Chapter 2: Descriptive Statistics

(Ott & Longnecker Sections: 3.3-3.5)

https://dzwang91.github.io/stat324/

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

What do we study?

This chapter is concerned with descriptive statistics. We will be particularly interested in descriptive statistics for numerical data.

Key Concepts: Histograms, measures of location and spread, median, quartiles, quantiles, mean, range, interquartile range, standard deviation, variance

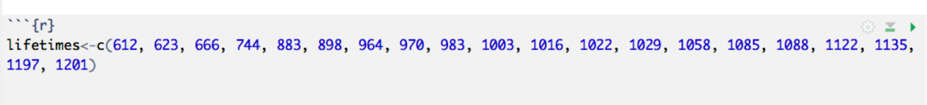
Outline

1. Graphical Summaries: Histogram
    - Frequency histogram
    - Relative frequency histogram
    - Other examples of histograms
    - How many bins to use

2. Numerical Summary Measures
    - Median
    - Quartiles
    - Mean
    - Range
    - Interquartile range
    - Standard deviation

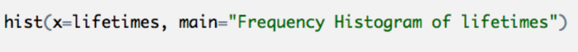
Graphical Summaries: Histogram

The following data represent the lifetimes (in hours) of 20 different lamps. The data was gathered as part of a quality control sample of lamps created at a large electronics manufacturer. They are ordered from smallest to largest for convenience:
612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135, 1197, 1201
The list is ordered – which gives us a a little insight – but a useful visual representation would be superior.

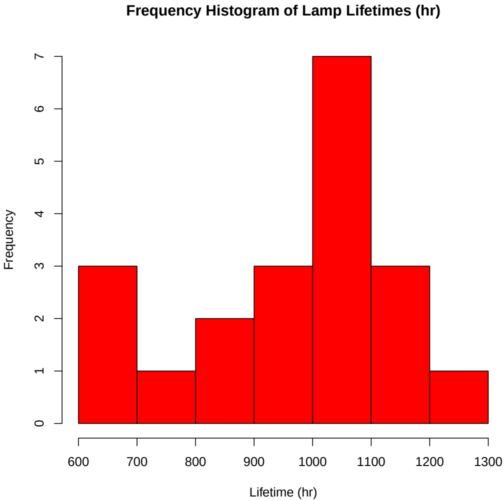


~~~
```{r}
lifetimes<-c(612, 623, 666, 744, 883, 898, 964, 970, 983, 1003, 1016, 1022, 1029, 1058, 1085, 1088, 1122, 1135,
1197, 1201)
~~~

Frequency histogram



```
hist(x=lifetimes, main="Frequency Histogram of lifetimes")
```

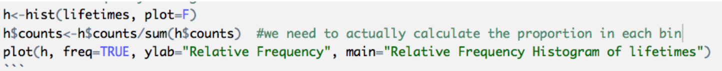


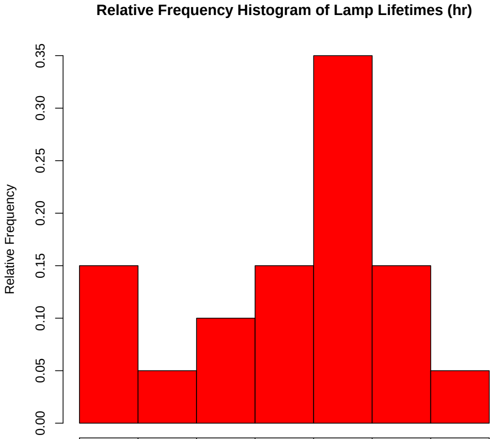
Frequency histogram

- The y-axis is frequency (# observations), and the x-axis is lifetime.
- Each bar sits atop a “bin” of data.
- Each observation goes into one, and only one, bin (for observations at boundaries, employ a consistent convention)
- **Example.** The first bin runs from 600 to 700. Since the bar atop it extends up to 3, this means there are three observations in the data set that fall between 600 and 700, and indeed this is true, the values being 612, 623, and 666. The second bin runs from 700 to 800 and contains only one observation, 744. And so on.

Relative frequency histogram

Another useful formulation is to express the y-axis as the relative number of observations in that bin:



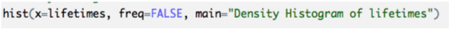
```
h<-hist(lifetimes, plot=F)
h$counts<-h$counts/sum(h$counts)  #we need to actually calculate the proportion in each bin
plot(h, freq=TRUE, ylab="Relative Frequency", main="Relative Frequency Histogram of lifetimes")
```
```


```

# Relative frequency histogram

- y-axis is now the proportion of observations in the bin
- **Example.** In this formulation, the first bar is at a height of 0.15, meaning that 15% of the data falls between 600 and 700. This is because there are 20 total data points, and $3/20 = 0.15$.

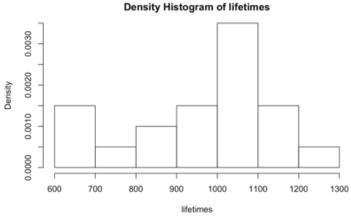
# Density frequency histogram

$density = \frac{relative\ frequency}{width\ of\ bins}$ (so that total area=1). Density*Width=Relative Frequency

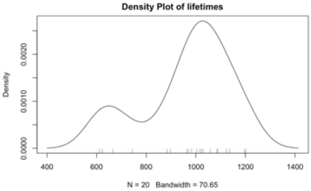


```
hist(x=lifetimes, freq=FALSE, main="Density Histogram of lifetimes")
```

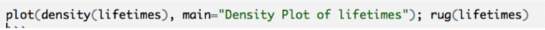
*or a density plot “smooths” out the bars



Density Histogram of lifetimes

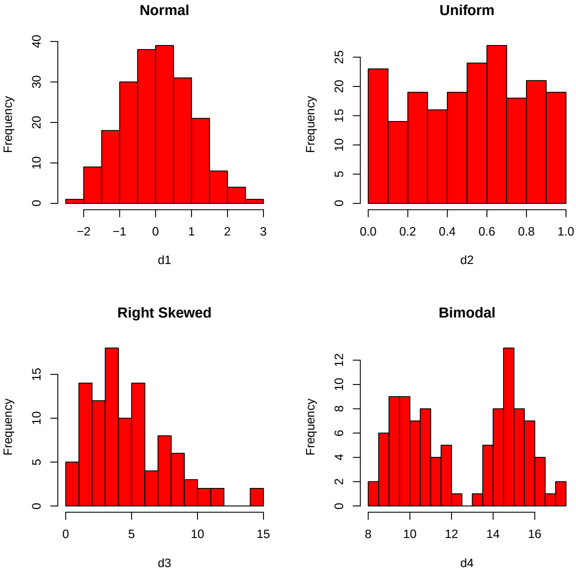


Density Plot of lifetimes



```
plot(density(lifetimes), main="Density Plot of lifetimes"); rug(lifetimes)
```

Other examples of histograms



Other examples of histograms

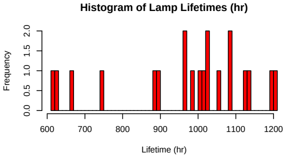
- In the upper left box, we can see that data follows a roughly bell-shaped curve, with most of the observations in the middle and fewer and fewer as we move away from the middle. We call it "normal".
- In the upper right box, we see that apart from small fluctuations, the frequency is about the same for all of the values on the x-axis. We call it "uniform".
- In the lower left box, "Right Skewed", we see that small values are common, but occasionally there are quite large values.
- In the lower right box, "Bimodal", we see a distribution that looks like two normal curves side-by-side. This might indicate data which mixes two kinds of things.

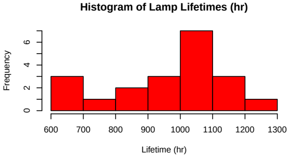
How many bins to use

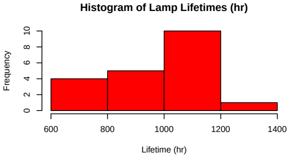
There is one important consideration, and that is the number of bins to use. If too many bins are used, almost every bin will have only a small number of observations, which is almost like not summarizing at all. And if too few bins are used, all the data will be piled together and two very different data sets might look exactly the same.

How many bins to use

Consider this series of histograms with different numbers of bins for the same set of data:







Outline

1. Graphical Summaries: Histogram
   - Frequency histogram
   - Relative frequency histogram
   - Other examples of histograms
   - How many bins to use

2. Numerical Summary Measures
   - Median
   - Quartiles
   - Mean
   - Range
   - Interquartile range
   - Standard deviation

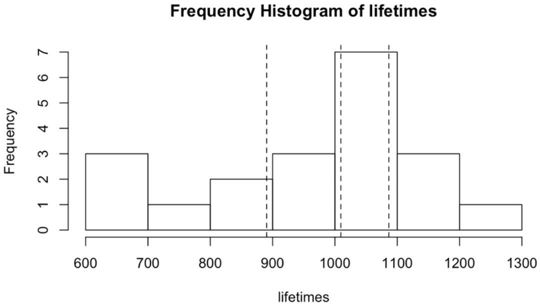
Median

- The median of a dataset is the value such that half of the data are smaller than the value and half are larger.
- Precise definition:
  - If a dataset consists of an odd number of observations, the median is the middle value in the sorted list.
  - If the dataset consists of an even number of observations, the median falls halfway between the two middle values in the sorted list (the average of the two middle values).
- **Example.** There are twenty observations in our lamp lifetime data. The data has been sorted, so the two middle values are just the 10th and 11th observations in the list: 1003 and 1016. Halfway between them is 1009.5 hrs, and that is the median.

Quartiles

- The quartiles are three values that divide the data into 4 approximately equal groups.
- The **first quartile** is defined as the value such that 75% of the observations are larger than that value, and 25% are smaller.
- The **third quartile** is defined as the value such that 25% of the observations are larger than that value, and 75% are smaller.
- (The median can also be called the second quartile.)

Quartiles


Frequency Histogram of lifetimes
Frequency
0
1
2
3
4
5
6
7
600
700
800
900
1000
1100
1200
1300
lifetimes

Quartiles

Precise definitions:

- If the dataset consists of an even number of observations, the first quartile is the median of the first half of the sorted list, and the third quartile is the median of the second half of the sorted list.
- If the dataset consists of an odd number of observations, use the following procedures.
  - **First quartile:** Make a new dataset including the median and every observation smaller than the median. The first quartile is the median of this new dataset.
  - **Third quartile:** Same as above, only use the median and all larger observations.

Quartiles

Example: The lamp data has 20 observations, an even number. Therefore the first quartile for this data is the median of the first 10 observations, which is 890.5 hrs, and the third quartile is the median of the second 10 observations, which is 1086.5 hrs.

Quartiles

Example: The lamp data has 20 observations, an even number. Therefore the first quartile for this data is the median of the first 10 observations, which is 890.5 hrs, and the third quartile is the median of the second 10 observations, which is 1086.5 hrs.

Note: our definition of quartile is not the only one! Different books and different software packages sometimes compute them differently.

Mean

- The **mean** is defined as the sum of the observations divided by the number of observations.
- Sometimes we call it the **average** or **expectation**.
- Let $y_1, y_2, ..., y_n$ denote the n observations in a dataset.

$$\text{Sum of the observations: } \sum_{i=1}^{n} y_i = y_1 + y_2 + ... + y_n$$

$$\text{Mean: } \bar{y} = \frac{\displaystyle\sum_{i=1}^{n} y_i}{n}$$

- **Example.** For the lamp data, the mean is

$$(612 + 623 + ... + 1201)/20 = 965 \text{ hrs}$$

Range

- Perhaps the simplest measure of spread is the **range**, which is the largest value minus the smallest value.
- **Example.** For the lamp data, the range is $1201 - 612 = 589$ hrs.
- The approximate value of the spread can be seen on a histogram (but not the exact value).
- **Example.** From the lamp histogram, we can see that the range must be somewhere between $1300 - 600 = 700$ hrs, and $1200 - 700 = 500$ hrs, and indeed it is.

Interquartile range

- Another relatively simple measure of spread is called the **inter-quartile range** or **IQR**.
- The IQR is computed as the third quartile minus the first quartile.
- The IQR is roughly the range of the middle 50% of the data.
- **Example.** For the lamp data, the IQR is $1086.5 - 890.5 = 196$ hrs.
- One benefit of the IQR is that it is less sensitive to changes in very large or very small values. Indeed, the smallest and largest quarters of the data are ignored in the calculation of IQR!
- **Example.** In the lamp data, if the largest data point had been 2000, the range would go up to $2000 - 612 = 1388$ hrs, but the IQR would be exactly the same.

Standard deviation

- The final measure of spread we will discuss is the **standard deviation**, or **SD**.
- The SD is a measure that tells us how far away a typical observation in the dataset is from the mean. You can think of standard deviation as "average distance from the mean".
- The formula and some notation for standard deviation is:

$$\text{Standard Deviation: } s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference?

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.
- Why divide by $n - 1$?

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.
- Why divide by $n - 1$? We'll give an answer when we talk about estimation.

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.
- Why divide by $n - 1$? We'll give an answer when we talk about estimation.
- Why take the square root?

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.
- Why divide by $n - 1$? We'll give an answer when we talk about estimation.
- Why take the square root? The reason has to do with the units in which the data are measured. Since we are squaring the differences, the units of the mean squared differences would be the original data units squared. We take the square root so that the standard deviation is in the original units of the data.

Standard deviation

There may be many puzzling aspects to this formula.

- Why take the square difference? For certain mathematical reasons, it remains so.
- Why divide by $n-1$? We'll give an answer when we talk about estimation.
- Why take the square root? The reason has to do with the units in which the data are measured. Since we are squaring the differences, the units of the mean squared differences would be the original data units squared. We take the square root so that the standard deviation is in the original units of the data.
- The un-square-rooted version of the standard deviation is called the **variance** and is usually denoted s_y^2.

Standard deviation

- **Example.** For the lamp data, the standard deviation works out to:

$$\sqrt{\frac{((612-965)^2+(623-965)^2+...+(1201-965)^2}{20-1}} = 178.30 \text{ hrs.}$$

Notice the units are in hours, not hours squared.

What's the next?

In the next chapter, we'll talk about concepts of probability.