STAT 400 Data Analysis September 23, 2005

Today:

Review of Bisquare (carbon data)
Slicing (Babinet Data)
Density Estimation (birth-death rate data)

We had a few questions left – mostly concerned with slicing:

15. In what context is **slicing** a useful strategy? (p 128)

16. Why are slices in Fig 3.40 (p 128) in a vertical direction? Generalize this.

17. In Fig 3.41 (p129) what role has the loess fit to the subsequent analysis?

18. Explain how Figure 3.43 (p131) is arrived at in the context of estimating the predictive relationship between Y=Babinet Point and X=Cube Root Concentration. (p 131)

19. What is Fig 3.44 (p 132) and what does the analyst look for in this display?

20. What use is the formula on p 133, and what is the basic idea behind it?

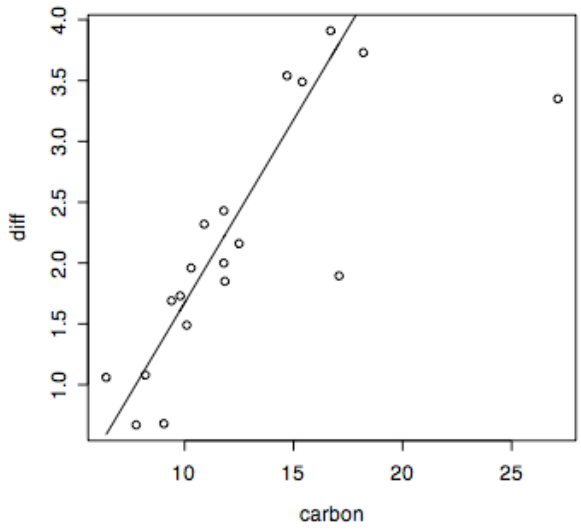
Review of Bisquare:

I sent out this e-mail :

Stat 400

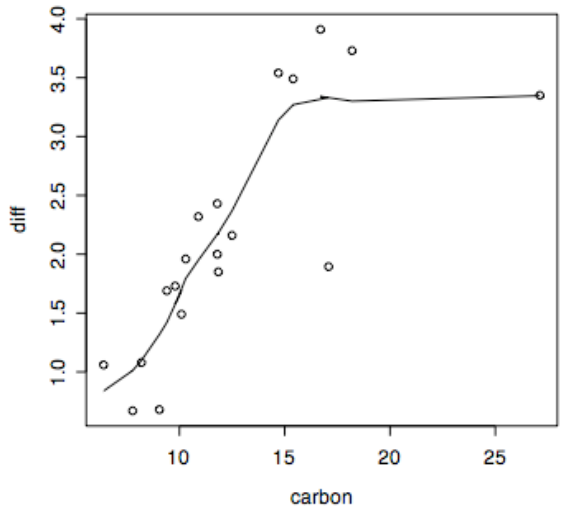
To get that graph on p 115 at the bottom, the result of using bisquare on a straight line fit, use

```
a=rlm(y~x,psi=psi.bisquare)    "rlm is robust linear model"
b=predict(a)
plot(x,y)
lines(x,b)
```



To do a similar thing based on a loess fit,

a=loess(y~x,psi=psi.bisquare)
b=predict(a)
plot(x,y)
lines(x,b)
If you do these both with the carbon data, you can see why Cleveland did not use loess for p 115. It makes the "outlier" look like a regular point.
For the carbon data, just put x=carbon and y=diff where diff = thorium-carbon.
(You will need to
attach(carbon.df)
first, if you want to avoid the carbon.df$carbon etc nuisance.



Overview of bisquare:

It is a method for making a fitting method more robust to outliers – in other words to automatically decrease the sensitivity of the fitting method/model to isolated points that do not fit the method/model well. The procedure is to fit , compute residuals, re-fit using weights on the data determined by the residuals (big residuals lead to small weight), then recalculate the residuals, and refit using new weights, etc until convergence.

Slicing:

In the same situation in which you might use regression (i.e. when you want to predict y from x), you can visualize the data appropriately by visualizing summaries of the slices. See p 128 for a slice. Note that if the x-data are not equally spaced (as is usual) the slices should probably be varying in width. One reasonable policy is to use x-intervals with equal numbers of points – this ensures that the information in each interval is equal, which is desirable when the trend across intervals is to be visualized.

Fig 3.43 on p 131 shows a set of equal-data-numbers intervals. Note that the intervals overlap – this is not usually done for histograms but is very sensible for visualization of trends.

How do you choose the intervals to overlap by a certain number of points and also contain equal numbers of points? The answer is in the formula on p 133. Can you see why it makes sense? Its explained on pp 134-135, but the important thing is to see that there must be a solution, and that it is not quite trivial.

Now go back and answer those last few questions (concerning slicing).

Next,

Density Estimation:

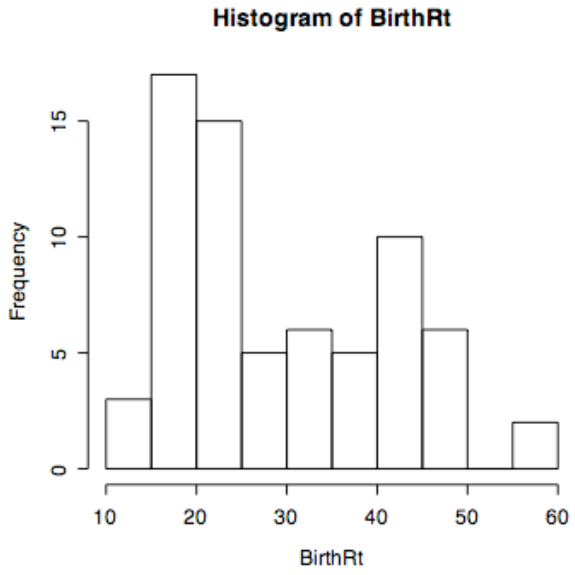
Here is a new data set: birth rates and death rates for 69 countries.

```
> bd.df
   BirthRt DeathRt Country Continent Cont.Code
1     36.4    14.6     alg    AFRICA         1
2     37.3     8.0     con    AFRICA         1
3     42.1    15.3     egy    AFRICA         1
4     55.8    25.6     gha    AFRICA         1
5     56.1    33.1     ict    AFRICA         1
6     41.8    15.8     mag    AFRICA         1
7     46.1    18.7     mor    AFRICA         1
8     41.7    10.1     tun    AFRICA         1
9     41.4    19.7     cam    AFRICA         1
10    35.8     8.5     cey    AFRICA         1
11    34.0    11.0     chi    AFRICA         1
12    36.3     6.1     tai    AFRICA         1
13    32.1     5.5     hkg      ASIA         2
14    20.9     8.8     ind      ASIA         2
15    27.7    10.2     ids      ASIA         2
16    20.5     3.9     irq      ASIA         2
17    25.0     6.2     isr      ASIA         2
18    17.3     7.0     jap      ASIA         2
19    46.3     6.4     jor      ASIA         2
20    14.8     5.7     kor      ASIA         2
21    33.5     6.4     mal      ASIA         2
22    39.2    11.2     mog      ASIA         2
23    28.4     7.1     phl      ASIA         2
24    26.2     4.3     syr      ASIA         2
25    34.8     7.9     tha      ASIA         2
26    23.4     5.1     vit      ASIA         2
27    24.8     7.8     can   AMERICA         3
28    49.9     8.5     cra   AMERICA         3
29    33.0     8.4     dmr   AMERICA         3
30    47.7    17.3     gut   AMERICA         3
31    46.6     9.7     hon   AMERICA         3
32    45.1    10.5     mex   AMERICA         3
33    42.9     7.1     nic   AMERICA         3
34    40.1     8.0     pan   AMERICA         3
```

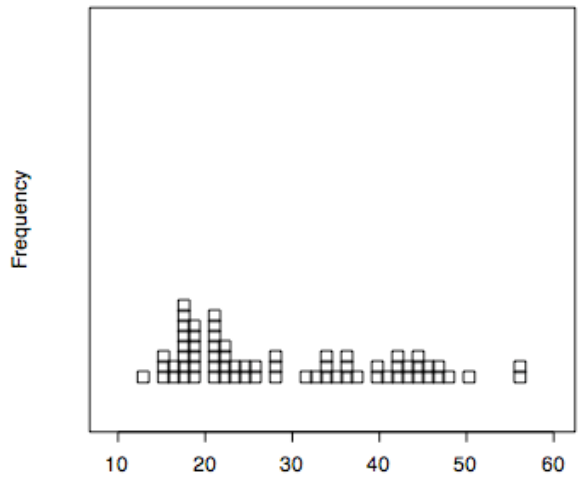
35	21.7	9.6	usa	AMERICA	3
36	21.8	8.1	arg	AMERICA	3
37	17.4	5.8	bol	AMERICA	3
38	45.0	13.5	bra	AMERICA	3
39	33.6	11.8	chl	AMERICA	3
40	44.0	11.7	clo	AMERICA	3
41	44.2	13.5	ecu	AMERICA	3
42	27.7	8.2	per	AMERICA	3
43	22.5	7.8	urg	AMERICA	3
44	42.8	6.7	ven	AMERICA	3
45	18.8	12.8	aus	EUROPE	4
46	17.1	12.7	bel	EUROPE	4
47	18.2	12.2	brt	EUROPE	4
48	16.4	8.2	bul	EUROPE	4
49	16.9	9.5	cze	EUROPE	4
50	17.6	19.8	dem	EUROPE	4
51	18.1	9.2	fin	EUROPE	4
52	18.2	11.7	fra	EUROPE	4
53	18.0	12.5	gmy	EUROPE	4
54	17.4	7.8	gre	EUROPE	4
55	13.1	9.9	hun	EUROPE	4
56	22.3	11.9	irl	EUROPE	4
57	19.0	10.2	ity	EUROPE	4
58	20.9	8.0	net	EUROPE	4
59	17.5	10.0	now	EUROPE	4
60	19.0	7.5	pol	EUROPE	4
61	23.5	10.8	pog	EUROPE	4
62	15.7	8.3	rom	EUROPE	4
63	21.5	9.1	spa	EUROPE	4
64	14.8	10.1	swe	EUROPE	4
65	18.9	9.6	swz	EUROPE	4
66	21.2	7.2	rus	EUROPE	4
67	21.4	8.9	vug	EUROPE	4
68	21.6	8.7	ast	OCEANIA	5
69	25.5	8.8	nzl	OCEANIA	5

Consider first the birthrates. Even though they are not a random sample from any population, it still is a reasonable descriptive task to describe the density.

First lets look at a histogram of the BirthWt variable.



Here is a better graph – a dotplot, more detail, and still easy to absorb visually.



But the histogram allows one to read off frequency and (if so scaled) area and proportion.

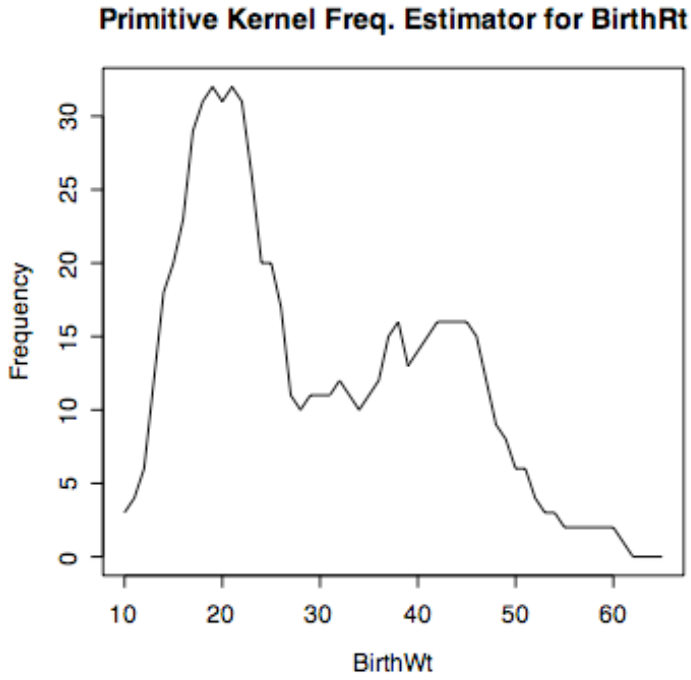
Consider another kind of histogram – created by a "kernel" smoother. Start with a grid of points along the x-axis. At each grid point, count the number of data values within a distance d of the grid point, and use this as the y-value to plot. But first rescale all these y-values (frequencies, really) so that the integral is 1. This is a crude density estimate, and the amount of smoothing is controlled by the choice of d.

Here is the R program:

```
> my.density
function (x,window.width=5,grid=c(10:65))
{
        #This program just counts the data values near the grid points
lg=length(grid)
for (i in 1:lg) {f[i]=length(x[abs(x-grid[i])<window.width])}
plot(grid,f,type="l",xlab="BirthWt",ylab="Frequency",main="Primitive Kernel Freq.
Estimator for BirthRt")
}
```

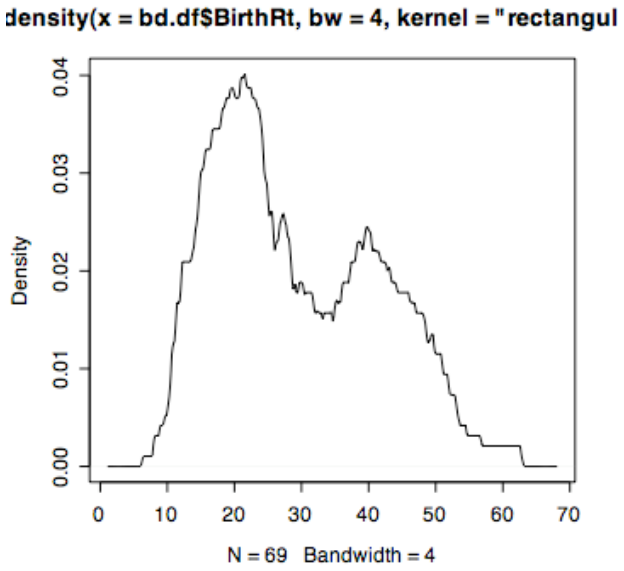
and for

```
my.density(BirthRt,window.width=5,grid=c(10:65))
```



Compare the built in density routine in R

```
plot(density(bd.df$BirthRt,kernel="rectangular",bw=4))
```

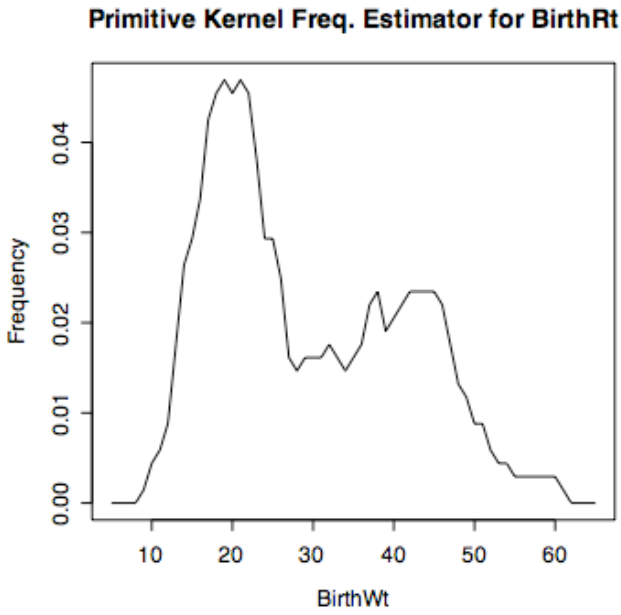


Note the scale difference. How do we get the scale of density right?
Units are relative frequency per unit.

Need to integrate our "relative frequencies", which are for overlapping intervals,
And divide by the integral. Add a couple of lines to the program …

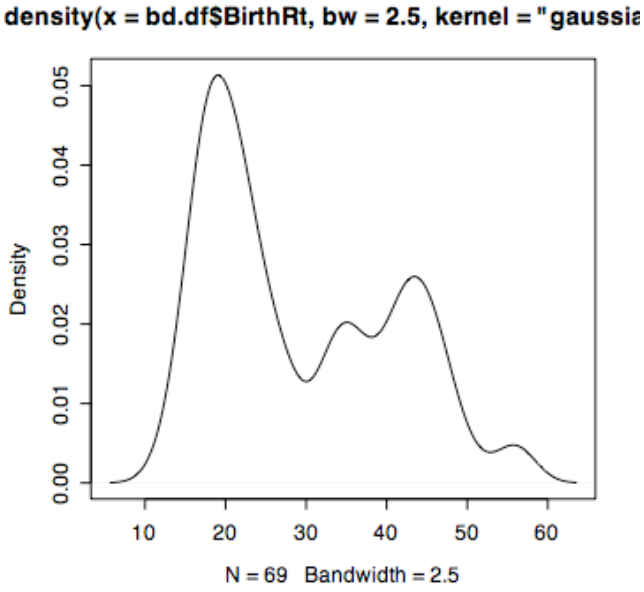
```
>my.density
function (x,window.width=5,grid=c(10:65))
{
        #This program just counts the data values near the grid points
lg=length(grid)
for (i in 1:lg) {f[i]=length(x[abs(x-grid[i])<window.width])}
integral=sum(f)
new.f=f/integral
plot(grid,new.f,type="l",xlab="BirthWt",ylab="Frequency",main="Primitive Kernel
Freq. Estimator for BirthRt")
return(list(grid,new.f))
}
```

and the result is



A result very similar to the R-built-in algorithm. However, smoother kernels are available .

plot(density(bd.df$BirthRt,kernel="gaussian",bw=2.5))



The point is: density estimation is not a black art – it is a simple extension of ideas involved in constructing a histogram, but a marked improvement on the histogram.

Next time we will look at the joint distribution of BirthRt and DeathRt.

Here is another little program I used to show the result of the parametric approach to density esxtimation in this BirthRt data.

```
x=5:65
y=dnorm(x,mean=29.25,sd=11.69)
z=rep(0,length(BirthRt))
plot(x,y)
lines(BirthRt,z,type="p")
```

Exercise to hand in for Friday, Sept 30:

You will be sent a data set "bimbo.txt". It involves deliveries and sales of loaves of bread to a single retail outlet. ("Bimbo" is the third largest bakery in the world, based in Mexico). The data gives 53 weeks of data, 6 days per week, a time series. We will be eventually examining the relationship between deliveries and sales, but to do a good job of this we need to eliminate the "seasonal" effect, if any. Prepare a version of the data for which the seasonal effect is eliminated. E-mail your adjusted data set to me, and hand in a paper copy showing a graph of the seasonal trend you use to adjust the data.