples with different modalities. Modality has to do with how many "peaks" a histogram has.

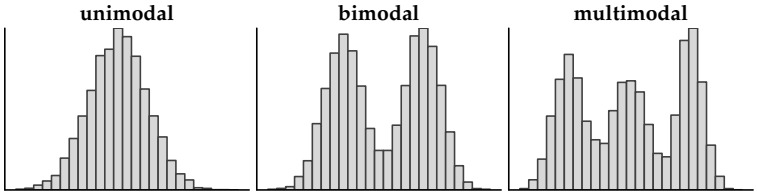


Figure 13.12: Histograms displaying distributions with three different *modalities*. Distributions with only one peak in are called *unimodal*. If we see two peaks in the histogram, we say the distribution is *bimodal*. Distributions with more than two peaks are called *multimodal*.

Data shape characteristics like skewness and modality are more qualitative than quantitative, but still important for better understanding of the dataset. You must be aware if you're dealing with multimodal or skewed data in order to choose appropriate statistical analysis procedures in later chapters.

13.3.5 Exercises

E13.1 Compute the numerical summary statistics for the `effort` variable that includes **Mean**, **Min**, **Max**, and **Range**.

E13.2 Compute the numerical summary statistics for the `effort` variable that includes $\mathbf{Q}_1$, **Med**, and $\mathbf{Q}_3$.

E13.3 Make a one-way frequency table for the `effort` variable. Use $(5, 7]$, $(7, 9]$, $(9, 11]$, $(11, 13]$ as the boundaries for bins.

13.4 Categorical data

Let's now look at the `background` variable of the student participants

```
arts, science, arts, ..., business, arts, science, science, arts
```

This is a categorical variable, meaning that `background` can take on a finite set of possible values (`arts`, `science`, or `business`).

We can plot categorical data using a bar chart. In a bar chart, each rectangle (or "bar") belongs to a category. The height of the bar represents a numeric measurement within the given category, such as a frequency (count) or a relative frequency (proportion). Unlike a histogram, the width of the bars has no meaning.

For categorical variables, there are fewer summary stats to calculate than for numerical variables. The most common are frequencies, relative frequencies, and mode.

Similar to numerical variables, frequencies for categorical variables are computed by counting the number of observations within each category. For example, the frequncy of students with an arts `background` is $\mathbf{Freq}_{\texttt{arts}} = 5$.

Relative frequencies are the fraction of the total number of observations for each category. If $\mathbf{Freq}_A$ is the frequency of category A, then the relative frequency is $\mathbf{Freq}_A/n$, the frequency divided by the number of observations. Relative frequencies can be expressed as a proportion (e.g. 0.5) or as a percentage (e.g. 50%). We can display these values in a one-way table.

For a discrete variable, the mode is the category with the most observations. The mode of the `background` variable is `science`, since

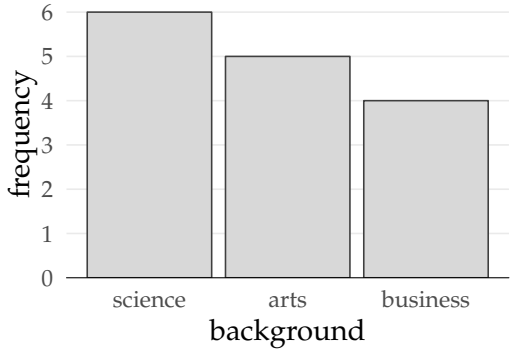


Figure 13.13: Bar chart showing the number of students per `background`. A bar chart is a visual representation of the number of observations in each category. The height of each rectangle represents the number students of each `background`.

background	frequency	relative frequency
arts	5	0.33
business	4	0.27
science	6	0.40

$\textbf{Freq}_{\texttt{arts}}/n = 5/15$

Figure 13.14: TODO: add caption

science was the most common `background` among the students that participated. Note that a variable could have more than one mode when there is a "tie" for the most common value.

13.4.1 Exercises

E13.4 Make a bar chart displaying the frequency of observations for the lecture vs. debate style classes in the `curriculum` variable.

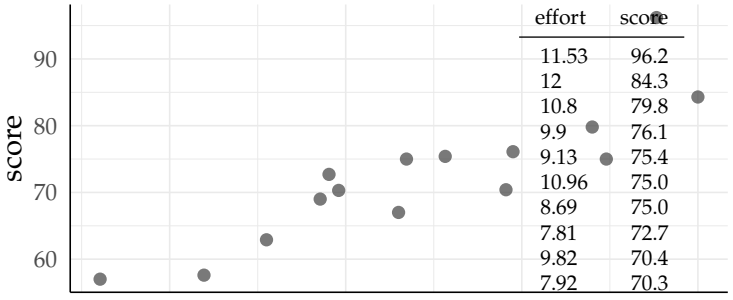
E13.5 Compute frequencies and relative frequencies for the `curriculum` variable. Display the results in a one-way table.

E13.6 What is the mode for `curriculum`?

13.5 Comparing two numeric variables

When exploring a dataset, we can also look at one variable's *association* with another. For example, we can look at the association between the `effort` and `score` variables. Let's use descriptive statistics examine whether `score` values tend to increase or decrease in relation to `effort`.

We can use a scatter plot to get a visual sense of association between two numerical variables, Figure 13.15 shows a scatter plot of `effort` vs. `score`.



effort	score
11.53	96.2
12	84.3
10.8	79.8
9.9	76.1
9.13	75.4
10.96	75.0
8.69	75.0
7.81	72.7
9.82	70.4
7.92	70.3
7.71	69.0
8.6	67.0
7.1	62.9
6.39	57.6
5.21	57.0

Figure 13.15: A scatter plot showing the association between `score` and `effort`. The tables of values is shown on the right for reference (the last two columns of Table 13.4). The dot is placed with horizontal position determined by the value of the first variable (x-axis) and vertical position determined by the second variable (y-axis).

In a scatter plot, two numerical values are mapped to locations that correspond to x and y coordinates, as shown in Figure **??**. If there is an associated between two variables, dots on the scatter plot will show a pattern. In Figure 13.15, the dots seem to scatter around an invisible line that points diagonally upward. This pattern indicates that higher `effort` values are associated with higher `score` values.

We can compute a numeric measure of association between the variables using the concepts of *covariance* and *correlation*. Covariance, denoted as $\mathbf{cov}(x, y)$, is the joint variability of two variables. If `effort` is x and `score` is y, then the covariance between these two variables is

$$\begin{aligned}\mathbf{cov}(x,y) &= \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \mathbf{Mean}_x)(y_i - \mathbf{Mean}_y) \\ &= \frac{(10.96-8.9)(75.0-72.6)+(8.69-8.9)(75.0-72.6)+\cdots+(12.00-8.9)(84.3-72.6)}{n-1} \\ &= 15.96 * 15/14 FIXME\end{aligned}$$

Because the result is a positive number, we say that there is a *positive association* between the `score` and `effort` variables. This means that students who put in more hours on the learning platform also performed better on the assessment. This corroborates what we saw in the scatter plot. If we had obtained a negative number instead, it would indicate a *negative association*, meaning that students who put in more `effort` tended to have lower `scores`. A zero covariance value would suggest that there is no relationship between the two variables, or at least no simple linear relationship.

Covariance is *not* a standardized measure; the units of measurement are the units for x multiplied by the units for y and values can range from $-\infty$ to $+\infty$. For this reason, covariance isn't a good measure of relationship *strength*; its value depends on the magnitude of the two variables. In other words, if either x or y have high variance, then the value of $\mathbf{cov}(x, y)$ will also be high, regardless of whether the association between the two variables is strong or weak.

Instead, we can look at *correlation* to judge the strength of the association. The correlation coefficient is denoted $\mathbf{corr}(x, y)$ and consists of a standardized version of covariance. It measures the degree of linear association between two variables. The correlation coefficient is a dimensionless quantity that ranges from -1 to $+1$. The closer the value is to -1 or 1, the stronger the relationship.

The correlation between `scores` and `effort` is

$$\mathbf{corr}(x,y) = \frac{\mathbf{cov}(x,y)}{\mathbf{Std}_x\,\mathbf{Std}_y} = \frac{15.96}{(9.6)(1.9)} = 0.875.$$

A correlation value of 0.875 indicates a high degree of correlation, meaning that `score` was pretty closely associated with `effort`.

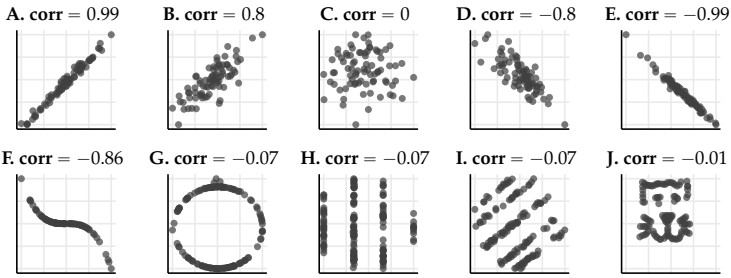


Figure 13.16: A correlation value close to 1 or -1 indicates that two variables *may* have a linear association, as shown in plots **A**, **B**, **D**, and **E**. A value of 0 indicates that there *may* be no relationship, as shown in plot **C**. However, variables may have a value close to 1 or -1 and *not* follow a linear relationship, as shown in plot **F**. They may also have a correlation of close to zero but show a strong, non-linear pattern, like plots **G**,**H**,**I**, and **J**.

Keep in mind the correlation coefficient is a very limited tool for describing the association between two variables. The correlation coefficient indicates a potential for a *simple linear association*, which might not always be a good fit for your data. Even if we find a high association between two variables, this doesn't necessarily mean that the variables follow a linear pattern. See the examples in Figure 13.16.

Even if there is a strong linear association, we cannot use descriptive statistics to say that one variable *causes* the other. Two variables can occur at higher values together without one having a direct influence on the other. The maxim "**correlation does not imply causation**" is fundamental concept in statistics. In this case, we cannot conclude that more effort lead to higher `scores`. It's equally plausible, for example, that some unaccounted for variable leads some students to both put in more effort and perform better on the assessment. Maybe some students were more interested in the subject matter to begin with, which motivated them to get high scores, and their higher interest also made them invest more hours of effort. To say anything about causation, we need carefully designed experiments and more involved statistical analyses.

13.5.1 Exercises

E13.7 Consider the following dataset of (x, y) pairs: $\{(2,5), (2,4), (3,3), (4,2), (5,2), (6,3), (7,4), (8,6), (4,8), (6,8)\}$. Calculate correla-

tion between x and y, then draw a scatter plot.

13.6 Comparing two categorical variables

The students that took part in this experiment decided for themselves which `curriculum` format they preferred to follow, lecture or debate. The teacher wants to see whether or not there was an even distribution of academic `backgrounds` in each of the two `curriculums` variants.

To visualize the relationship between two categorical variables, we can make a stacked bar chart. In a stacked bar chart, a rectangle for each category in one variable is made up of smaller blocks that represent a second variable. Figure 13.17 is a stacked bar plot that shows the frequencies of each of the two `curriculum` types within each of the three academic `backgrounds`. For example, there are four students with `science` backgrounds who are enrolled in the `lecture` curriculum. The height of the corresponding block spans four units on the y-axis.

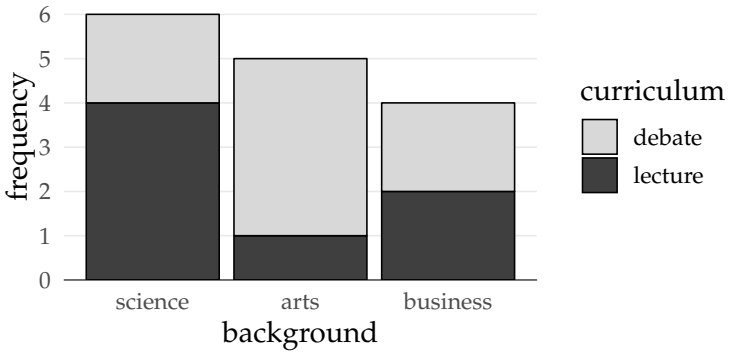


Figure 13.17: A stacked bar chart showing the distribution of `curriculum` choices by academic `background`.

From the stacked bar chart, it appears that more `arts` students chose the `debate` curriculum, while `science` students seemed to prefer the `lecture` style curriculum. The `business` students were evenly split between the two types of curriculum.

Remember, because "**correlation does not imply causation**," we don't know whether the students' `backgrounds` influenced their decision of `curriculum`. Their choice could have been completely random! With descriptive stats, we can observe a pattern but we can't give an explanation for it.

We can also display bivariate categorical frequencies in a *two-way table*. A two-way table shows the observed frequency of every combination of categories from the two variables, as well as the totals for each category.

	background			
curriculum	arts	business	science	TOTAL
lecture	1	2	4	7
debate	4	2	2	8
TOTAL	5	4	6	15

Freq`science` students in `lecture` class

$\textbf{Freq}_{\texttt{lecture}}$ = row sum $= 1 + 2 + 4$

$\textbf{Freq}_{\texttt{science}}$ = column sum $= 4 + 2$

A two-way table can also show relative frequencies.

	background			
curriculum	arts	business	science	TOTAL
lecture	0.07	0.13	0.27	0.47
debate	0.27	0.13	0.13	0.53
TOTAL	0.33	0.27	0.40	1.00

$\frac{\textbf{Freq}_{\texttt{science+lecture}}}{n} = \frac{4}{15}$

$\frac{\textbf{Freq}_{\texttt{lecture}}}{n} = \frac{7}{15}$

$= \text{row sum}$

$= 0.07 + 0.13 + 0.27$

$\frac{\textbf{Freq}_{\texttt{science}}}{n} = \frac{6}{15}$

$= \text{column sum}$

$= 0.27 + 0.13$

Table 13.13: TODO: add caption

Relative frequency tables for bivariate data can be misleading. For instance, the proportion of business and science students that chose the debate `curriculum` were both 0.13. A viewer could wrongly interpret this to mean that both groups had the same preference for that style of learning. To compare `curriculum` preference between the three academic `backgrounds`, we need to look at *conditional relative frequencies*. Instead of dividing the number of observations per category combination by the total number of observations, we divide it by the number of observations for the *variable of interest*, in this case, academic `background`.

	background			
curriculum	arts	business	science	TOTAL
lecture	0.20	0.50	0.67	0.47
debate	0.80	0.50	0.33	0.53
TOTAL	1.00	1.00	1.00	1.00

Table 13.15: TODO: add caption

The conditional relative frequencies for columns reveal the different preferences between student `backgrounds`. To show frequencies that are conditional on *curriculum*, we divide each frequency by the total observation *per row*.

$\frac{\text{Freq}_{\texttt{science+lecture}}}{n} = \frac{4}{15}$

$\frac{\text{Freq}_{\texttt{lecture}}}{n} = \frac{7}{15}$

$= \text{row sum}$

$= 0.07 + 0.13 + 0.27$

$\frac{\text{Freq}_{\texttt{science}}}{n} = \frac{6}{15}$

$= \text{column sum}$

$= 0.27 + 0.13$

	background			
curriculum	arts	business	science	TOTAL
lecture	0.14	0.29	0.57	1.00
debate	0.50	0.25	0.25	1.00
TOTAL	0.33	0.27	0.40	1.00

Table 13.17: TODO: add caption

This table shows the relative composition of student `backgrounds` per `curriculum`. For example, the second row indicates the debate style class comprised half arts students, and a quarter business, and a quarter science students. Because these proportions are different from one another, we can say that there is some association between `background` and `curriculum` in our data. The larger the difference, the stronger the association.

To visualize association, we can make a stacked bar chart using conditional relative frequencies instead of absolute frequencies.

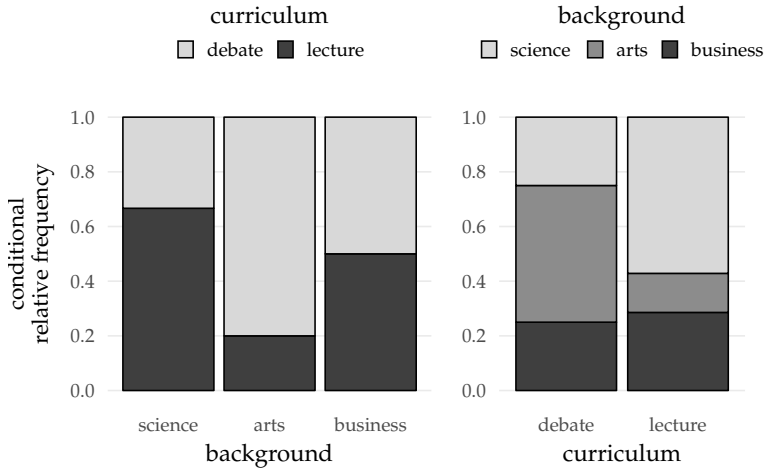


Figure 13.18: Stacked bar charts showing relative frequencies conditional on curriculum (left) and background (right).

13.6.1 Exercises

E13.8 given the following data, make a frequency, relative frequency, and two conditional relative frequency tables.

E13.9 question testing concept of correlation != causation

13.7 Comparing a categorical and a numerical variable

We can reuse all the numerical and visual summaries know for describing a single numeric variable (Section 13.3), when studying the relationships between a numeric variable and a categorical variable. We follow the usual procedures for computing numerical summaries, and drawing strip charts, histograms, and box plots by repeating them for each of the possible values of the categorical variable.

Suppose the teacher wants to see how student scores (numerical) compare between the two curriculum variants (categorical). One way to compare a categorical and a numerical variable, is to make a table of summary statistics within categories.

	score				
curriculum	min	median	max	mean	std
lecture	57.00	70.30	75.40	68.14	7.18
debate	62.90	75.55	96.20	76.46	9.84

Table 13.19: This table shows that for the `debate` curriculum, scores were overall higher but also more variable than for the `lecture` variant.

Let's visualize this difference with a paired strip chart, histogram, and box plot. These three types of numerical plots can be placed side-by-side to compare a numeric variable within categories, as shown in Figure 13.19.

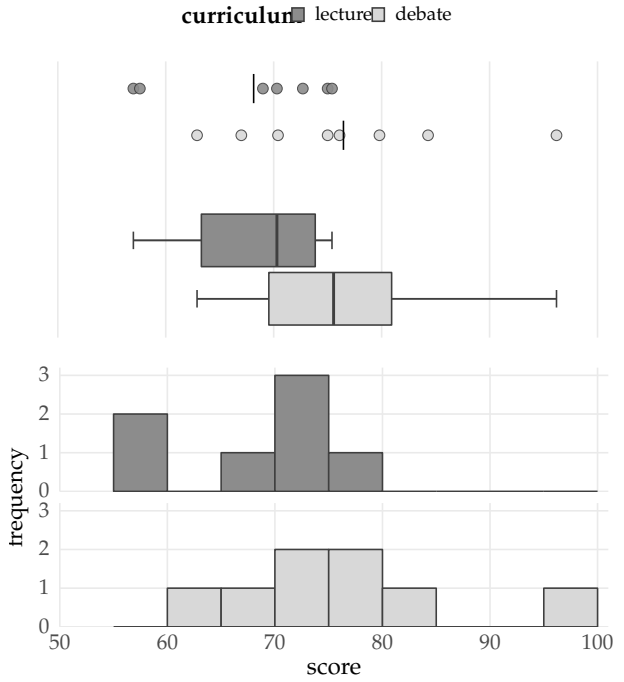


Figure 13.19: Strip charts, histograms, and box plots comparing students' `scores` in the debate vs. lecture style class. The vertical lines in the strip charts indicate the positions of the means: **Mean**(`lecture`) = 68.14 and **Mean**(`debate`) = 76.46. The boxplots are Tukey boxplots.

We can also compare debate and lecture `scores` with a bar chart. We draw one bar for debate and one for lecture, then set the height of

each bar to the mean `score` within the respective `curriculum` type. To give viewers an idea about dispersion within each category, we can also add *error bars* I . Error bars represent the variability of data. In Figure 13.20, the error bars stretch from **Mean** − **Std** to the **Mean** + **Std**.

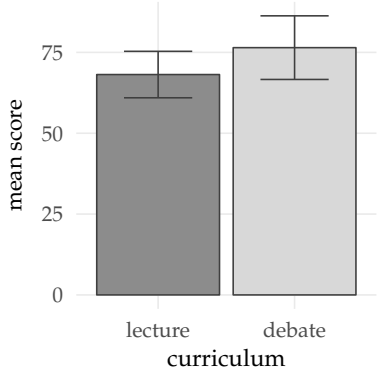


Figure 13.20: Bar charts can be used to compare a numerical value between two categories, such as the mean `score` between debate and lecture styles `curriculums`. Error bars stretching from **Mean** ± **Std** are a common way of showing how reliable the mean is as a representative value.

13.7.1 Exercises

E13.10

13.8 Explanations

13.8.1 Histogram binning

To visualize the distribution of `score` data as a histogram, we divided data into five convenient bins, each 10 units wide. If we had chosen fewer or more bins, our histogram would have appeared very different.

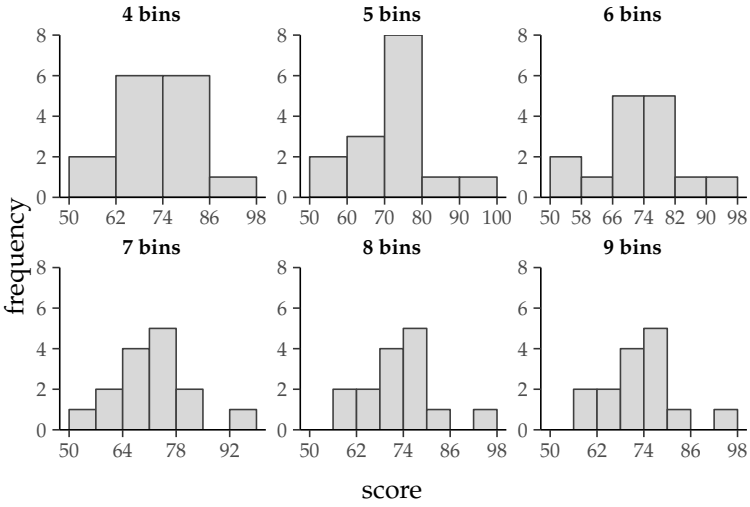


Figure 13.21: The same `score` data displayed as six histograms with varying number of bins. The number and width of bins you choose will change the appearance of your histogram.

If you choose fewer bins, the bins will be wider. The wider the bin, the less detail the histogram shows. In contrast, if you use more bins, the bins will be narrower. However, bins that are too narrow show too much detail and distracts from the overall shape of the data. Generally, you can afford to have narrower bins when you have a larger **n**.

The start point of each bin also impacts the overall shape of the histogram, even when the number and width of each bin is the same.

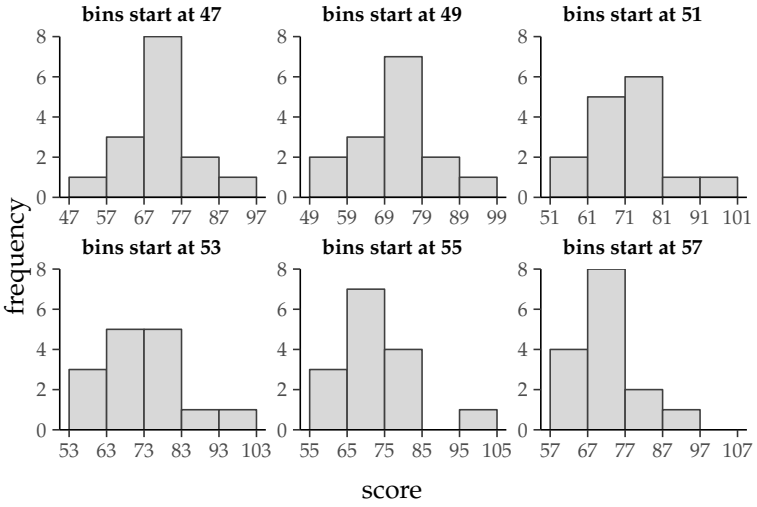


Figure 13.22: The same `score` data displayed as six histograms each with five bins of width 10, but with bin boundaries starting at a different number.

Using integers for bin boundaries makes a histogram easier to interpret. In terms of the number of bins, there are formulas that you may use as a guideline choosing the correct amount, k, such as $k = \sqrt{n}$. Regardless, you should experiment by making histograms with varying numbers of bins and varying bin boundaries, and use your judgement. A strong pattern in the shape of your data will show up in different histograms with varying bin widths.

13.8.2 Dealing with outliers

Outliers are observations that are so much larger or smaller than the other data points. These points can have undue leverage on the value of your summary statistics. Statistics that are calculated using all of the data points can be pushed to much higher or lower values because of just one or two exceptional values. For example, the mean of the dataset $x = \{x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 50\}$ is $\textbf{Mean}_x = 14$. This is very misleading since the values x_1, x_2, and x_3 are small numbers, but the presence of the much large outlier x_4 pushes the mean to a very large value. If we remove the outlier x_4 the mean drops drastically:

$$\textbf{Mean}_{x_1,x_2,x_3} = \frac{1}{3}(1 + 2 + 3) = 2.$$

Because outliers can completely change the conclusions your draw from your data, statisticians tend to pay a lot of attention to them.

Sometimes these extreme observations result from mistakes in data collection, like when a tool used to make measurements malfunctions, a typo is made in entering data, or when the wrong subject is measured. For example, when you intended to study dogs but mistook a grizzly bear for a Tibetan mastiff. These mistakes are usually identified when a measurement is obviously impossible, like a 1600 lbs dog. In these cases, you should first try to fix the mistake (for example, if the data was transcribed from paper to computer, see if you can find what was originally written down and fix the typo). Otherwise, it's fair to remove these faulty observations from your analyses.

Some statisticians argue that outliers should be excluded even if they aren't mistakes. They reason that because outliers skew results so drastically, those results won't be generalizable. Consider though that outliers may indicate an unexpected phenomenon, or a previously unsuspected variable that influences the measurement. It would be a shame to just discard a potentially valuable finding.

Instead of rejecting the validity of these observations, you could instead investigate the unusual cases more closely and look for an explanation. Alternatively, you could use a statistical procedure that gives less weight to outliers, or you could repeat the experiment to obtain a new dataset and compare your two sets of results. Finally, you could choose to report results both with and without outliers and let your audience decide.

Whatever strategy you prefer, it's best to decide how you will deal with outliers *a priori*—before you collect your data. If you do end up eliminating a data point from your analyses (for any reason), always report what you removed and why you removed it.

13.9 Discussion

13.9.1 Choosing stats and plots

We learned about four characteristics of data that we can express with descriptive statistics: central tendency, dispersion, shape, and position. Each characteristic can be visualized using several statistics and plots. Which you choose depends on the purpose of your analyses as well as on the data itself.

When it comes to choosing summary statistics for central tendency, here are some general guidelines:

- Use the **mean with standard deviation** (most common) when

  Data are symmetrical; and

  There are no outliers;

- Use **median along with the interquartile range** when

  There are outliers; or

  Data is skewed;

  Data is ordinal; or

  Some data is missing

In terms of displaying the distribution of your data using the plots we've discussed so far:

- Stripcharts are best for small datasets with non-overlapping points.
- Use histograms to show the shape of a distribution. Histograms are better than boxplots for displaying the details of a distribution, especially one with very high or low variance.
- Use a **boxplot** to show the exact position of the median, quartiles, and outliers. Boxplots beat histograms for quick comparisons between different distributions.

Scatter plots (for numerical data) and stacked bar charts (for categorical data) are two standard visuals for comparing two variables, though other bivariate plots exist (see Section 13.9.4).

Whenever possible, generate a few different types of plots and summary statistics to get a more comprehensive understanding of your data.

13.9.2 Always plot your data

Summary statistics are a useful tool for getting a glimpse into patterns in your data, but don't trust them on their own. Wildly different datasets can have identical summary statistics, let us show you:

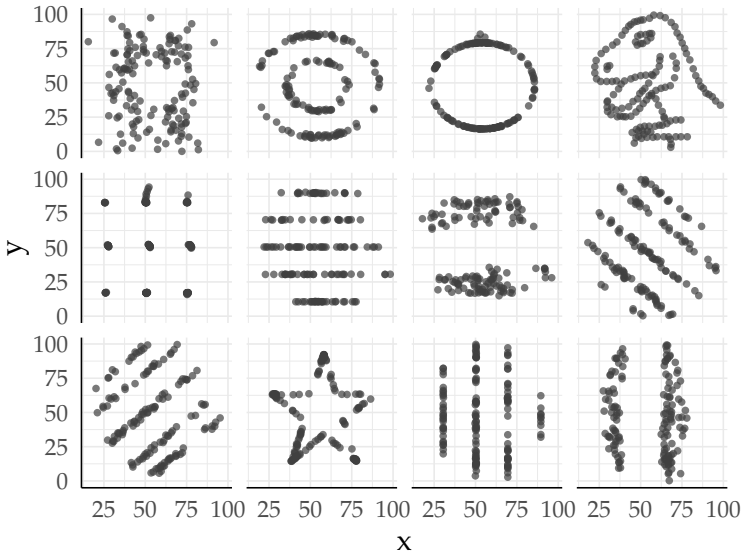


Figure 13.23: All the plots shown here have $\textbf{Mean}_x = 54.3$, $\textbf{Mean}_y = 47.8$, $\textbf{Std}_x = 16.8$, $\textbf{Std}_y = 26.9$, $\textbf{corr}(x, y) = -0.1$. Read more about how these datasets were created here `https://www.autodeskresearch.com/publications/samestats`

See more examples in the paper `https://www.autodeskresearch.com/publications/samestats`.

The lesson here is to always visualize your data.

13.9.3 Computing quartiles and percentile

With our data sorted in ascending order, the quartile $\mathbf{Q}_k$ is given by the formula:

$$\mathbf{Q}_k = x_{\lfloor i \rfloor} + (i - \lfloor i \rfloor)(x_{\lfloor i \rfloor + 1} - x_{\lfloor i \rfloor}),$$

where $k \in 1, 2, 3$, $i = \frac{k}{4}(n-1)$, and $\lfloor i \rfloor$ (the *floor* of i) is the greatest integer that's less than i. Note this formula uses linear interpolates between the values, so $\mathbf{Q}_k$ might not be values that are observations in your dataset. Remember the quartiles are fenceposts that divide the dataset into four chunks chunks.

Quartiles are just one of many ways of dividing up the distribution of our data. *Percentiles*, $\mathbf{P}_k$ are like quartiles but divide data into 100 chunks instead of four. $\mathbf{Q}_1 = \mathbf{P}_{25}$, $\mathbf{Q}_2 = \mathbf{P}_{50}$, and $\mathbf{Q}_3 = \mathbf{P}_{75}$. It only makes sense to compute percentiles when you have a large dataset.

13.9.4 More stats and more plots

In this chapter we introduced some fundamental summary stats and plots, but the world of descriptive statistics is vast!

The term *mean*, unless otherwise specified, refers to the *arithmetic mean*. That's the one we introduced to you in this chapter. We didn't even get a chance to explain the other types of means, like the *geometric mean* (good for variables that are multiplicative) or a *weighted mean* (useful when some data points are more important than others). There are also different ways to calculate correlation and covariance. We introduced the most common method, referred to as *Pearson correlation coefficient*, but other situations may call for different methods.

The options for plotting are even more varied. Graphical visualization of data is an ever evolving art/science, especially with the advent of video, 3D capabilities, and interactive tools. Clearly it's not possible to cover the entire rich ecosystem of data visualizations in a mere x pages. We encourage you to explore many different ways to visualize your data. As inspiration, here are a few other ways we could have visualized the education data

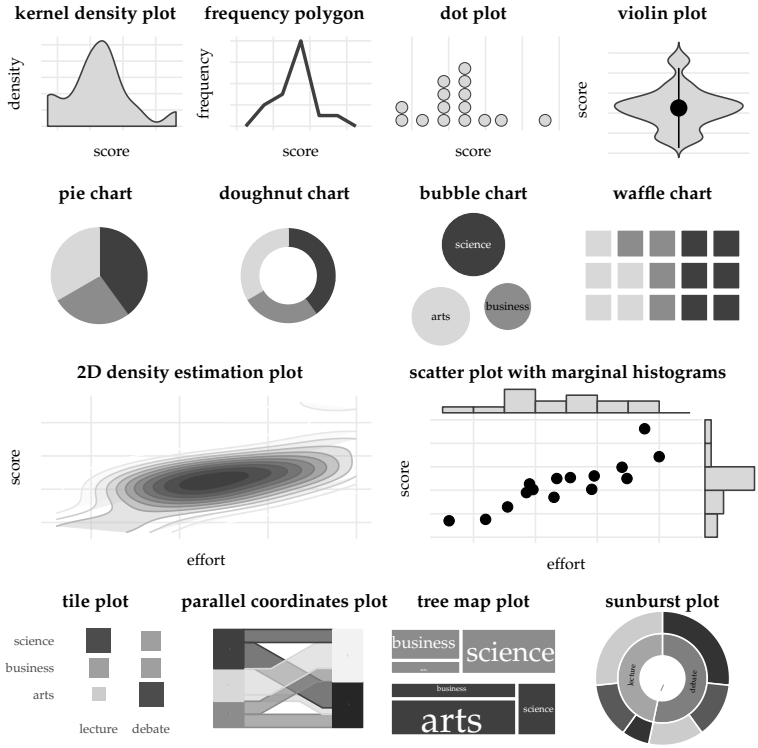


Figure 13.24: Examples of the multitude of data visualization options.

Here are some basic ground rules that will help you present your information efficiently:

- Avoid distorting your data: using 3D effects, for example, reduce interpretability.
- Minimize the ink to data ratio: only include elements that are necessary for conveying information.
- Use the complete axis: in most cases, it's best to start numerical axes at zero and use consistent intervals, rather than ...
- Avoid dual axis charts: a chart with two y axes can be misleading. Try plotting data side by side or use a single index scale instead.
- Use colour wisely: if you use colour to differentiate between categories, choose hues that are distinguishable and accessible.
- Keep your target audience in mind: ...

13.9.5 The importance of descriptive statistics

Doing descriptive statistics is like viewing an apartment before you sign a year long lease. Always get to know the basic qualities of your data before you commit to a long-term analysis. Let me make this clear: I don't care how excited you are to test out that fancy machine learning algorithm you just heard about. It doesn't matter if your super geeky stats friend already told you, you should definitely use a generalized linear mixed model with a poisson error structure, or whatever. *Always* start by plotting and summarizing your data. I promise this will make your life easier and potentially save you lots of future heartaches. You're welcome.

In this chapter, we learned how to summarize and compress data into easily interpretable summary statistics and plots. Later, we'll apply these summary statistics to answer questions about things we haven't measured directly. For that, we're going to need some background on probability.

13.9.6 Tidy data

Tidy data is... Each row represents a single observation and each column one variable (in other words, the data is tidy).

13.10 Links