pieandbar-day1

kim :

August 28, 2023

NOTES for class: A.TONS of data out there! Need to be able to make sense of the data and use it honestly. Two components to this class:
1) statistics and probability 2) R -to understand and use statistics it is essential to look at data. To look at dataset of any substantial size absolutely need a good computer software tool.Too difficult to enter into a calculator and the calculator does not have all the functionality. R is one the two top statistics software packages in the industry.Data is complex and so a program to use it has some complexities. With lots of thoughtful practice you will get good at R.

First, we will practice some examples of code by copying and altering code from the spreadsheet rtasks1300 into the console, which is in the lower lefthand window of the Rstudio big window. You enter commands at the > and press enter.

This document is an example of an rmarkdown. It is a way to create reworkable code in R and its output and documentation in one file. You should save this after changes by typing command-s or going to 'save' under 'file'.It may seem like a bit extra at first but it is well worth that for the reworkability. You could think of it as a notebook to save your r work. And you can do additional practice in it.

To run this you will need the package "rmarkdown". Install this by going to packages in the lower right hand corner and entering rmarkdown in the box.

The following is called a code snippet.Note that it is highlighted in gray. What is outside a code snippet is just text, not code. The comments in the code snippet, which don't get run, are marked off with a #. To run (and test) this code snippet you click the green arrow to the right). You can then change the code and retry it.

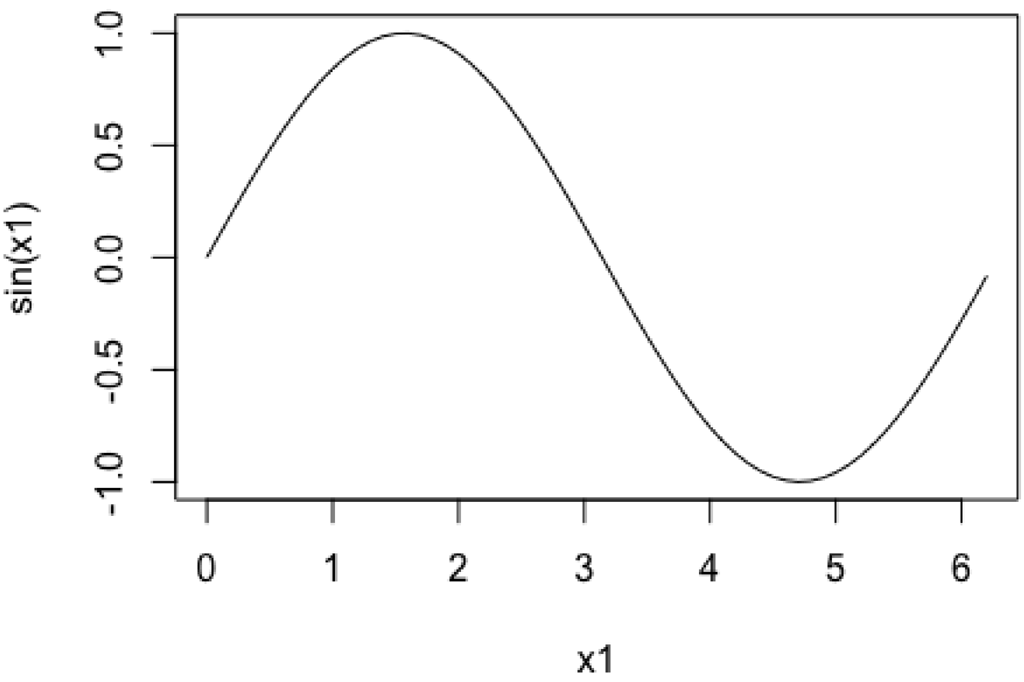
```r
#enter the numbers 2,3,7,9 in the variable x
x=c(2,3,7,9)
#sum the elements in x
sum(x)

## [1] 21

#determine what the fourth element of x is
#change the third element of x to 12
#see what x is now
#multiply x by 4 and see what the result is.

myvec=1:9
myvec
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

```
mean(myvec)
```

```
## [1] 5
```

```
x1<-seq(0,2*pi,0.1) #<- can be used in place of =
#seq(start,finish, increment)
plot(x1,sin(x1),'l') #plot x, function of x, connect the dots.
```



Create another code snippet by typing option-command- I.Then create a variable named trees with the values 15, 22, 45.Run the snippet.

To the above snippet add another variable treenames with entries oak, chestnut, hickory. You need to put each of the treenames in quotes.

```
#remove the hatch marks and run the following code
#pie(trees,treenames)
#pie(trees,treenames,radius = 4)
```

B. Populations and individuals and variables:(Some examples from my environment: titanic_train, census1880 (See notes later on how to get these files)

Types of variables: categorical and numerical 1.categorical variable: usually not a number. could be intervals (example) Examples from above.

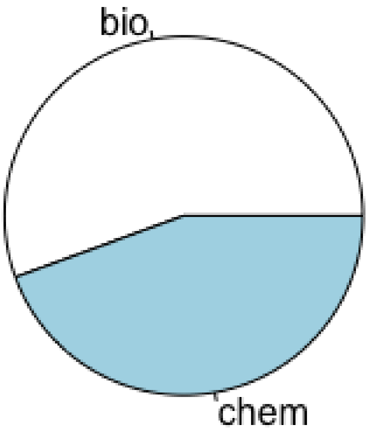
Give some examples of why we are interested. Basic types of questions for categorical variables: How many? Proportion of each value for the category? (show example)

Can you think of some population that would have at least three categorical variables attached to it? Try to make it something from your major or another interest.

Pie chart. What proportion of the whole does each value represent?

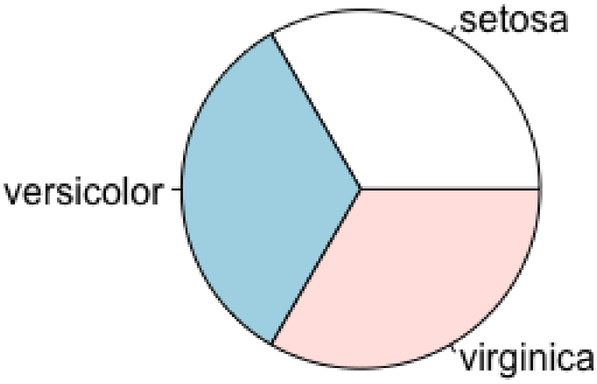
example 1:In a certain class, 5 students are taking biology, 4 are taking Chemistry, 12 are taking Physics, and 6 are taking no science course. Represent this via a pie chart. For the pie chart to be valid what needs to be true?

```r
#Typical code for creating the variables and the pie chart
sci.type<-c('bio','chem') # complete this list of labels
num.stu<-c(5,4)   # complete this list or vector of number of students in each science course.
pie(num.stu,sci.type)    #format of pie: pie(list/vector of numbers in each category, list/vector of labels)
```



example 2: built-in data set iris

```
pie(table(iris$Species))  #iris is the dataset; Species is the variable.We
need to provide pie with number of each value, so use table for that.
```

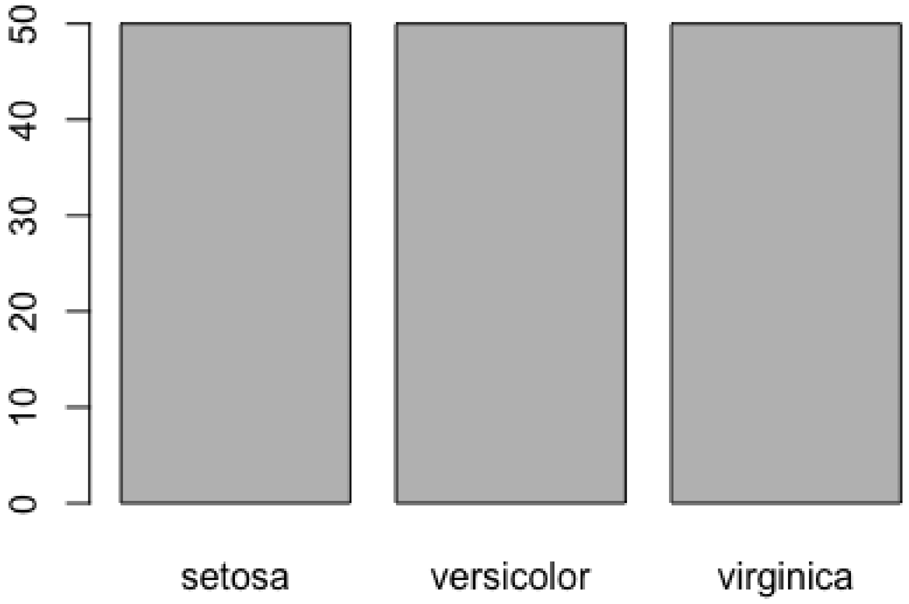


```
table(iris$Species)

##
##     setosa versicolor  virginica
##         50         50         50

#create your own pie chart
```

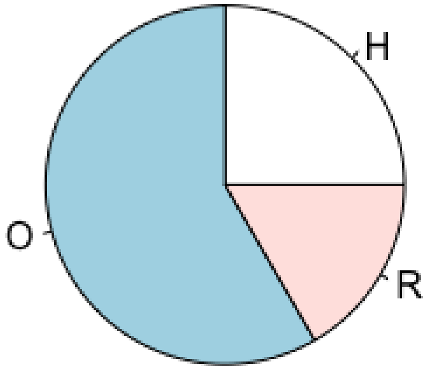
What does that tell you?

bar graphs

```
barplot(table(iris$Species)) # what does this graph tell you?
```



```
#you have the following. How will you get a pie chart form it? What is wrong
with the code pie(mytrees)?
mytrees=c("O","O","H" ,"O","O","H","R","R","O","O","H" ,"O")
#pie(mytrees) -does not work
pie(table(mytrees))
```



```
#barplot(trees, names.arg=treenames)
#barplot(trees)
```

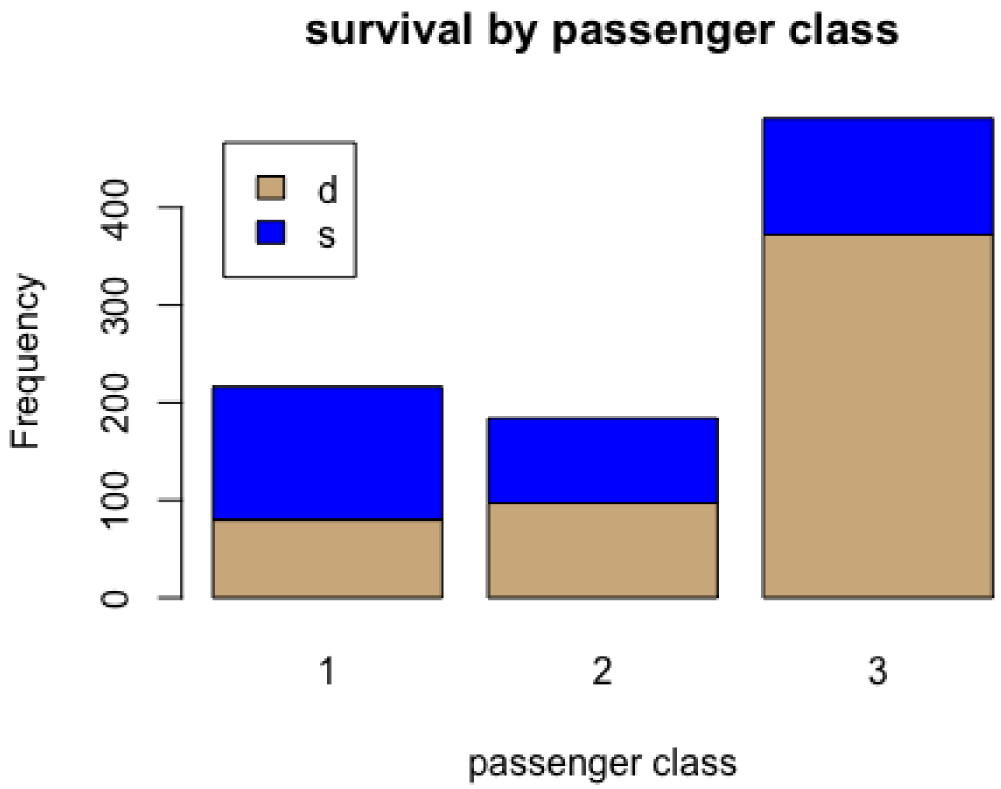
Download the files titanic_tr.rda and census1880midlancpa.rda from canvas or the source your teacher gives you. Then click on each of them to load them into R.

```
#the code lines allow me to knit my document. You can ignore them for now.
You would replace them with your own load statement if you were knitting
this.
load("~/Desktop/Rnotes/census1880midlancpa.rda")
load("~/Desktop/Rnotes/titanic_tr.rda")

#View(titanic_train)# The dataset contained in the file titanic_tr.rda is
called titanic_train. Deactivate to see the data

#View(thefam2) #The dataset contained in the file census1880midlancpa.rda is
thefam2.
```

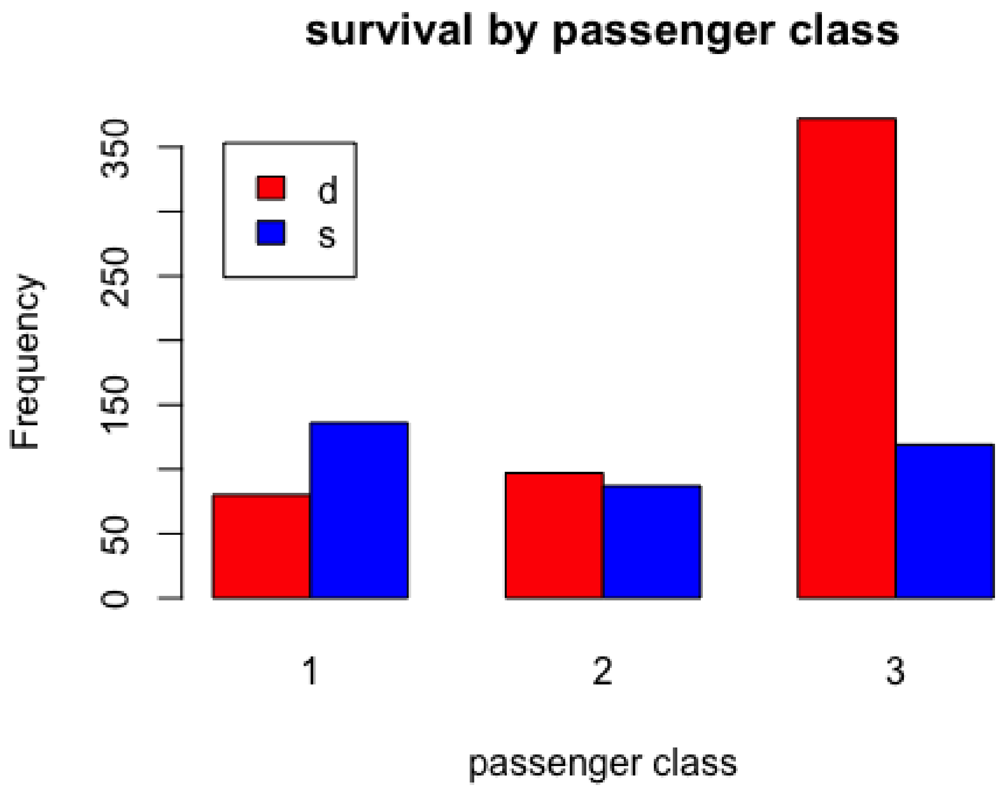
b. Create a pie chart of the variable titanic_train$Pclass
c. Create a bar graph of the same variable.
d. You can do the same for thefam2 and the variable relationship. Which is the biggest group within relationship? Why do you think that is?

The following is to give you examples of stacked and grouped bar plots.I would not expect you to reproduce this without the example.

```
v=table(titanic_train$Survived,titanic_train$Pclass)
v

##
##       1   2   3
##   0  80  97 372
##   1 136  87 119

#note two directions, six possible outcomes
 barplot(v,col=c('tan','blue'),main='survival by passenger class',
xlab='passenger class',ylab='Frequency')
 legend('topleft',inset=.05,c('d','s'),fill=c('tan','blue'))
```



```
 #what do you notice from this graph?
 #color list -type colors at prompt

barplot(v,col=c('red','blue'),main='survival by passenger class',
xlab='passenger class',ylab='Frequency',beside=T)
legend('topleft',inset=.05,c('d','s'),fill=c('red','blue'))
```



e.Modify the above code to plot the number of male /female by class.

Added-> f. Try modifying the above code, inserting t(v) (the transpose of v) instead of v. What do you get? What all do you need to change?

Summary:R: basic vectors-declaring, accessing, altering,View, table, pie, barplot

Notes about rmarkdowns: First, an rmarkdown can be just used as a place to keep code and keep trying it again.

Second, an rmarkdown can be knitted to either an html, a pdf, or a doc document to create cool reports(see the little knit icon at the top). When you do that, it needs to know where exactly a datafile is. To do that, load your datafile using files in the lower right hand window. Then copy the load statement - which you'll find in the console - into a code chunk in your rmarkdown. Make sure your load statement precede any code that uses that dataset.