Chapter 1 Statistical Vocabulary Notes

Name: Chenghao Wu
Date: 1/24/2022
Chapter Number: #1
Title of Chapter: Statistical Vocabulary

1. Descriptive Statistics

- Descriptive Statistics: It summarizes a **vector** of data by calculating a scalar value summarizes those data.
- Vector: a list of scalar values that all quantify one attribute.(as height, weight. age, etc.)
- Scalar: a single number

1.1 Measure of central tendency

- MOCT summarizes a vector of data by figuring out the
  - the location of the typical
  - the middle
  - or the most common point of data

Three most common measures of central tendency are the mean, the median, and the mode.

- **The mean (arithmetic mean, aka average)**
  - The mean is computed by adding together all of the values in a vector, by summing them up to compute a total, and then dividing the total by the number of observations in the vector.
  - Example in R:

    rainfall <- c(0,0,0,2,1,1,4) # Amount of rain each day this week
    sum(rainfall) / length(rainfall) # Compute the mean the hard way
    mean(rainfall) # Use a function to compute the mean

```
> rainfall <- c(0,0,0,2,1,1,4)   # Amount of rain each day this week
> sum(rainfall) / length(rainfall)  # Compute the mean the hard way
[1] 1.142857
> mean(rainfall)   # Use a function to compute the mean
[1] 1.142857
>
```

The first line creates a vector of seven values using the c() command, the second two lines has the same function.

- **The median**
  - If we sorted all of the elements in the vector from smallest to largest and then picked the one that is right in the middle, we would have the median.
  - Outliers: a case that someone who looks at the data considers to be sufficiently extreme (either very high or very low) that it is unusual.
  - Benefit of the median: **it is less impacted by, and more resistant to outliers.**
  - The mean is somewhat enlarged by the presence of that one case, whereas t**he median represents a more typical middle ground.**
  - Example in R:

    median(rainfall)

```
> median(rainfall)
[1] 1
>
```

- **The mode(modal value, the most typical evalue, statistical mode.)**
  - The **mode** is the value in the data that occurs most frequently.
  - **It is also resistant to outliers.**
  - Example in R:
    - **Instead of using mode() function, the best practice is using mfv() in modeest package.**

```
install.packages("modeest")   # Download the mode estimation package
library(modeest)              # Make the package ready to use
mfv(rainfall)                 # mfv stands for most frequent value
```

```
> mode(rainfall)
[1] "numeric"
>
> install.packages("modeest")
Error: unexpected input in "install.packages(?"
> library(modeest)
> mfv(rainfall)
[1] 0
> >
```

Conclusion of measure of central tendency:

- The mean provides a precise, mathematical view of the arithmetic center point of the data, but it will impacted by the outliers.
- The median provides a way to find out which observation lies right at the balancing point.

- By comparing the mode and the mean, we can learn which direction the outlier are pulling the mean to.
- Before manipulating the data, it is best to take a glance of the central tendency.

1.2 Measure of dispersion

- The difference between of c(299,300,301), and c(200, 300, 400), although they all have same mean and median, but they have the different dispersion of data. The first one is higher and the second one is lower.
- So the how dispersed, or “spread out,” the data are, is the same important as the central tedency.
- The most commonly used measure of dispersion is “standard deviation”

- **The Range**
  - You can calculate the range simply by subtracting the smallest value in the data from the largest value.
  - Cons: Because one isolated value on either the high end or the low end can completely throw things off.
  - Example in R(range() function will give you the highest and the lowest value in the vector)

```
> disdf<-c(200,300,400)
> disdf2<-c(299,300,301)
> disdf3<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,99)
> range(disdf)
[1] 200 400
> range(disdf2)
[1] 299 301
> range(disdf3)
[1]  1 99
> 
```

- **Deviations from the mean**
  - The DFM of a scalar from vector = the scalar - the mean.
  - The sum of the DFMs in a vector always equal to zero.
  - It is basically sooooo unhelpful to use the sum of abs.

- **Sum of squares**
  - The quantity is the sum of the squared deviations from the mean, also commonly called the "**sum of squares.**"
  - {DFM1)^2+ (DFM2)^2+.....+(DFMn)^2
  - it is the totality of how far points are from the mean, with lots more oomph given to the points that are further away.

- **Variance (mean square)**
  - It is t**he sum of squared deviations from the mean divided by the number of observations.**
  - As the number of scalars in the vector increased, the sum of squares will keep growing , so we want a more accurate metric.
  - Sum of squares/ number of scalars
  - You can think of the variance as an average or "mean" squared deviation from the mean.
  - Example in R (also be careful of the parentheses, it tells you the order of the calculation)

```
> (votes-mean(votes))^2  # Show a list of squared deviations
[1] 10000     0 10000
> sum( (votes - mean(votes)) ^ 2)   # Add them together
[1] 20000
> sum( (votes - mean(votes)) ^ 2) / length(votes)  # Divide by the number of observations
[1] 6666.667
> ›
```

- **Standard Deviation**
  - Its' **the square root of the variance**, less un-wieldly and more intuitive and plain than variance.
  - Example in R: (As we can see, the scenario one deviations are a hundred times as big as the deviations in scenario two.)
    - The code below using the **population of the data sets**, which stands for the entire data set.
    - The R also provides us with var() and sd() for calculating the variance and the standard deviation. But they **use the sample of data,** which will be a slight diff from the result from the population.
    - According to the **Central limit theorem， normally when N >50, the result of sample and population will be more likely the same. If their a bigger dataset, we can use var() and sd().**

Using population

```
> votes1 <- c(200,300,400) # Here is scenario one again
>
> sqrt( sum((votes1 - mean(votes1))^2) / length(votes1) ) # Here is the standard deviation
[1] 81.64966
>
> votes2 <- c(299,300,301) # Here is scenario two
>
> sqrt( sum((votes2 - mean(votes2))^2) / length(votes2) ) # And the same for scenario 2
[1] 0.8164966
> ›
```

Using sample

```
> var(votes1)
[1] 10000
> var(votes2)
[1] 1
> sd(votes1)
[1] 100
> sd(votes2)
[1] 1
> 
```

- **Mathematic formula for mean and sd**

Mean: $$\mu = \frac{\sum x}{N}$$

Standard deviation: $$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

2. DISTRIBUTIONS AND THEIR SHAPES

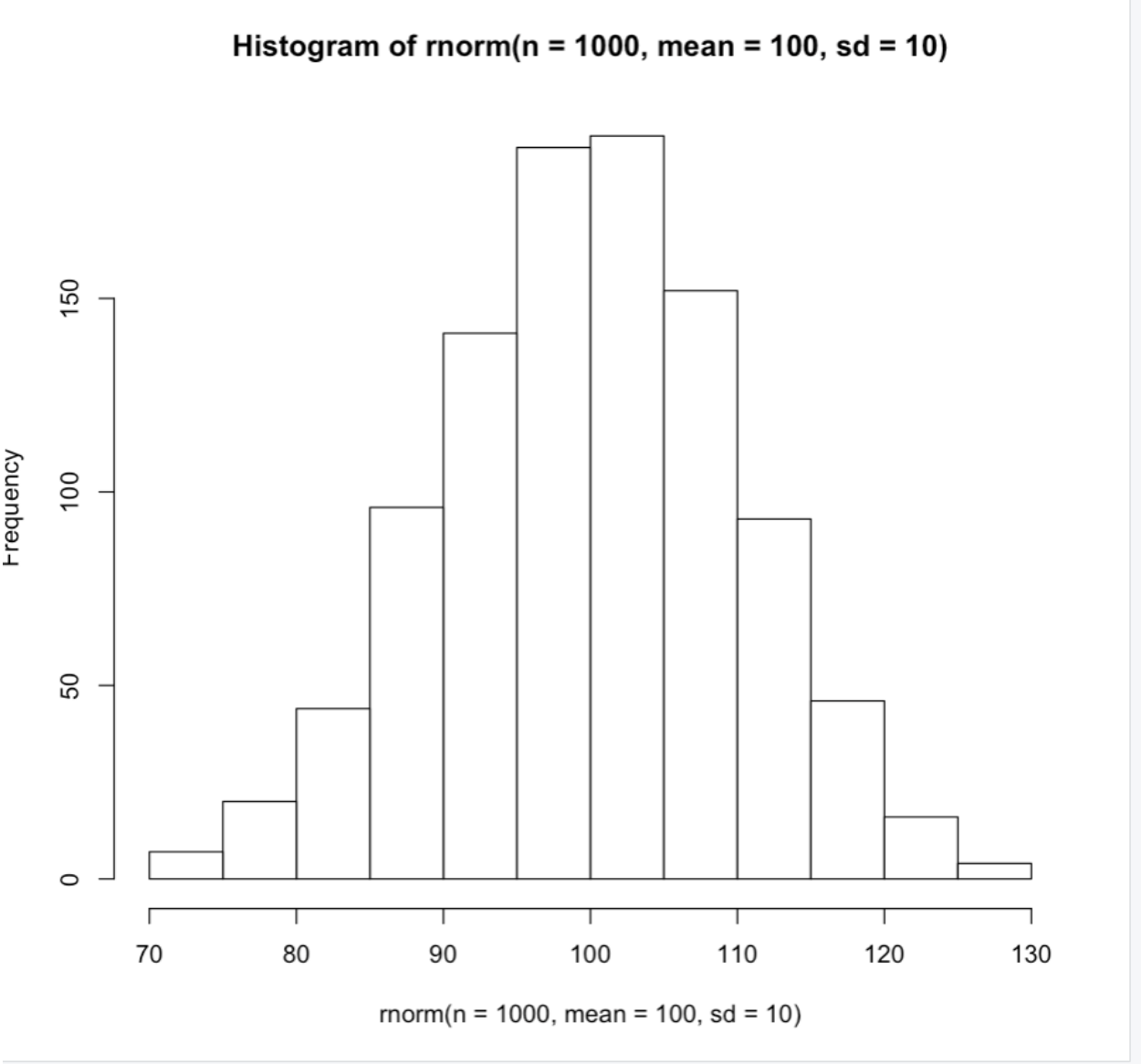
If we can use the measure of the central tendency and of dispersion depends on the distribution of dataset. It is the amount of observations at different spots along a range or continuum, the more intuitive way is observe the shape.

2.1 The normal distribution (the bell curve)

- The common explanation for why a distribution of any measurement takes on a bell shape is that the underlying phenomenon has many small influences that add up to a combined result that has many cases that fall near the mean.
- The variance(extra bigger/ shorter) will cancel each out and many will near the mean.

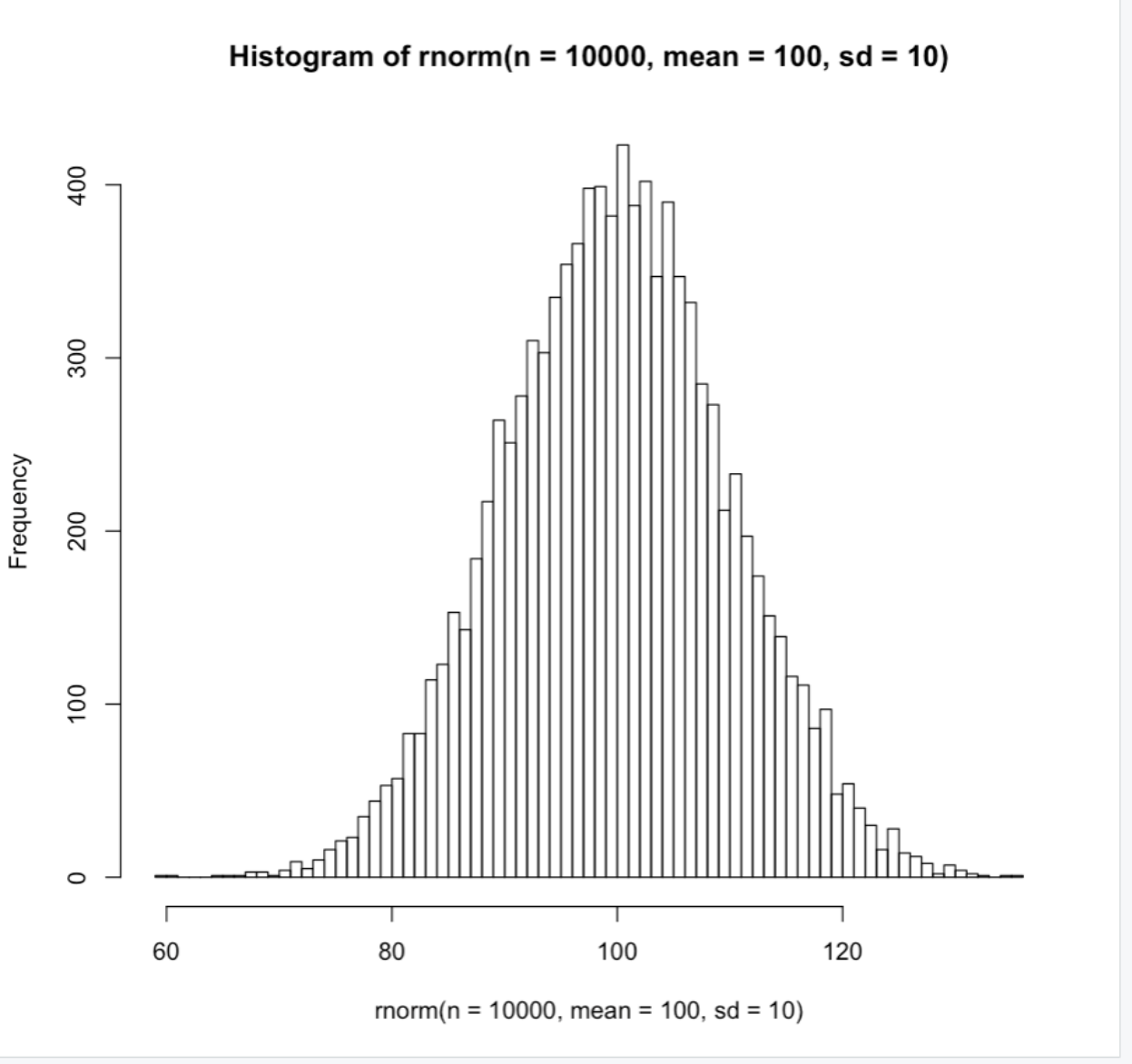
- Example in R:
  - rnorm() : It will random generate a list of 1000 numbers in normal distribution, you can specify the mean and sd.
  - hist(): It takes that list of points and creates a histogram from them.
  - This histogram is considered a **univariate** display, because in its typical form it shows the shape of the distribution for just a single variable.
  - x-axis: Each of the groupings must have the same range so that each bar in the histogram covers the same width. from small values to large values.
  - y-axis: where the height of each bar indicates that there are a certain number of observations, as shown on the *Y*-axis (the vertical axis).
  - The two or three bars above the low end of the *X*-axis are the **left-hand tail,** while the top two or three bars on the high side are the **right-hand tail.**

hist(rnorm(n=1000, mean=100, sd=10))



You can also define the number of groups using the parameter" breaks= "

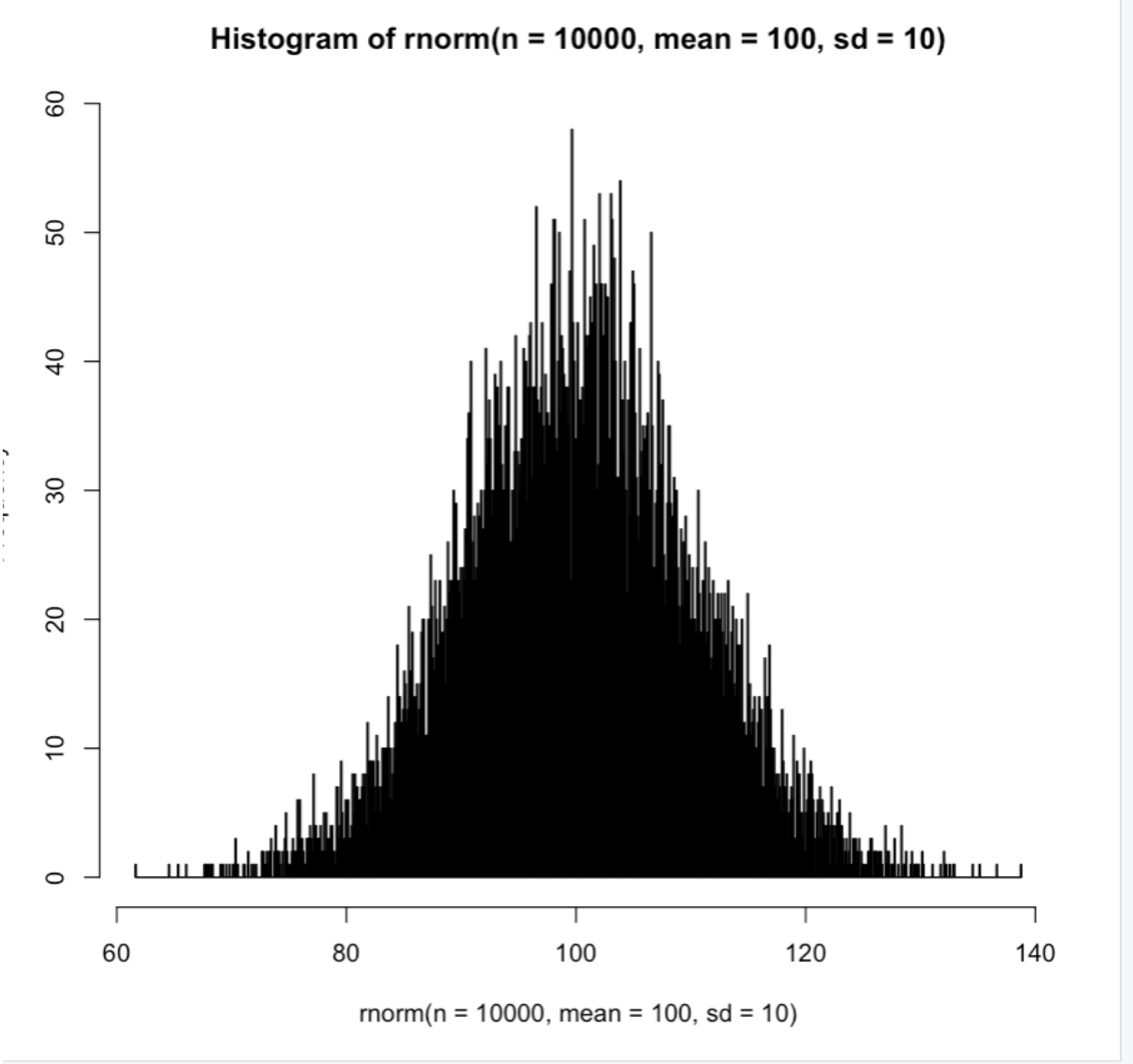
hist(rnorm(n=10000, mean=100, sd=10), breaks=100)



The more the breaks, it will be more smooth and more like a bell curve.

With an infinite number of categories and an infinite amount of normally distributed data, the curve would be per- fectly smooth. That would be the “ideal” normal curve and it is that ideal that is used as a model for phenomena that we believe are normally distributed.

hist(rnorm(n=10000, mean=100, sd=10), breaks=1000)

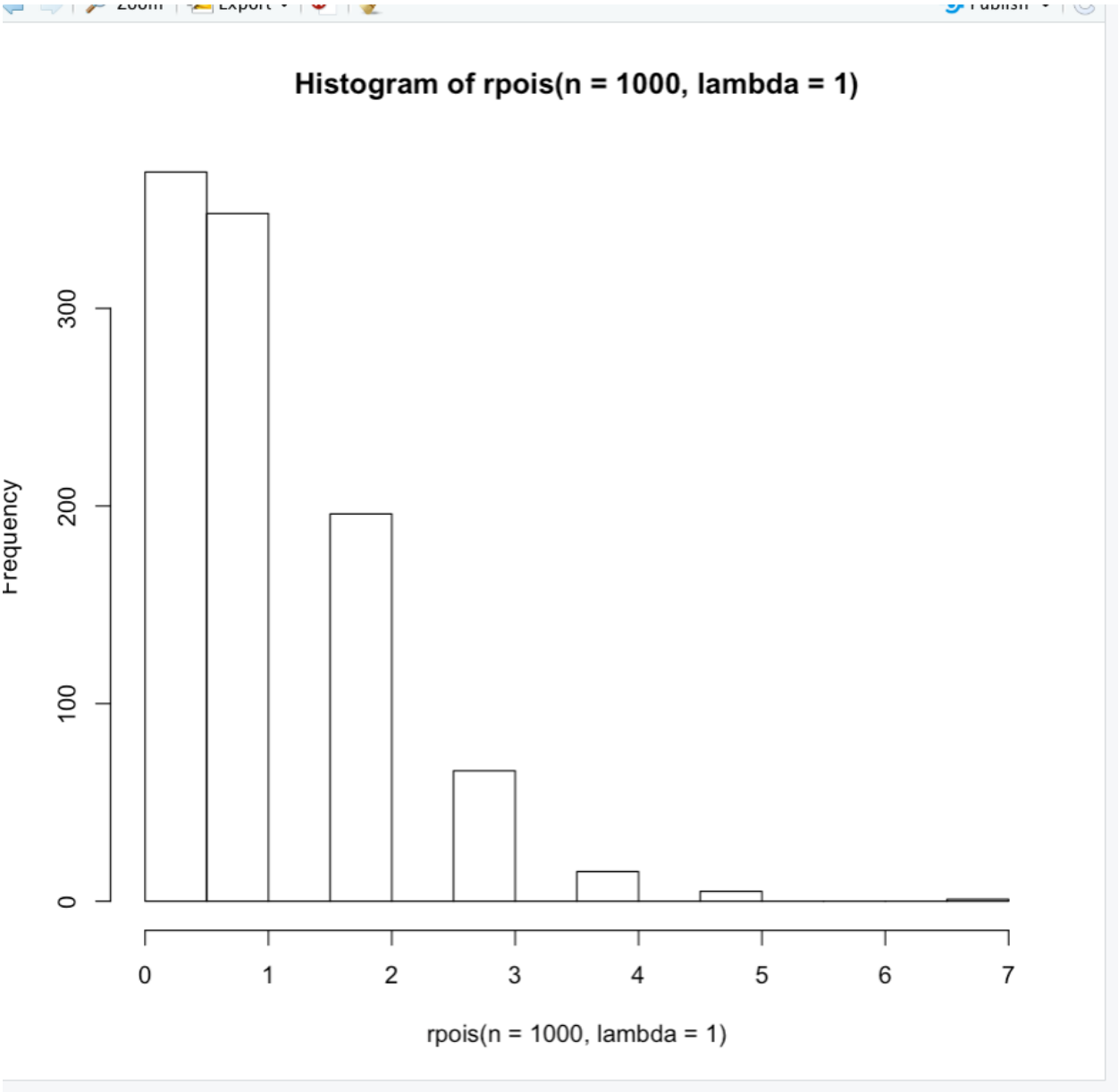


Statistical significance referred to thinking about chance occurrences and likelihoods in terms of a model. The normal distribution is one of these models—in fact prob- ably the single most popular model for reasoning about statistical significance.

2.2 The Poisson Distribution(more about arrival time similarity)

- Poisson distribution fits a range of natural phenomena, and in particular works well for modeling arrival times.
- The distribution of these “arrival” times turns out to have many similarities to other kinds of arrival time data, such as the arrival of buses or subway cars at a station, the arrival of customers at a cash register, or the occurrence of telephone calls at a particular exchange.
- The reason that these phenomena share this similarity is that in each case a mixture of random influences impacts when an event occurs.
- Example in R:
    - rpois() :generates a list of random numbers that closely match a Poisson distribution, “n=” tells the function how many data points to generate, “lambda=” refers to the **mean that we expect. BUT IT WILL NEVER EQUALS TO 1 UNLESS WE HAVE INFINITE DATA POINTS.**

```
hist( rpois(n=1000, lambda=1) )
```



one way of thinking about values that make a Poisson distribution is thinking of them as delays between arrival times—in other words the interval of time between two neighboring arrivals.

eg: Most cars will have 1-2 mins delay passing a specific location, and if we have a turtle, it will fall in 7 mins category.

We can also check the mean of this poisson distribution(expected 1 but never will be 1, it will change every time)

```
> mean(rpois(n=1000, lambda=1))
[1] 1.025
> >
```

One way to set seed is assigning it to a variable.

```
> myPoiSample <- rpois(n=1000, lambda=1)
> >
```

then each time you run the code below, it will give you the same mean.

```
> mean(myPoiSample)
[1] 0.978
> >
```

Also we can check the sd of a poisson distribution:(it's sampled from 1000 using sd())

```
> sd(myPoiSample)
[1] 0.9982547
> >
```

Chapter Wrap Up

variable= an individual list(vector)

All terms above summarizes the variable

univariate descriptive statistics: using statistics to describe single variables. Single variables are the main building blocks of larger statistical data sets,

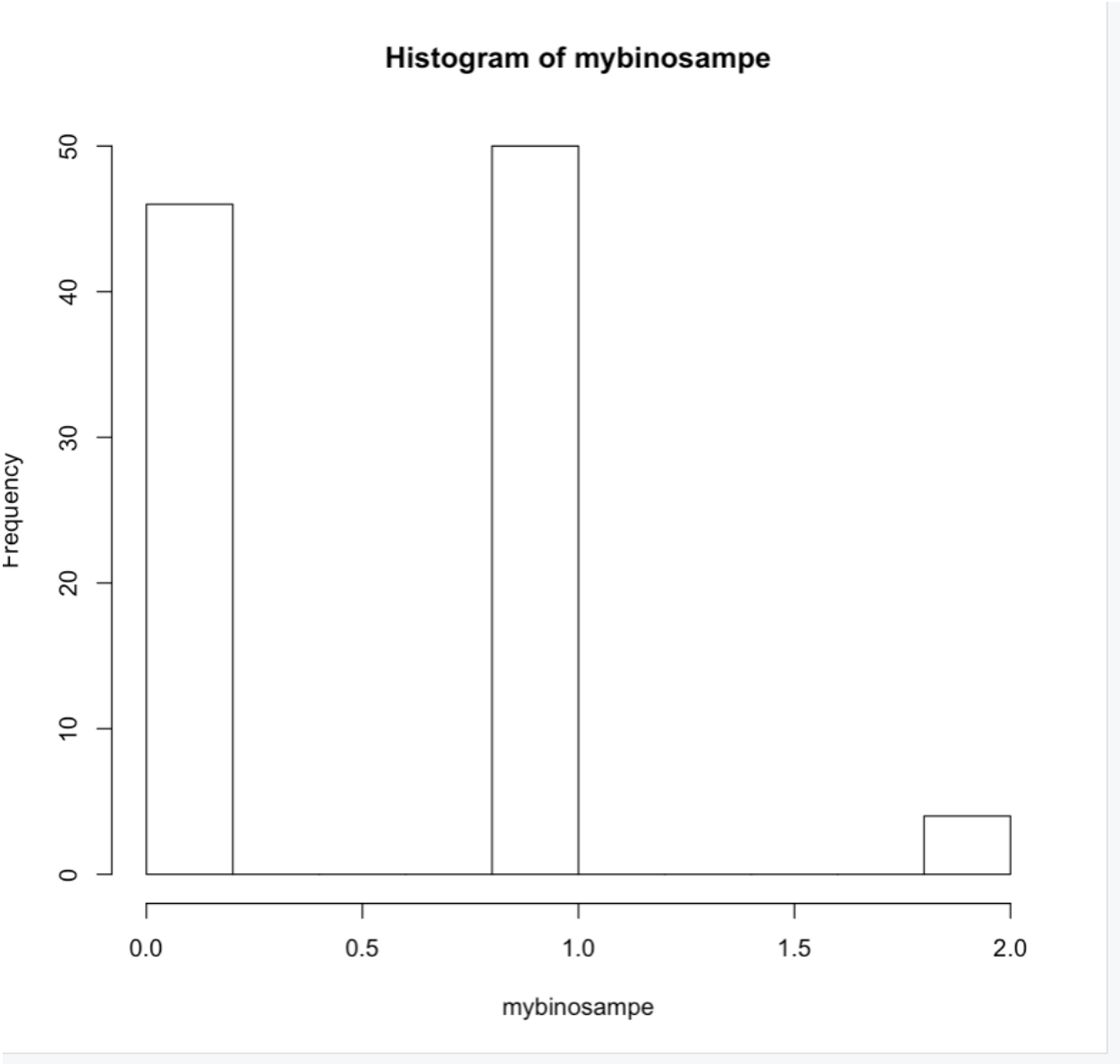
EXERCISES

1. **Using the material from this chapter and possibly other information that you look up,**
   write a brief definition of these terms in your own words: mean, median, mode, vari-
   ance, standard deviation, histogram, normal distribution, and Poisson distribu-
   tion.

(please see the first bullet points for the each term above)

1. **Write the equations, using the appropriate Greek letters, for the population mean and**
   population standard deviation. Explain briefly what each Greek letter means. The R
   environment offers about 20 different kinds of statistical distributions. Choose any one
   of these distributions other than the normal distribution or the Poisson distribution. (The
   help system in R can assist you with finding a description of these distributions and
   their commands: type “?distributions” at the command line. For a hint about one distri-
   bution you might choose to study, read the beginning of the next chapter!) Write some
   R code that generates 100 random points in that distribution, displays a histogram of
   those 100 points, calculates the mean of those points, and calculates the standard
   deviation. Make sure to use the technique shown just above that begins with assigning
   the 100 points to a vector that can be reused for all of the other commands.

```
> ?distribution
> mybinosampe<- rbinom(100, 2, 0.3)
> mean(mybinosampe)
[1] 0.58
> sd(mybinosampe)
[1] 0.5717243
> hist(mybinosampe)
> 
```



1. **Use the data() function to get a list of the data sets that are included with the basic installation of R: just type “data()” at the command line and press enter. Choose a data set from the list that contains at least one numeric variable—for example, the Bio-chemical Oxygen Demand (BOD) data set. Use the summary() command to summarize the variables in the data set you selected—for example, summary(BOD). Write a brief**

description of the mean and median of each numeric variable in the data set. Make sure you define what a “mean” and a “median” are, that is, the technical definition and practical meaning of each of these quantities.

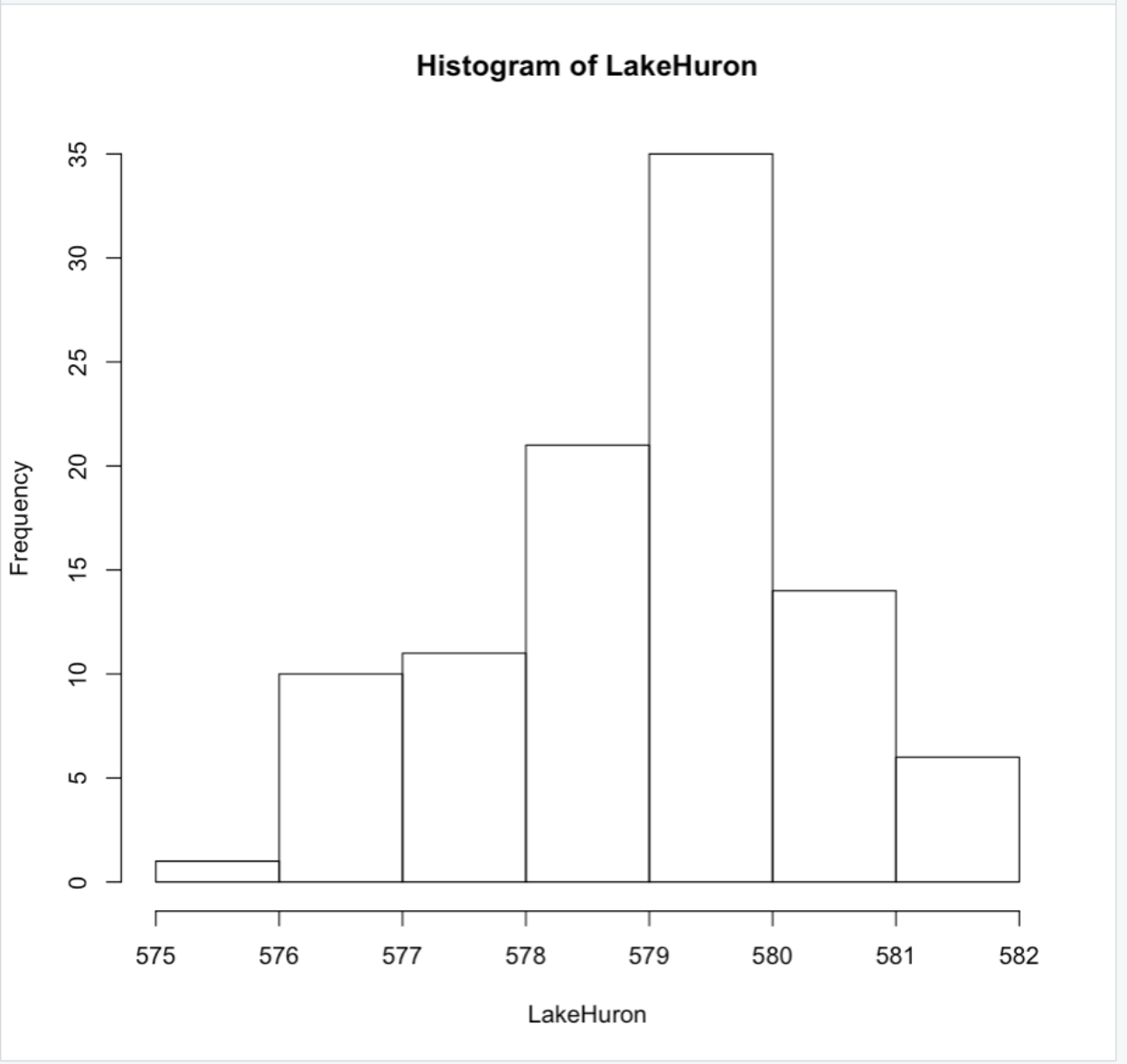
```
> summary(ToothGrowth)
      len        supp         dose
 Min.   : 4.20   OJ:30   Min.   :0.500
 1st Qu.:13.07   VC:30   1st Qu.:0.500
 Median :19.25           Median :1.000
 Mean   :18.81           Mean   :1.167
 3rd Qu.:25.27           3rd Qu.:2.000
 Max.   :33.90           Max.   :2.000
>
```

The average tooth growth length in our dataset is 18.81 and the number lies on the central location is 19.25, which means we may have a lower outliers drag the mean to the left direction.

The average dose of support vitamin in our dataset is 1 and the number lies on the central location is1.167 , which means we mean have a more samples using dose bigger than 1 than less than 1.

1. **As in the previous exercise, use the data() function to get a list of the data sets that are included with the basic installation of R. Choose a data set that includes just one variable, for example, the LakeHuron data set (levels of Lake Huron in the years 1875 through 1972). Use the hist() command to create a histogram of the variable—for example, hist(LakeHuron). Describe the shape of the histogram in words. Which of the distribution types do you think these data fit most closely (e.g., normal, Poisson).**

Speculate
on why your selected data may fit that distribution.



It’s more fits normal distribution.