Chapter 11: More Plots (Histograms and Stem Plots)

Overview: In Chapter 10, we looked at 3 different kinds of plots: Pie Charts, Bar Graphs and Line Plots. The first two types are meant for categorical data and the third type is meant for time series data. Today, we look at 2 more ways to visualize data from numerical variable; histograms and stem and leaf plots.

Motivating Example: In a survey conducted by Statistics Canada from September to October 2020, individuals were asked to identify whether or not their mental health had improved, stayed the same, or worsened since the COVID-19 pandemic began. They were also asked to select which population groups applied to them. The data set titled MentalHealthData.csv available in Brightspace gives the percentages of people who replied that their mental health worsened. How can we visualize the distribution of the data?

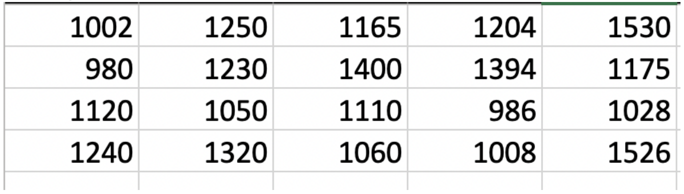
A **histogram** is an effective visualization tool when you have a numerical variable that takes on many (possibly infinite)) values

A histogram groups values together and then displays how many individuals in the sample belong to each group of values.

For example, the distribution of the following variables could be displayed using a histogram:

height of palm trees, weight of pandas, etc

Suppose you were trying to determine the distribution of SAT scores. If you have a sample of 20 people with the following scores:



1002	1250	1165	1204	1530
980	1230	1400	1394	1175
1120	1050	1110	986	1028
1240	1320	1060	1008	1526

How to Make a Histogram (by hand):

1. Determine the range of the data and then divide it into groups of equal width:

min 980
max 1530

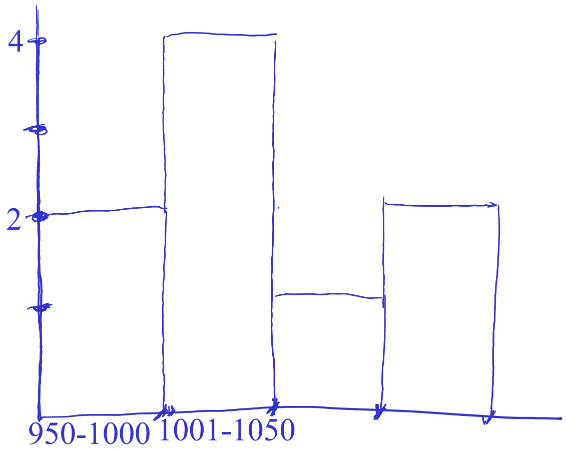
range = max-min = 1530-980 = 550

group by difference of 50
951- 1000
1001- 1050
1051- 1100
1101- 1150
etc.

2. Count the number of individuals in each group (this is called a frequency table):

Group	Frecuancy
951-1000	2
1001-1050	4
1051-1100	1
1101-1150	2
1151-1200	2
.	.
.	.

3. Draw the histogram. The x-axis contains the groups (in order) and the y-axis represents the frequency.



Note: This is a lot like a bar graph except that the x-axis has numerical groups rather than categories.

How to make a Histogram in R:

The hist() function takes in a vector of numerical values, automatically groups the values and computes the corresponding frequencies and then plots a histogram of the values.

Example: Download the MentalHealthData file that is available in Brightspace and read it into R. We are going to plot a histogram of the percentages of various Canadian population groups who feel that their mental health has declined throughout the pandemic.

```
MH.data = read.csv ("                              ", header = True)
```

1. Create a vector containing the data from your variable of interest:

```
percents = MH.data$Percentage
```

2. For the most basic histogram, simply use hist(vector):

```
hist (percents)   -> only for numerical vectors
```

3. You can add titles by using the *main* and *xlab* parameters:

```
hist (percents, main = "Percent of Population whose mental health has declined",
      xlab = "Percent of Population Group")
```

4. You can adjust the limits on the x and y axis by using *xlim* and *ylim* parameters:

```
hist (percents, xlim = c(min(percents), max(percents)+10), ylim = c(0,12))
```

5. You can adjust how many groups there are by using the *breaks* parameter:

```
hist (percents, breaks = 20)
```

6. You can also add colour to the histogram (both on the border and inside the bars):

```
hist (percents, border = "red", col= "blue")
```

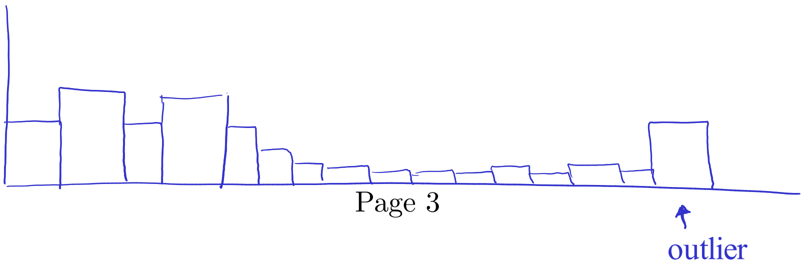
What can we tell from a Histogram?

We use a histogram as a way to visualize the distribution of a variable. When describing a distribution, there are certain characteristics that we look for which we now define below.

Definitions: When looking at the distribution of a variable, we look to see if any of the following apply:

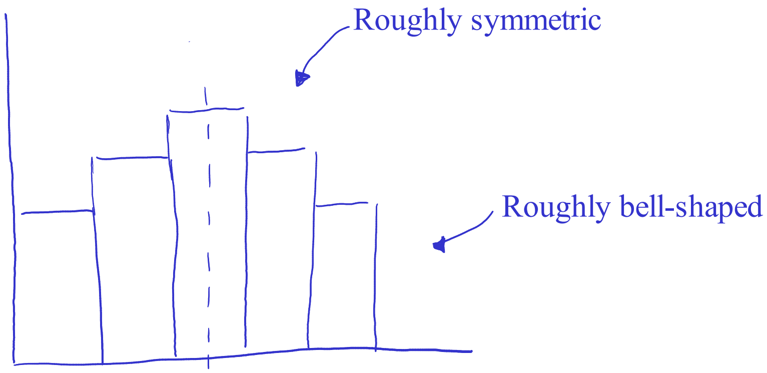
- An **outlier** in any graph of data is on individual observation that falls outline the overall pattern of the graph. This is a value in the data set that lies really for away from the rest of the data.

For example:



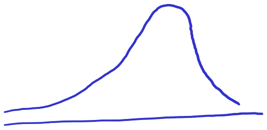
- A distribution is **symmetric** if the right and left sides of the histogram are approximatly mirror images of each other

For example:



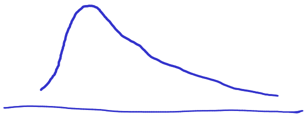
- A distribution is **skewed to the left** if

the left side of the histogram extends much further out than the right side (there is a long left tail)



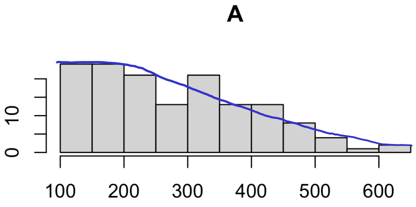
- A distribution is **skewed to the right** if

the right side of the histogram extends much further out than the left side

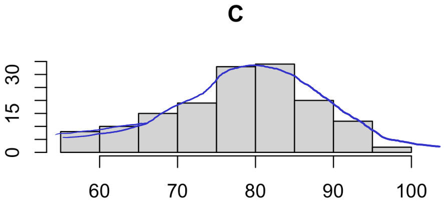


Practice Question:

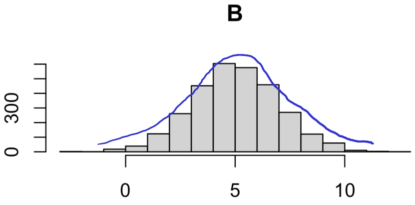
Consider the following histograms and determine which one is symmetric, right-skewed, and left-skewed.



right-skewed



more left-skewed than symmetric



symmetric

One last characteristic of a distribution that we can visualize using a Histogram is:

- The **variability** of a distribution describes the spread out values that the variable takes on. Looking at the histogram, we can see the range that the values take on (smallest group and largest group) which tells us someting about variablity

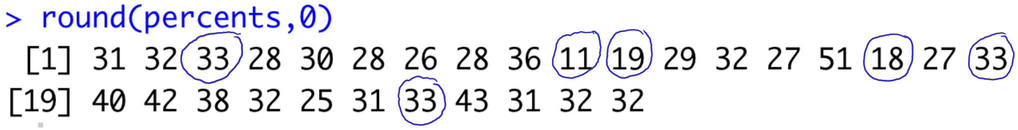
Note: We will learn a precise way to compute variability in Chapter 12.

Stemplots: One more type of plot that we will see in this chapter is called a Stemplot.

While histograms tend to be the most commonly used tool to visualize a distribution of a numerical variable, for small data sets, we can also create something called a stemplot.

How to make a Stemplot by hand:

We will illustrate the procedure with an example. Consider again the (rounded) percentages of various population groups who feel that their mental health has been negatively impacted since the pandemic:



```
> round(percents,0)
 [1] 31 32 33 28 30 28 26 28 36 11 19 29 32 27 51 18 27 33
[19] 40 42 38 32 25 31 33 43 31 32 32
```

1. Separate each observation into a **stem** consisting of all but the final (right-most) digit and a **leaf** (the final digit).

   Note: Stems may have as many digits as needed, but each leaf contains only a single digit.

e.g. $\underset{\text{stem}}{3}\underset{\text{leaf}}{1}$ $\qquad$ $\underbrace{21}_{\text{stem}}\underset{\text{leaf}}{0}$

2. Write the stems in a vertical column with the smallest at the top and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order from the stem.

	stem	leafs
[10-19]	1	1, 8, 9
[20-29]	2	5, 6, 7, 7, 8, 8, 8, 9
[30-39]	3	0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 6, 8
[40-49]	4	0, 2, 3
[50, 59]	5	1

Notice: If you turn your stemplot on it's side, the appearance is similar to that of a histogram.

In R, there is stem() function

```
stem(round(percents, 0))
```