Stats 110A, Spring 02

Graphical Summaries

Variable: an attribute of a collection that varies from individual to individual, or even from measurement to measurement.

Statistics is concerned with how to quantify and describe variation, and we therefore focus on attributes that show variation. We are concerned with two basic characteristics of such variables:
a) the variables' values and
b) the values' frequencies
In other words: what values does the variable take on and with what frequency?

The handout you have is similar to one given to people participating in a UCLA study. We're going to examine the responses of that study and compare them to our own.

To do this, we'll make graphical summaries of the data. Most graphical summaries present these to aspects of the variable.

Data: Respondents were asked to assess the "risk" associated with various activities on a scale of 0 (no risk) to 100 (maximum risk). The activities we'll consider today are:
1) using household appliances
2) living near a nuclear power station
3) pool: swim in an indoor public pool each weekend
4) plane: fly on commercial airplanes every month
5) xray: receive diagnostic xrays every six months

Getting Data into R

Your best bet is to start with your data in an ascii (text) file, with each item separated by tab, and each observation on one line. Here's a copy of the first few lines of the data set:

ID	appliances	nuclear	pool	plane	xray	gender
625	0	100	50	50	0	male
526	0	100	10	0	100	female
684	0	95	10	5	80	female

Let's suppose this data is saved in a file called "risk.txt". R looks for files in its working directory. You can set the working directory with one of the menu items: Tools: Change Working Directory. Change this to the directory in which risk.txt is stored.

The command to load the data into R is

```
risk <- read.table("risk.txt", header=T)
```

We now have what's called a "data table" of this data. It's rows are the observations and it's columns are the variables. You can see the names of the variables by typing:

```
> names(risk)
[1] "ID"         "appliances" "nuclear"    "pool"       "plane"
[6] "xray"       "gender"
```

To refer to variables by name, we must type:

```
attach(risk)
```

To see the first 10 entries of a variable, say appliances:

```
> appliances[1:10]
 [1] 0 0 0 0 0 0 0 0 0 0
```

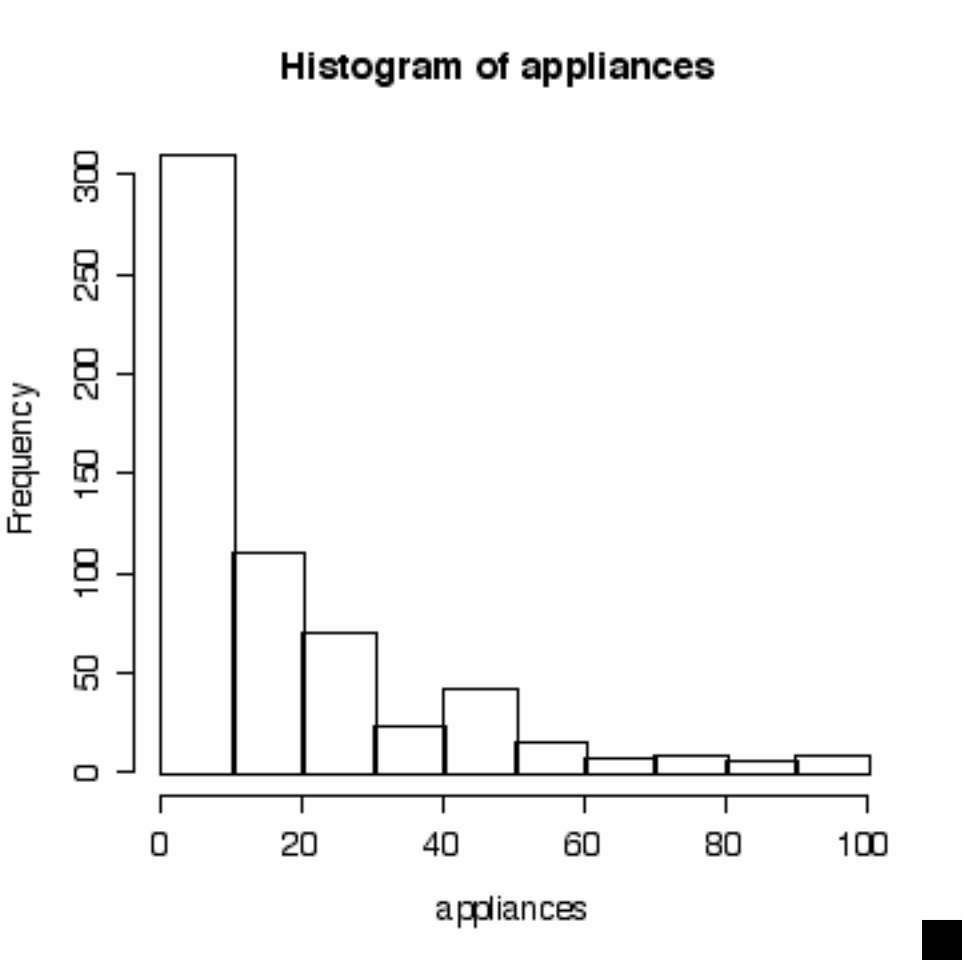
We could see the entire list by typing just "appliances". But as you can see, that would be along list:

```
> length(appliances)
[1] 611
```

Histogram

There are several ways of "viewing" these data. The most common is the histogram:

```
hist(appliances)
```



The histogram creates 'bins' and counts the number of observations falling in each of these bins.

Variations

You can change the number of bins:

```
> hist(appliances, breaks=50)
> hist(appliances, breaks=2)
```

What is the affect of this on the shape of the histogram? What is most informative?

You can change the y-axis from frequency (counts) to percentages:

```
 hist(appliances, freq=F)
```

Stem and Leaf Plots

Really these are meant to be done by hand, but R does it too. The most basic version does this: the last digit of each observed value is the "leaf". Preceding digits are the "stem". Write the stems on the left, the leafs on the right. And you get a histogram-like picture.

```
> stem(appliances)
```

Categorical Variables

Some variables have values that are categories, not numbers. Two useful plots for these are bar charts and pie charts. Their construction is somewhat tedious in R, for reasons that are not clear, and so we won't go over them here. Consult your book for details.

Comparisons

Do males and females assess the risk of household appliances differently?

One of the faults of R is that this is not as easy to do as in some packages. But the basic idea is that we want to put our numerical displays as close together as possible. Here are some different steps:

a) look at males, females alone, sequentially

hist(appliances[gender == "male"])

hist(appliances[gender=="females"])

b) look at them side-by-side

```
> par(mfrow=c(1,2))
> hist(appliances[gender==
+ "male"])
> hist(appliances[gender=="female"])
```

c) Look at them one on-top the other:

```
> par(mfrow=c(2,1))
> "male"])
Error: syntax error
> hist(appliances[gender=="male"])
> hist(appliances[gender=="female"])
```

d) restore originall settings
par(mfrow=c(1,1))

Questions for Discussion

1) The distribution of "appliances" is "skewed right" . This means that it has a "tail" that points towards the right. This is common when the possible values are bounded below (by 0 here) and the typical score is a low one.

a) Are there any variables you would expect to be right-skewed? Symmetric? Try it and see.

2) Uses graphically techniques, which, if any, would you say is the most risky, and why?

3) What would you say is a "typical" risk level for

a) appliances
b) nuclear
c) pool

4) Which activities, if any, do you think show the most disagreemet in the risk ratings? Why do you say this/

5) Without looking, which activities do you think will differ most by gender. Check your hunches.