Observation Studies, Sampling Strategies, and Experiments

Lt Col Ken Horton Professor Bradley Warner Lt Col Kris Pruitt

11 May, 2020

Objectives

1)

2)

Numerical Data

This lesson introduces techniques for exploring and summarizing numerical variables, and the `email50` and `county` data sets from the `openintro` package provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical.

Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In a Figure 1, we present again a scatterplot used to examine how federal spending and poverty were related in the `county` data set.

Another scatterplot is shown in Figure 2, comparing the number of line breaks `line_breaks` and number of characters `num_char` in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 2.

To put the number of characters in perspective, this paragraph has 363 characters. Looking at Figure 2, it seems that some emails are incredibly long! Upon further investigation, we would actually find that most of the long emails use the HTML format, which means most of the characters in those emails are used to format the email rather than provide text.

> Exercise What do scatterplots reveal about the data, and how might they be useful?[1]

> Example Consider a new data set of 54 cars with two variables: vehicle price and weight.[2] A scatterplot of vehicle price versus weight is shown in Figure 3. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen which show relationships that are very linear.

[1] Answers may vary. Scatterplots are helpful in quickly spotting associations between variables, whether those associations represent simple or more complex relationships.

[2] Subset of data from http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

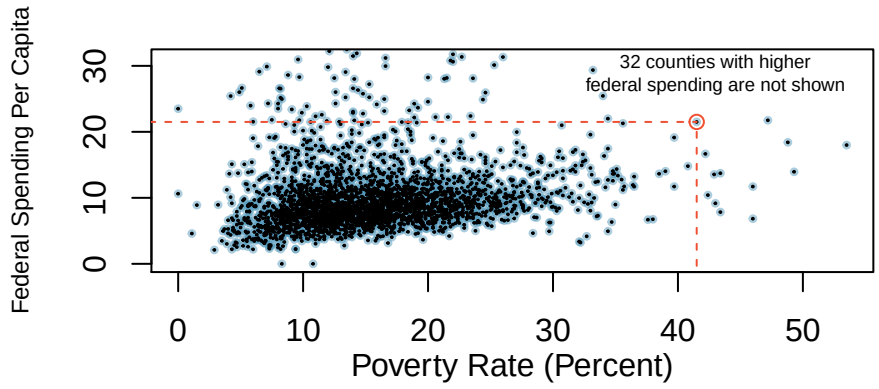


Figure 1: A scatterplot showing fed_spend against poverty. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

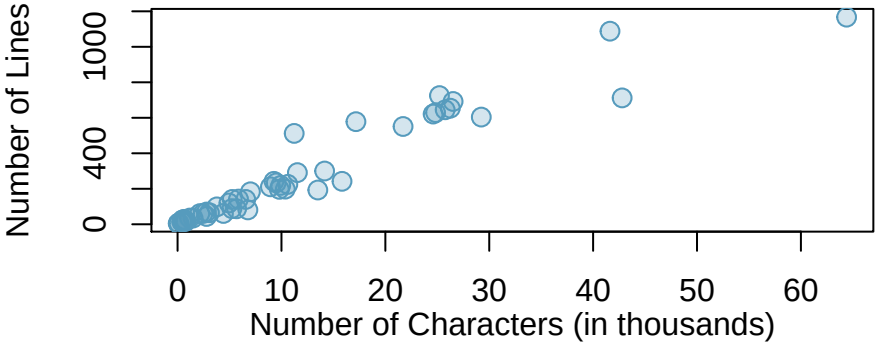


Figure 2: A scatterplot of *line_breaks* versus *num_char* for the `email50` data.

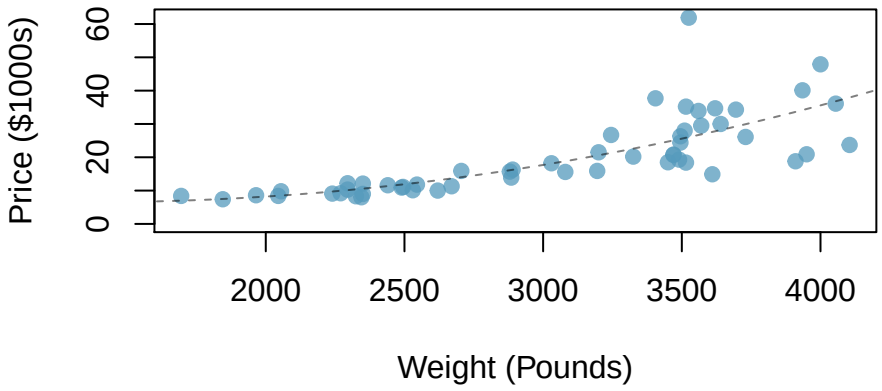


Figure 3: A scatterplot of *price* versus *weight* for 54 cars.

Exercise
Describe two variables that would have a horseshoe shaped association in a scatterplot.[3]

Dot plots and the mean

Sometimes two variables is one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the number of characters from 50 emails is shown in Figure 4.

The **mean**, sometimes called the average, is a common way to measure the center of a **distribution** of data. To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6$$

The sample mean is often labeled $\bar{x}$, and the letter x is being used as a generic placeholder for the variable of interest, `num_char`.

Mean The sample mean of a numerical variable is the sum of all of the observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

Exercise Examine the two equations above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[4]

[3] Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description since water becomes toxic when consumed in excessive quantities.

[4] x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the i^{th} email in the data set.

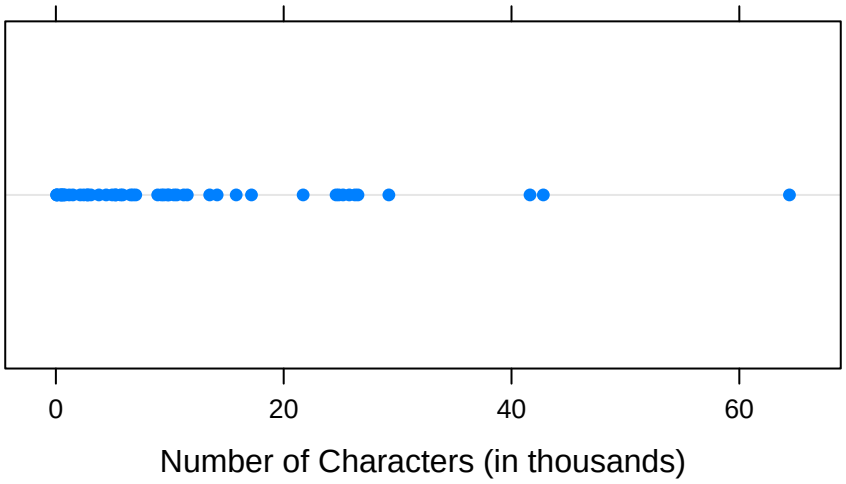


Figure 4: A dot plot of num_char for the `email50` data set.

Exercise What was n in this sample of emails?[5]

The `email50` data set is a sample from a larger population of emails that were received in January and March. We could compute a mean for this population in the same way as the sample mean. However, there is a difference in notation: the population mean has a special label: μ. The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x.

Example The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x, the mean number of characters in all emails in the `email` data set? (Recall that `email50` is a sample from `email`.)

The sample mean, 11.6, may provide a reasonable estimate of μ_x. While this number will not be perfect, it provides a **point estimate** of the population mean. Later in the semester, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

Example We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes from the 3,143 counties in the `county` data set. What would be a better approach?

The `county` data set is special in that each county actually represents many individual people. If we were to simply average across the `income` variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the `county` data, we would find that the per capita income for the US is $27,348.43. Had we computed the *simple* mean of per capita income across counties, the result would have been just $22,504.70!

This previous example used what is called a **weighted mean**, which will a key topic in the probability section. As a look ahead, the probability mass function gives the population proportions of each value and thus to find the population mean μ, we will use a weighted mean.

[5]The sample size was $n = 50$.

Histograms and shape

Dot plots show the exact value of each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin. This tabulation is shown below.

```
##
##   (0,5]  (5,10] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50]
##      19      12       6       2       3       5       0       0       2       0
## (50,55] (55,60] (60,65]
##       0       0       1
```

These binned counts are plotted as bars in Figure 5 into what is called a **histogram**.

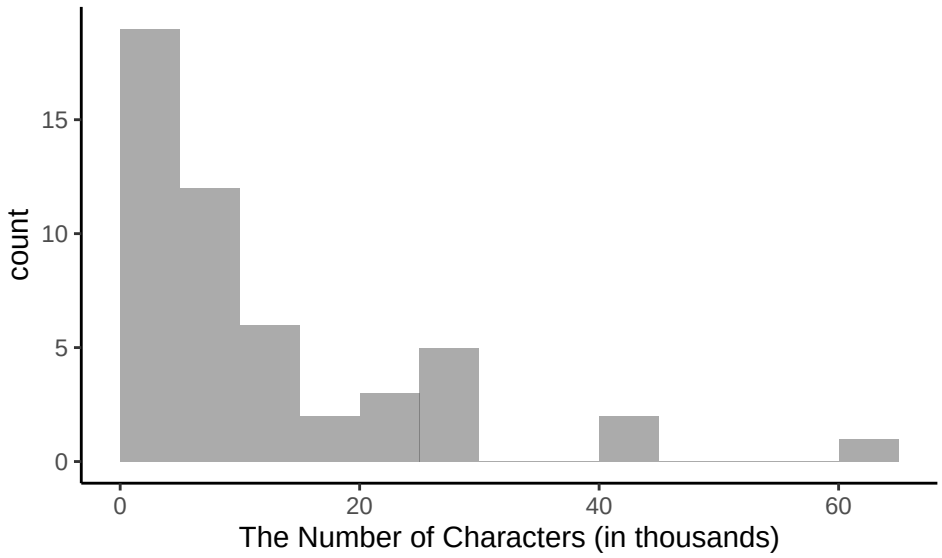


Figure 5: A histogram of `num_char`. This distribution is very strongly skewed to the right.

Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more dense. For instance, there are many more emails between 0 and 10,000 characters than emails between 10,000 and 20,000 characters in the data set. The bars make it easy to see how the density of the data changes relative to the number of characters.

Histograms are especially convenient for describing the shape of the data distribution. Figure 6 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right **tail**, the shape is said to be **right skewed**.[6]

Data sets with the reverse characteristic – a long, thin tail to the left – are said to be **left skewed**. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off in both directions are called **symmetric**.

[6]Other ways to describe data that are skewed to the right: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

Long tails to identify skew When data trail off in one direction, the distribution has a **long tail**. If a distribution has a long left tail, it is left skewed. If a distribution has a long right tail, it is right skewed.

Exercise Take a look at the dot plot above. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[7]

Exercise Besides the mean, what can you see in the dot plot that you cannot see in the histogram?[8]

Making our own histogram Let's take some time to make a simple histogram. We will use the `ggformula` package which is a wrapper for the `ggplot` package.

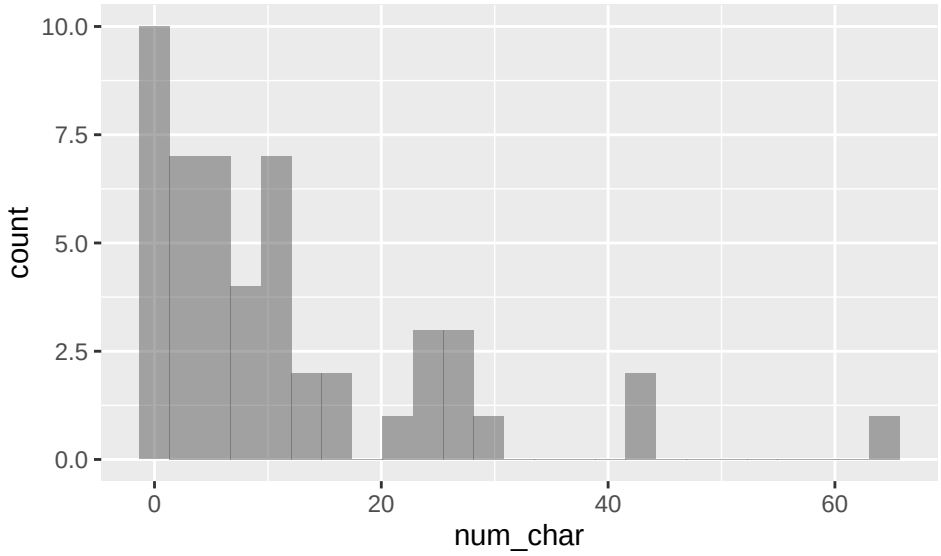
Here are two questions: *What do we want `R` to do?* and
What must we give `R` for it to do this?

We want `R` to make a histogram. In `ggformula` the plots have the form `gf_XXXX` so we will use the `gf_histogram`. To find options and more information type:

```
?gf_histogram
```

To start we just have to give the formulas and data to `R`.

```
gf_histogram(~num_char,data=email50)
```



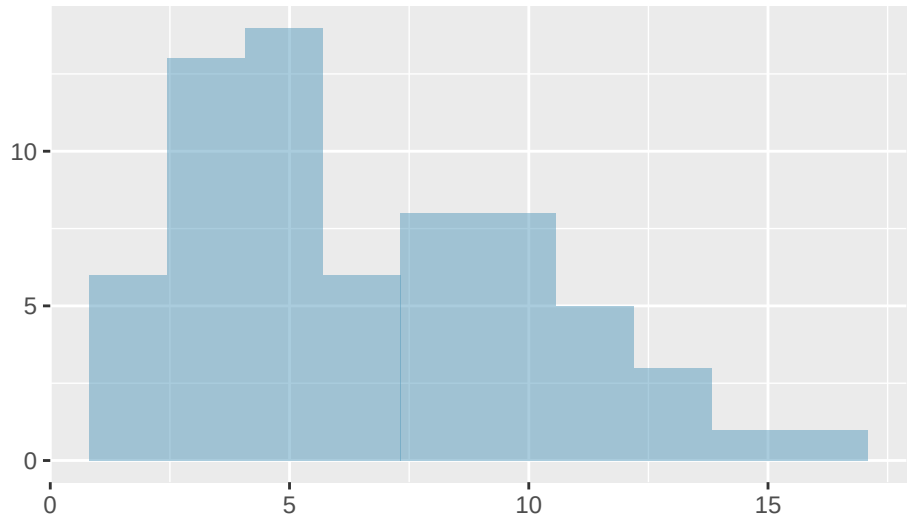
Exercise Look at the help menu for `gf_histogram` and change the x-axis label, change the bin width to 5, and have the left bin start at 0.

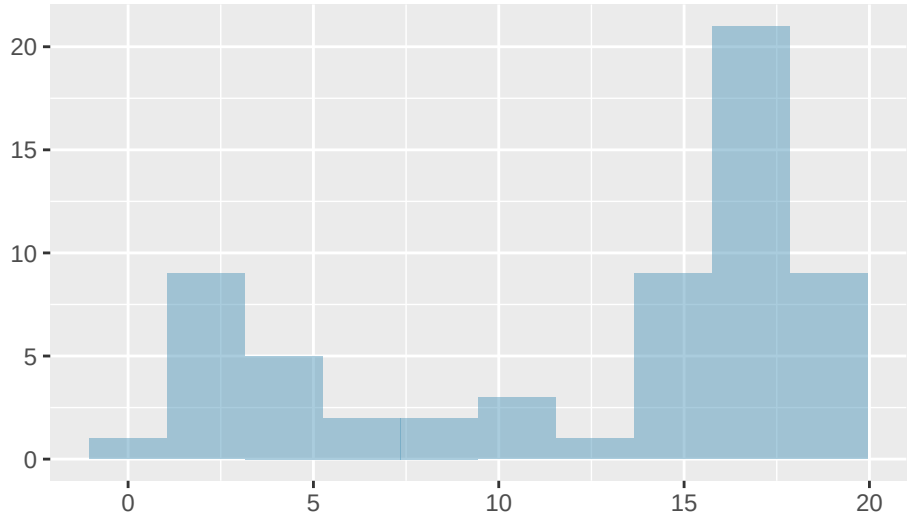
```
email50 %>%
   gf_histogram(~num_char,binwidth = 5,boundary=0,xlab="The Number of Characters (in thousands)") %>%
   gf_theme(theme_classic())
```

[7] The skew is visible in all both plots, though the dot plot is the least useful.
[8] Character counts for individual emails.

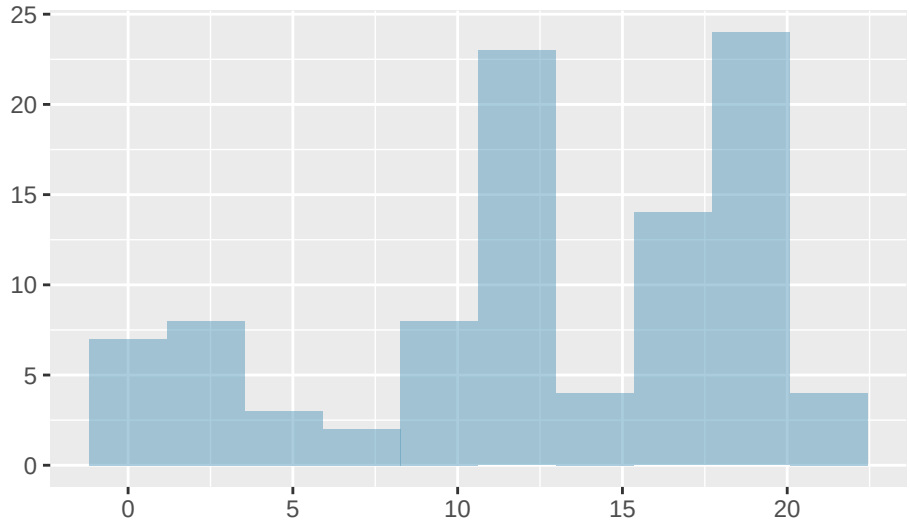
In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A **mode** is represented by a prominent peak in the distribution.[9] There is only one prominent peak in the histogram of `num_char`.

The next three figures show histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.





[9] Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real data sets.



Exercise Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[10]

Looking for modes Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why **prominent** is not rigorously defined in these notes. The important part of this examination is to better understand your data and how it might be structured.

Variance and standard deviation

The mean is used to describe the center of a data set, but the *variability* in the data is also important. Here, we introduce two measures of variability: the **variance** and the **standard deviation**. Both of these are very useful in data analysis, even though the formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to conceptually understand, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\begin{aligned} x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\ x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\ x_3 - \bar{x} &= 0.6 - 11.6 = -11.0 \\ &\vdots \\ x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2 \end{aligned}$$

If we square these deviations and then take an average, the result is about equal to the **sample variance**,

[10] There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal. But it could be multimodal because within each group we may be able to see a difference in males and females.

denoted by s^2:

$$\begin{aligned} s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44 \end{aligned}$$

We divide by $n-1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance yet. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

The **standard deviation* s is the square root of the variance:

$$s = \sqrt{172.44} = 13.13$$

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

> Variance and standard deviation The variance is roughly the average squared distance from the mean. The standard deviation is the square root of the variance and describes how close the data are to the mean.

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[11] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

> Tip: standard deviation describes variability Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as we have seen, these percentages are not strict rules.

> Exercise Earlier the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using the three figures as an example, explain why such a description is important.[12]

> Example Describe the distribution of the `num_char` variable using the histogram above. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.[13]

In practice, the variance and standard deviation are sometimes used as a means to an end, where the *end* is being able to accurately estimate the uncertainty associated with a sample statistic. For example, later in the course we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[11] The only difference is that the population variance has a division by n instead of $n-1$.

[12] Starting with Figure 6, the three figures show three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a~distribution.

[13] The distribution of email character counts is unimodal and very strongly skewed to the high end. Many of the counts fall near the mean at 11,600, and most fall within one standard deviation (13,130) of the mean. There is one exceptionally long email with about 65,000 characters.

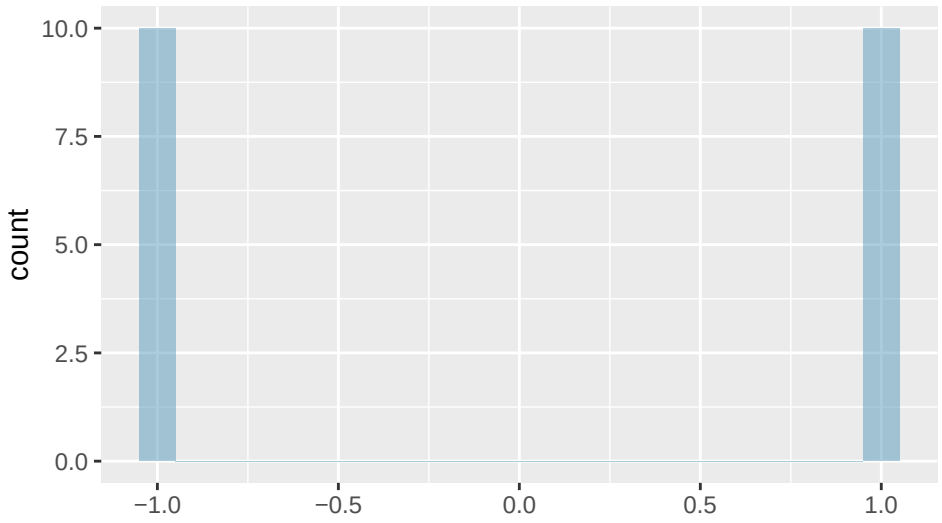


Figure 6: The first of three very different population distributions with the same mean, 0, and standard deviation, 1.

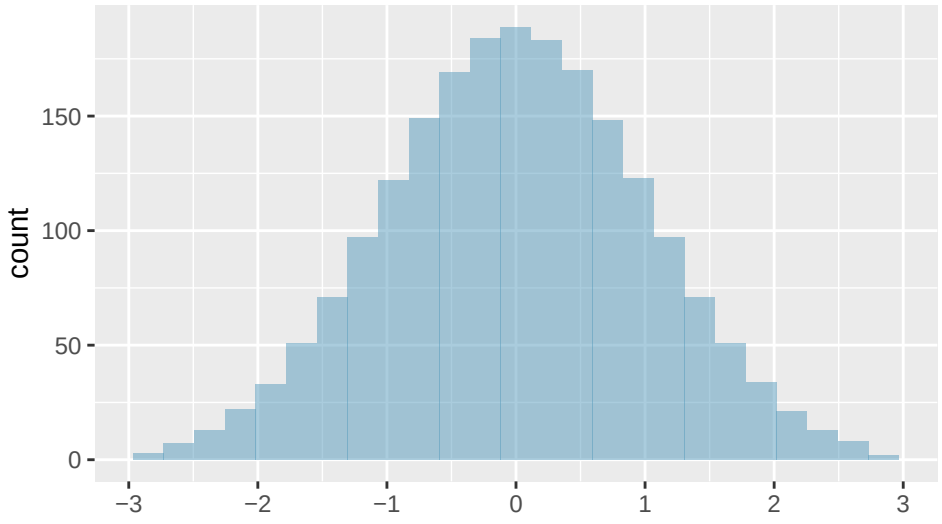


Figure 7: The second plot with mean 0 and standard deviation 1.

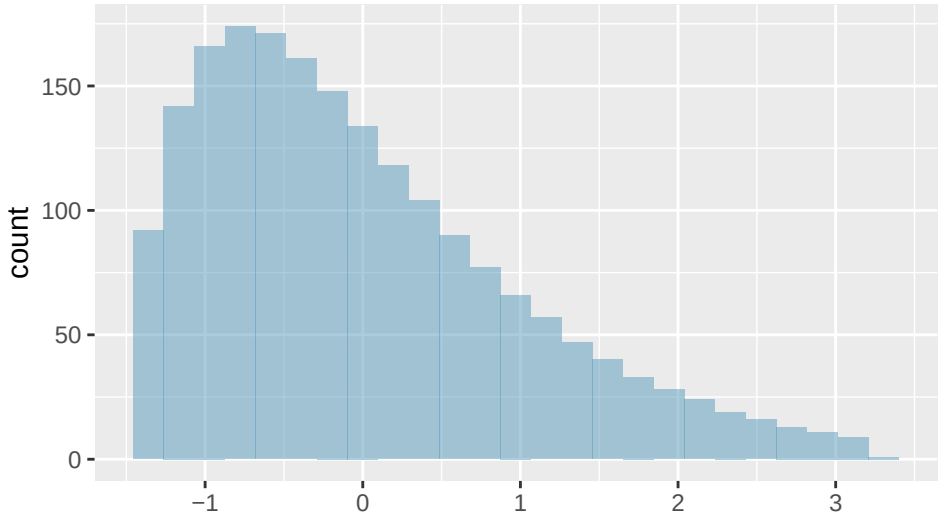


Figure 8: The final plot with mean 0 and standard deviation 1.

Box plots, quartiles, and the median

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 9 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 9 shows 50% of the data falling below the median (red dashes) and the other 50% falling above the median (blue open circles). There are 50 character counts in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50^{th} percentile: $(6{,}768 + 7{,}012)/2 = 6{,}890$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

> Median: the number in the middle If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 9, is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR. The two boundaries of the box are called the **first quartile** (the 25^{th} percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75^{th} percentile), and these are often labeled Q_1 and Q_3, respectively.

> Interquartile range (IQR) The IQR is the length of the box in a box plot. It is computed as
>
> $$IQR = Q_3 - Q_1$$
>
> where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

> Exercise What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[14]

[14]Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3.

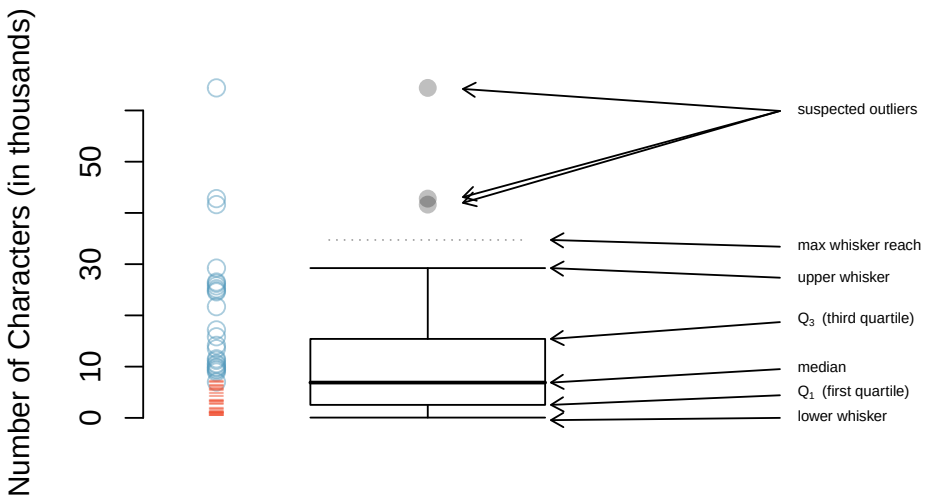


Figure 9: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

Extending out from the box, the **whiskers** attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.[15] They capture everything within this reach. In Figure 9, the upper whisker does not extend to the last three points, which are beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation that lies beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of just extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. In this case, it would be reasonable to classify the emails with character counts of 41,623, 42,793, and 64,401 as outliers since they are numerically distant from most of the data.

> Outliers are extreme An **outlier** is an observation that is extreme relative to the rest of the data.
>
> Why it is important to look for outliers Examination of data for possible outliers serves many useful purposes, including 1. Identifying **strong skew** in the distribution. 2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64,401 characters to ensure this value was accurate. 3. Providing insight into interesting properties of the data.
>
> Exercise The observation 64,401, an outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[16]

[15] While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.
[16] That occasionally there may be very long emails.

Exercise Using Figure 9, estimate the following values for `num_char` in the `email50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[17]

Of course R can calculate these summary statistics for us. First we will do these them individually and then in one function call. Remember to ask what you want R to do and what it needs.

```
mean(~num_char,data=email50)
```

```
## [1] 11.59822
```

```
sd(~num_char,data=email50)
```

```
## [1] 13.12526
```

```
quantile(~num_char,data=email50)
```

```
##       0%      25%      50%      75%     100%
##  0.05700  2.53550  6.88950 15.41075 64.40100
```

```
iqr(~num_char,data=email50)
```

```
## [1] 12.87525
```

```
favstats(~num_char,data=email50)
```

```
##    min     Q1 median       Q3    max     mean       sd  n missing
##  0.057 2.5355 6.8895 15.41075 64.401 11.59822 13.12526 50       0
```

Robust statistics

How are the *sample statistics* of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these *summary statistics* if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 10, and sample statistics are computed in R.

```
p1 <- email50$num_char
p2 <- p1[-which.max(p1)]
p3 <- p1
p3[which.max(p1)] <- 150

robust <- data.frame(value= c(p1,p2,p3),group=c(rep("Original",50),rep("Dropped",49),rep("Increased",50)

favstats(value~group,data=robust)
```

```
##       group   min     Q1 median       Q3     max     mean       sd  n missing
## 1   Dropped 0.057 2.4540 6.7680 14.15600  42.793 10.52061 10.79768 49       0
## 2 Increased 0.057 2.5355 6.8895 15.41075 150.000 13.31020 22.43436 50       0
## 3  Original 0.057 2.5355 6.8895 15.41075  64.401 11.59822 13.12526 50       0
```

[17] These visual estimates will vary a little from one person to the next: $Q_1 \sim 3{,}000$, $Q_3 \sim 15{,}000$, IQR=$Q_3 - Q_1 \sim 12{,}000$. (The true values: $Q_1 = 2{,}536$, $Q_3 = 15{,}411$, IQR $= 12{,}875$.)

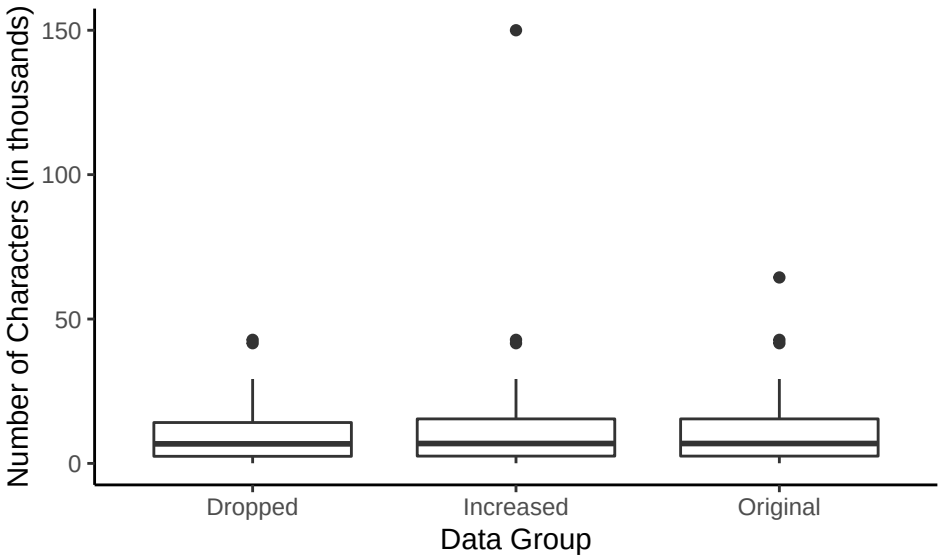


Figure 10: Box plots of the original character count data and two modified data sets.

Notice by using the formula notation, we were able to calculate the summary statistics for each group.

> Exercise
> (a) Which is more affected by extreme observations, the mean or median? The data summary may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[18]

The median and IQR are called **robust estimates** because extreme observations have little effect on their values. The mean and standard deviation are much more affected by changes in extreme observations.

> Example The median and IQR do not change much under the three scenarios above. Why might this be the case?[19]

> Eexercise The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[20]

Transforming data

When data are very strongly skewed, we sometimes transform them so they are easier to model. Consider the histogram of salaries for Major League Baseball players' salaries from 2010, which is shown in Figure 10.

> Example The histogram of MLB player salaries is useful in that we can see the data are extremely skewed and centered (as gauged by the median) at about $1 million. What isn't useful about this plot?[21]

[18] (a) Mean is affected more. (b) Standard deviation is affected more.

[19] The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are relatively stable – there aren't large jumps between observations – the median and IQR estimates are also quite stable.

[20] Buyers of a *regular car* should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

[21] Most of the data are collected into one bin in the histogram and the data are so strongly skewed that many details in the data are obscured.

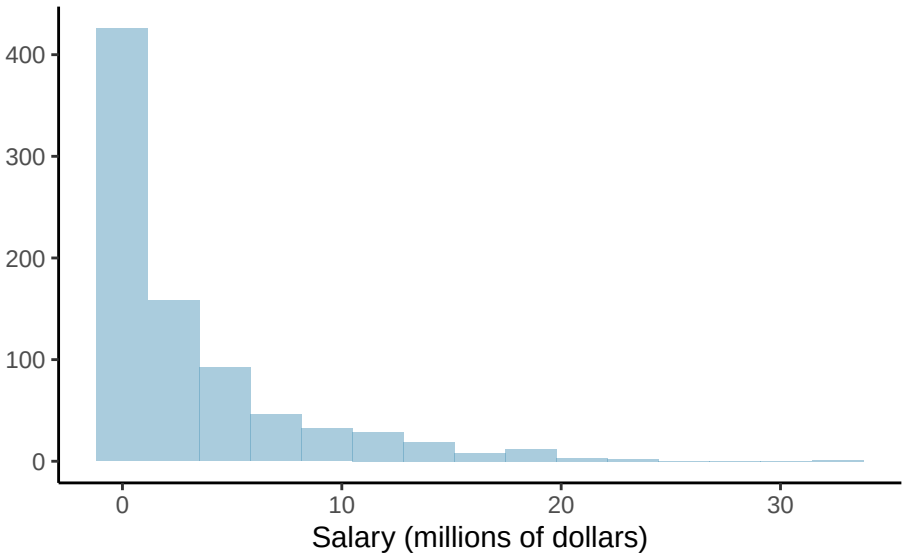


Figure 11: Histogram of MLB player salaries for 2010, in millions of dollars.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[22] of player salaries results in a new histogram in Figure 11. Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

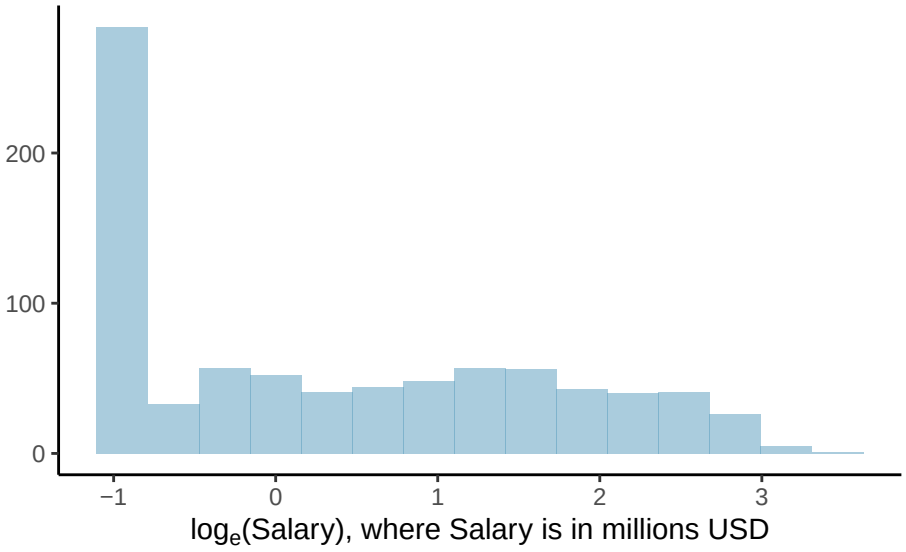


Figure 12: Histogram of the log-transformed MLB player salaries for 2010.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 2 above. We can see a positive association between the variables and that many observations are clustered near zero. Later in this course, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be

[22]Statisticians often write the natural logarithm as log. You might be more familiar with it being written as ln.

modeled very well. Figure 12 shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

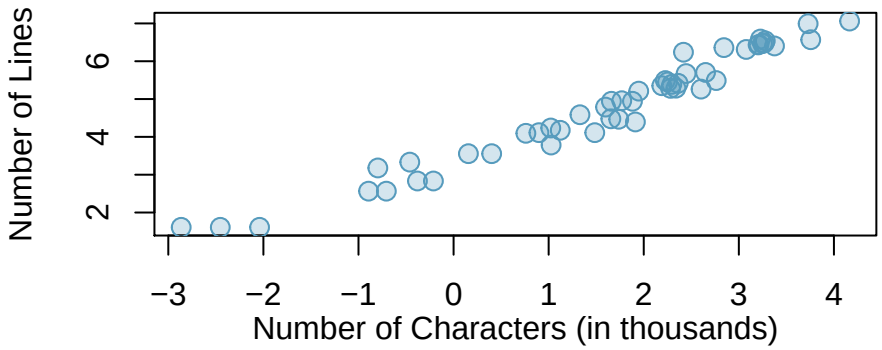


Figure 13: A scatterplot of *line_breaks* versus *num_char* for the `email50` data but where each variable has been log-transformed.

Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are used by statisticians. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

File creation information

- File creation date: 2020-05-11
- Windows version: Windows 10 x64 (build 17763)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.6.0
- `tidyverse` package version: 1.3.0
- `openintro` package version": 1.7.1