STAT 100 Notes March 24, 2010

Today:

Big Ideas of Statistics

The Nature of Statistical Software (R)

Introduction to Week 11

Big Ideas of Statistics:

The content of the following comments are from a paper by Stephen Stigler, the most famous current researcher of the History of statistics. It is available at

http://news.harvard.edu/gazette/story/2008/10/key-statistical-ideas-celebrate-birthdays/

I am including a summary of it here so we can see if our STAT 100 course has introduced these big ideas.

1. Averaging
2. Root n Rule
3. Hypothesis Testing
4. Percentiles and the Normal Distribution
5. Regression
6. Random Sampling
7. Statistical Design
8. Graphical Display of Data
9. Chi-squared Distribution
10. Modern Computation and Simulation

1. Averaging: The idea is that there is more information in an average than in the numbers that are averaged. Another way to say this is that averages are less variable than the numbers averaged. Our discussion of "theory of averages" certainly involved this idea. It also relates to points 2 and 4.

2. Root n Rule: In a way, the $\sqrt{n}$ rule quantifies how much more information there is in an average than in the numbers averaged. We discussion how a sample average had a proportion $1/\sqrt{n}$ of the variability of the numbers averaged. Recall that we measured the variability of numbers using the SD.
3. Hypothesis testing: The idea that a small p-value suggests that the assumption used to compute it sheds doubt on the assumption – this was the logical breakthrough that allows researchers to make good decisions about which effects are likely real and which are transient effects due to random variability.
4. Percentiles and the Normal Distribution: The Normal distribution in our course was introduced to describe the distribution of sample averages, and also as a model of a distribution that is symmetrical about its mean, and whose likely values is indicated by its SD. In talking about 1, 2, and 3 SDs from the mean having certain "tail" probabilities of 16%, 2.3%, and .13%, we were actually talking about "percentiles". The 16^{th} percentile of the standard normal distribution is "-1". The 97.7^{th} percentile is +2.
5. Regression: This is the method we used to predict a Y from an X. We also discussed articles in which a Y was predicted from X1, X2, X3, … The reason it is called "regression" has an historical root. Francis Galton noticed that sones of tall fathers were shorter than their fathers, and sons of short fathers were taller than their fathers. So it looked like heights were "regressing" to the population average height. But a closer analysis shows that any imperfectly correlated variables demonstrate this phenomenon, and the continued "regression" does not occur.
6. Random Sampling: All our discussion of theory of means and of survey polls, and many of the readings, use the concept of random sampling. It is a sampling method that has predictable behaviour (in a probabilistic sense) and most other methods of sampling, like "convenience" sampling such as a mall survey would provide, do not have predictable behaviour.
7. Statistical Design: The planning of experiments and observational studies has been discussed in detail in several of the articles: turkey mail, school choice, tiger prey, and others. The idea is that we need to plan how the data is extracted from the population of interest, in order to describe the quality of the information implied by the data about the population.
8. Graphical Display of Data: I started the course with the gas consumption example because I think it was a good example of how a simple graphing technique extracted information from data. Of course, we have had many other examples as well: Africanized bee invasion, Randomness in the stock market, Advertising as an Engineering Science, …
9. Chi Squared Distribution: Well, we missed this one – but it may come up yet!
10. Modern Computation and Simulation: we had lots of simulation, and of course it requires a lot of computation, but I have kept the details in the background. I hope you have been able to see the usefulness of the technique even though you have not had to cope with the software use.

So even though the course has focused on several data analysis case studies, the main ideas of statistics have been introduced. But there are some aspects of statistical software that I want to make explicit.

The Nature of Statistical Software (R)

Why has it changed the discipline of statistics?

Simulation
Graphics
Regression
Trial and Error Modeling
Less parameterization

What is easy to do, and what is hard to do?

Easy
Common procedures with good data. Repeated analyses.
e.g. simple calculations
re-running a program

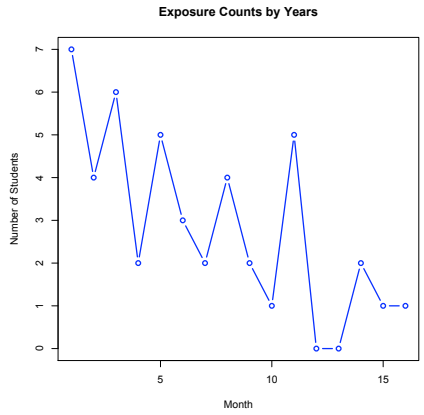
Hard
Data organization, data input/output
non-standard procedures (programming)
Self-evident displays (Captions, defintions, …)
e.g. data cleaning,
writing a program

What are the most basic procedures?

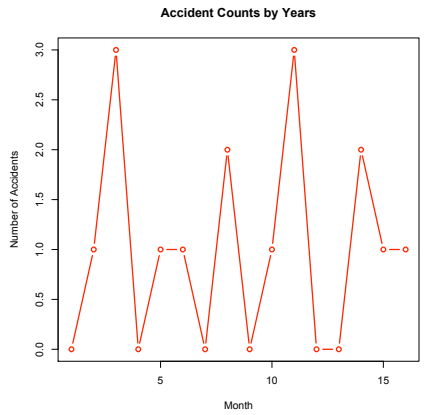
c() and read.table()
mean()
sd()
sum()
plot()
runif()
rnorm()
pnorm()

```
> accidents.df
  exposure count accidents a.over.c
1        6     7         0    0.00
2       18     4         1    0.25
3       30     6         3    0.50
4       42     2         0    0.00
5       54     5         1    0.20
```

```
6      66    3       1    0.33
7      78    2       0    0.00
8      90    4       2    0.50
9     102    2       0    0.00
10    114    1       1    1.00
11    126    5       3    0.60
12    138    0       0     NaN
13    150    0       0     NaN
14    162    2       2    1.00
15    174    1       1    1.00
16    186    1       1    1.00
> mean(accidents)
[1] 1
> sd(accidents)
[1] 1.032796
> plot(count,type="b",lwd=2,col="blue",main="Exposure Counts by
Years",xlab="Month",ylab="Number of Students")
```



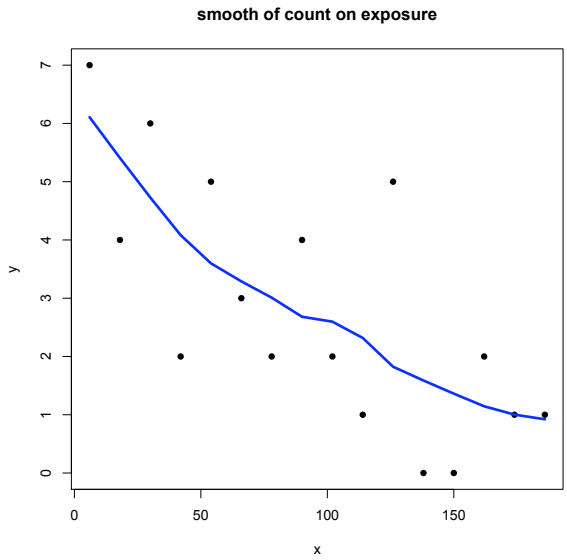

```
> plot(accidents,type="b",lwd=2,col="red",main="Accident Counts by
Years",xlab="Month",ylab="Number of Accidents")
```



In neither case is the mean and SD a reasonable summary. Why?

Although programming is “hard”, once done it is easy to use. For example, I have made up a program called “smo” that will automatically smooth a time series.

- smo(exposure,count,main="smooth of count on exposure")



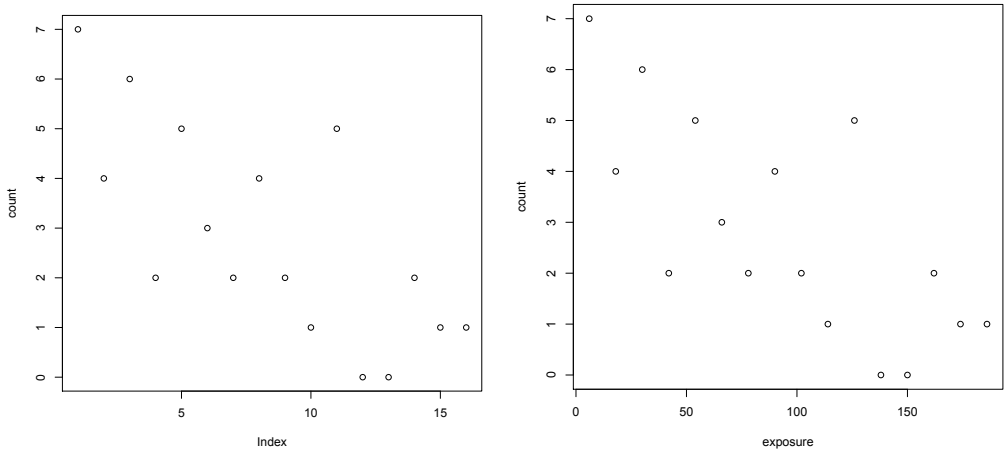
In this case the program is only:

```
> smo
function (x,y,span=.5,col="blue",main="",degree=1,...)
{
yl=loess(y~x,span=span,degree=degree)
plot(x,y,col="black",type="p",main=main,pch=16)
smooth=predict(yl)
lines(x,smooth,col=col,lwd=3)
invisible(smooth=smooth)
}
```

but you do have to know lots of R syntax to make even this simple program.

Q: What role would be played by a residual plot of the blue smooth?

One very nice feature of the best modern software is “Object Oriented Response”. For example, I can say
>plot(count) or
> plot(exposure,count)



The first plot command recognizes that only one variable is provided and so it supplies "index". The second plot command recognizes that it is given two variables and plots them appropriately. The one "plot" command knows how to deal with different objects – it is object oriented.

Software that does not have this feature is much harder to use.

One more example of this object-oriented feature:

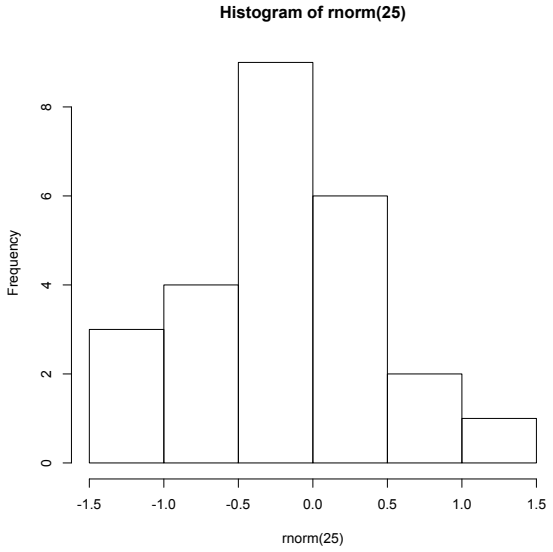
```
> summary(count)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.812   4.250   7.000
>> summary(accident.df)
    exposure       count          a.over.c
 Min.   :  6   Min.   :0.000   Min.   :0.0000
 1st Qu.: 51   1st Qu.:1.000   1st Qu.:0.0500
 Median : 96   Median :2.000   Median :0.5000
 Mean   : 96   Mean   :2.812   Mean   :0.4738
 3rd Qu.:141   3rd Qu.:4.250   3rd Qu.:0.9000
 Max.   :186   Max.   :7.000   Max.   :1.0000
                               NA's   :2.0000
```

Note: a "Quartile" is a particular "Percentile". The first quartile is the 25^{th} percentile, the second quartile is the 50^{th} percentile, and the third quartile is the 75^{th} percentile. The second quartile is also called the median.

Simulation:

We have had many examples of simulation in this course so far. Simple simulations are very easy to do:

```
> runif(5)
[1] 0.9181037 0.6643069 0.8369041 0.2435384 0.4474388
> rnorm(5)
[1]  0.06736807  0.05439180  1.91053709 -1.28081992
0.14323804
> hist(rnorm(25))
```



rnorm() generates data as if it were selected at random from a population of numbers that have a N(0,1) distribution.
runif() generates data as if it were selected at random from a population of numbers that have a uniform distribution over the interval (0,1).

```
> rdunif(5)
[1] 4 4 9 2 4
```

I wrote a little program to generate random digits:

```
> rdunif
function (n=10,k=10,start0=T)
{
     sign=-1
     if (start0==F) {sign=1}
```

```
    a=round(k*runif(n)+0.5*sign,digits=0)
    return(a)
}
```

Why is simulation so useful for statistics?

1. It allows one to study how statistical procedures perform when the population is known exactly. (e.g. sample means distribution, random walk)
2. It allows one to demonstrate the effects of randomness. (e.g. leagues demo, random walk, spatial patterns)
3. It provides the means to arrange random samples from real populations. (e.g. surveys like HIV study)
4. It provides a way to generate data quickly for demonstrating statistical techniques. (e.g. spatial data, random walk)
5. It allows one to mimic real life processes and to use the simulation models so generated to study those real life processes. (e.g. stock market)

Intro to week 11:

11. **Quality Control**. Snee 323-338: Improving the Accuracy of a Newspaper: A Six Sigma Case Study of Business Process Improvement. Doganaksoy, Hahn & Meeker 339-358: Assuring Product Reliability and Safety.

W. Edwards Deming tried to interest the large American Companies (like GE, Ford, Du Pont, ..and many others) in implementing his product improvement process. This was in the 1940s. These companies were doing well and did not want his advice. So Deming went to Japan in 1950 and initiated a revolution in industrial style there that had the effect of seriously hurting American Industry. Eventually, other world economies inclusing those in North America heeded the changes that were needed. This story is partly told in the Snee article (pp 323-337). The role of Deming himself is covered in the Wikipedia entry http://en.wikipedia.org/wiki/W._Edwards_Deming and many other internet sites.
The Snee article talks about "Six Sigma" strategies which is really the modern version of the strategies that Deming initiated. While some of the details may seem a bit ponderous, there are a few ideas that are both very statistical and very useful. One is that a focus on variance reduction in manufacturing is an excellent way to increase profit. Another is that "management by exception" allows an worker to monitor and enhance many complex processes at once. A third is that graphical displays can be usefully employed by workers with minimal statistical education. I recommend that you read the Snee article as preparation for Week 11.