**Descriptive Statistics**

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Variables are things that we measure, control or manipulate in our research which assumes different values for different elements. They differ in many respects, most notably in the role they are given in our research and the type of the measures that can be done to them.

**There are two types of variables**

Quantitative: A variable that takes numerical values for which the arithmetic operations make sense is called a quantitative variable. These variables can be divided in two parts

a. Discrete Variable: A variable whose values are countable is called discrete. e.g. number of homes, number of cars, number of accidents etc.

b. Continuous Variable: A variable that can assume any numerical value over a certain interval or intervals is called a continuous variable. e.g. income, age, height, time etc.

Qualitative: A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a qualitative (categorical) variable .e.g. gender, hair color, pain level etc.

**Qualitative Data**

We often organize categorical data with tables but we may also look at the data graphically with bar graphs or pie charts.

**Using tables**

The ***table*** command allows us to look at tables. Its simplest usage looks like ***table(x)*** where x is a categorical variable.

A survey asks people if they smoke or not. The data is

Yes, No, No, Yes, Yes

We can enter this into R with the c() command, and summarize with the table command as follows

> x=c("Yes","No","No","Yes","Yes")

> table(x)

The table command simply adds up the frequency of each unique value of the data.

**Factors**

Categorical data is often used to classify data into various levels or factors. For example, the smoking data could be part of a broader survey on student health issues. R has a special class for working with factors which is occasionally important to know as R will automatically adapt itself when it knows it has a factor. To make a factor is easy with the command factor or ***as.factor.***

>x=c("Yes","No","No","Yes","Yes")

>factor(x)

**Bar charts**

A bar chart draws a bar with height proportional to the count in the table. The height could be given by the frequency, or the proportion. The graph will look the same, but the scales may be different. Suppose the grade distribution of the students in MA 345 is as below:

|  |  |
| --- | --- |
| Grade | Frequency |
| A | 12 |
| B | 20 |
| C | 10 |
| D | 3 |
| F | 5 |

In order to draw a bar graph we use the following R code:

>grade<-c(12,20,10,3,5)

> barplot(grade)

It produces a bar graph as below



In order to place the title, color and labels we use the R code below

> grade<-c(12,20,10,3,5)

> names(grade)=c("A", "B", "C", "D", "F")

>barplot(grade,col=c("1","2","3","4","5"),main="Grade Distribution")



Note that don't use barplot with the raw data we need to create a frequency table.

We can draw a relative frequency bar graph using the following code

> grade<-c(12,20,10,3,5)

> names(grade)=c("A", "B", "C", "D", "F")

> barplot(grade/(sum(grade)),col=c("1","2","3","4","5"),main="Grade Distribution")

**Pie charts**

A circle divided into sectors that represent the percentages of a population or a sample that belongs to different categories is called a pie chart. To construct a pie chart, we multiply by 360 the relative frequency for each category to obtain the degree measure or size of the angle for the corresponding category.

Suppose the distribution of the students enrolled in a small college is as below

|  |  |
| --- | --- |
| Major | Frequency |
| Maths | 120 |
| Science | 200 |
| Engineering | 350 |
| Education | 230 |
| Technology | 125 |

>students<-c(120, 200, 350, 230,125)

>names(students)=c("Maths","Science","Engineering","Education","Technology")

>pie(students,col=c("1","2","3","4","5"),main="Student Enrollment”)



**Graphical Display of Quantitative Data**

**Histogram**

A histogram is a graphic that gives an idea of the “shape" of a sample, indicating regions where sample observations are concentrated and regions where they are sparse. A histogram is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. When the number of observations n is large (several hundred or more), some have suggested that reasonable starting points for the number of classes can be given by Sturge's formula is c = 1 + 3.3 log n (note it is the common logarithm with base 10) where n is the number of observations in the data.

Construct a histogram for minimum temperature recorded in Chicago for last 20 days

25, 37, 20, 31, 31, 21, 12, 25, 36, 27, 38, 16, 40, 32, 33, 24, 39, 26, 27, 19

Below is the R code to draw a histogram

> x<-c (25,37,20,31,31,21,12,25,36,27,38,16,40,32,33,24,39,26,27,19)

> hist(x)



> x<-c (25,37,20,31,31,21,12,25,36,27,38,16,40,32,33,24,39,26,27,19)

>hist(x,main="Histogram of Chicago Temp.",xlab="Temperature",ylab="# of Days", c=3)



>hist(x, breaks=15) it suggests 15 breaks

>hist(x, breaks = c(10, 20, 30, 40))# uses these breaks

>hist(x, breaks="scott") Use "Scott" algorithm

Also note that the 'Sturges" algorithm is the default one.

**Density estimation for continuous variables**

The problems associated with drawing histograms and the density function of continuous variables is much more challenging. The subject of density estimation is an important issue for statisticians. You can get a feel for what is involved by browsing the?***density*** help window. The algorithm used in ***density.default*** disperses the mass of the empirical distribution function over a regular grid of at least 512 points, uses the fast Fourier transform to convolve this approximation with a discretized version of the kernel, and then uses linear approximation to evaluate the density at the specified points. The choice of bandwidth is a compromise between smoothing enough to rub out insignificant bumps and smoothing too much so that real peaks are eliminated.

The rule of thumb for bandwidth is



Where *n* is the number of data points.

Old faithful geyser eruption data are available in the data set ***faithful*** which can be accessed as

***> attach(faithful)***

***>(max(eruptions)-min(eruptions))/(2\*(1+log(length(eruptions),base=2)))***

[1] 0.192573

>par(mfrow=c(1,2))

>hist(eruptions,15, freq=FALSE,main="Density Curve I",col=2)

>lines(density(eruptions,width=0.6,n=200))

>truehist(eruptions,nbins=15, main="Density Curve II",col=3)

>lines(density(eruptions,n=200))



Remark: Note that ***truehist*** is in the package MASS

**Steam and leaf plot:**

Stem-and-leaf plot is a simple way of summarizing quantitative data. In a stem-and-leaf plot each data value is split into a "stem" and a "leaf". The "leaf” is usually the last digit of the number and the other digits to the left of the "leaf" form the "stem". Usually there is no need to sort the leaves, although computer packages typically do.

> x<-c(78,74,82,66,94,71,64,88,55,80,91,74,82,75,96,78,84,79,71,83)

> stem(x)

The decimal point is 1 digit(s) to the right of the |

5 | 5

6 | 46

7 | 11445889

8 | 022348

9 | 146