STAT 400 Data Analysis September 17, 2003

Today: Density Estimation
Minitab Macros and Programming

Density Estimation:

A univariate histogram of sample data gives a rough idea of the shape of the density function. However, the need to use rectangular frequency indicators produces graphs that do not look like density functions at all (unless the data set is huge and the number of bins huge). There is a simple way to improve on this. First I want to introduce a new data set:

B-D Rates:

Row	birth-rt	death-rt	country	group
1	36.4	14.6	alg	1
2	37.3	8.0	con	1
3	42.1	15.3	egy	1
4	55.8	25.6	gha	1
5	56.1	33.1	ict	1
6	41.8	15.8	mag	1
7	46.1	18.7	mor	1
8	41.7	10.1	tun	1
9	41.4	19.7	cam	1
10	35.8	8.5	cey	1
11	34.0	11.0	chi	1
12	36.3	6.1	tai	1
13	32.1	5.5	hkg	2
14	20.9	8.8	ind	2
15	27.7	10.2	ids	2
16	20.5	3.9	irq	2
17	25.0	6.2	isr	2
18	17.3	7.0	jap	2
19	46.3	6.4	jor	2
20	14.8	5.7	kor	2
21	33.5	6.4	mal	2
22	39.2	11.2	mog	2

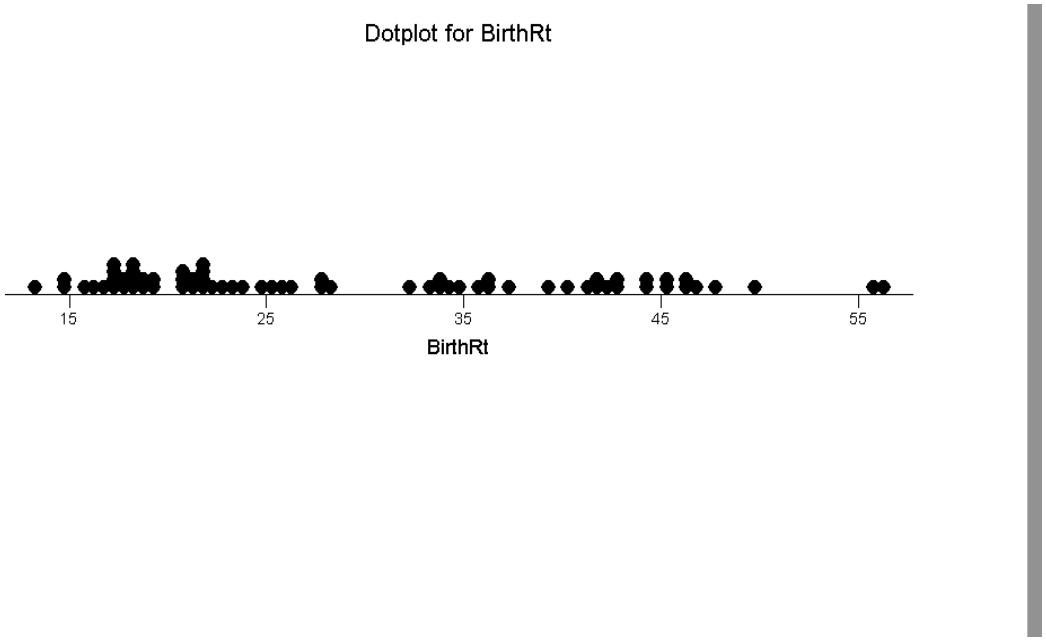
23	28.4	7.1	phl	2
24	26.2	4.3	syr	2
25	34.8	7.9	tha	2
26	23.4	5.1	vit	2
27	24.8	7.8	can	3
28	49.9	8.5	cra	3
29	33.0	8.4	dmr	3
30	47.7	17.3	gut	3
31	46.6	9.7	hon	3
32	45.1	10.5	mex	3
33	42.9	7.1	nic	3
34	40.1	8.0	pan	3
35	21.7	9.6	usa	3
36	21.8	8.1	arg	3
37	17.4	5.8	bol	3
38	45.0	13.5	bra	3
39	33.6	11.8	chl	3
40	44.0	11.7	clo	3
41	44.2	13.5	ecu	3
42	27.7	8.2	per	3
43	22.5	7.8	urg	3
44	42.8	6.7	ven	3
45	18.8	12.8	aus	4
46	17.1	12.7	bel	4
47	18.2	12.2	brt	4
48	16.4	8.2	bul	4
49	16.9	9.5	cze	4
50	17.6	19.8	dem	4
51	18.1	9.2	fin	4
52	18.2	11.7	fra	4
53	18.0	12.5	gmy	4
54	17.4	7.8	gre	4
55	13.1	9.9	hun	4
56	22.3	11.9	irl	4
57	19.0	10.2	ity	4
58	20.9	8.0	net	4
59	17.5	10.0	now	4
60	19.0	7.5	pol	4
61	23.5	10.8	pog	4
62	15.7	8.3	rom	4

63	21.5	9.1	spa	4
64	14.8	10.1	swe	4
65	18.9	9.6	swz	4
66	21.2	7.2	rus	4
67	21.4	8.9	vug	4
68	21.6	8.7	ast	3
69	25.5	8.8	nzl	3

This gives birth-rates and death-rates for 69 different countries (identified by 3-letter code and continent).

Lets focus on the birth rates first:

Distribution of birth rates?

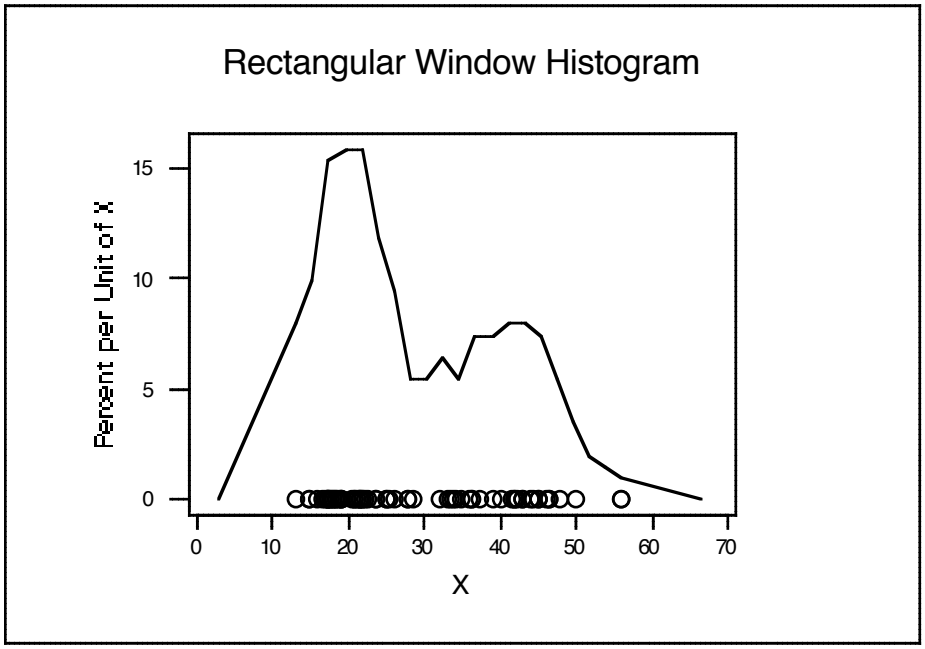


Density? Although the data is not sample data, it is still reasonable to describe the distribution as a continuous curve.

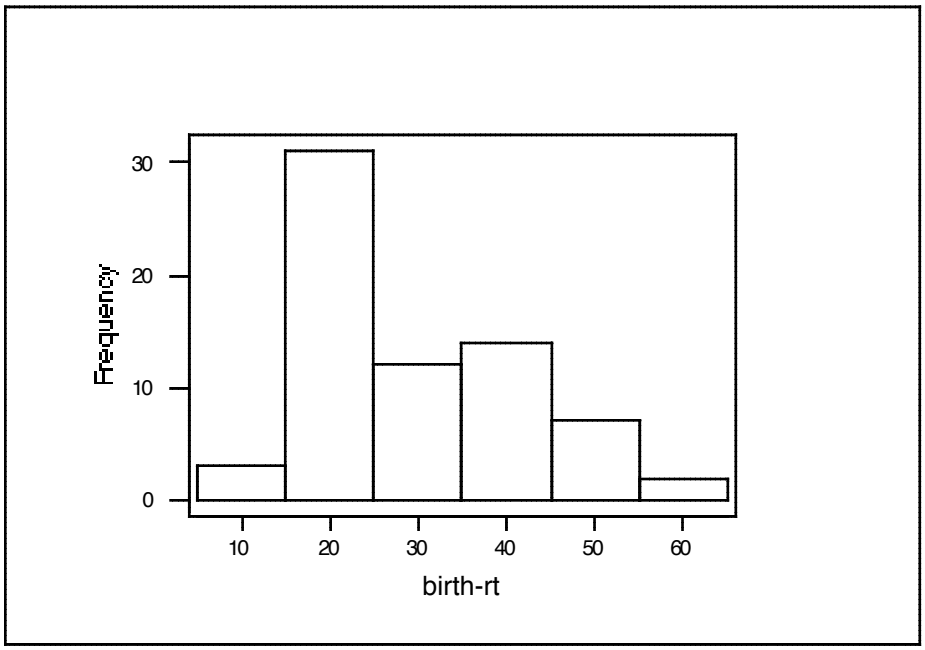
Here is a procedure:

1. choose as grid of **m** x-values
2. choose a distance **d** within which you will count frequencies
3. At each grid point, plot as a y value the number of data values within d of the grid point.

This is what you get for **m**=23, **d**=5 for the birth rates
(The program chooses m and d automatically-however, you can override this.)



Compare the MINITAB default histogram:



Here is that histo program again:

```
Gmacro
histo.mac             # this program is a primitive density estimator
        # it assumes data is in c1 (it wipes out some other columns)
         # it expects the data to be in C1, any length.
erase c2-c4         # it simply uses a rectangular window to accumulate
brief 0                          #frequency
                      # the frequency close to each of 23 grid points
mini c1 k1                    # these commands set up the range for the grid
maxi c1 k2
n c1 k10                       # get sample size
let k3=(k2-k1)/k10**.5  # window half-width: use a larger DENOM for
                                                              #LESS smoothing
let k13=20
set c2
0:k13                             # increase this for more grid points
end
let c2=k1+(c2/k13)*(k2-k1)
let k6=k1-2*k3
let k7=k2+2*k3
stack k6 c2 k7 c2
let k14=k13+3
do k4=1:k14          # For each grid point, accumulate the frequency
let c3=abs(c1-c2(k4))
let c3=c3 LE k3
sum c3 k5
let c4(k4)=k5                  # store the accumulated frequency
enddo
brief 2
sum c4 k9     # these commands standardize the pseudofrequencies so
#n c1 k10     # they "integrate" to 1 (this n command already done)
let c4=100*c4*k10/k9*(1/(k2-k1))
name c4 'pctperX' c2 'X'      # these commands set up the graph of the
                                              #resulting density est.
let c5=0*c1
plot c4*c2 c5*c1;
connect;
type 1 0;
symbol;
type 0 1;
overlay;
```

```
jitter .025 .025;
title 'Windogram';
Axis 1;
Label "Birthweight (gms)";
Axis 2;
Label "Percent per gm".
endmacro
```

Some questions:

How does the program determine d, the half-window width?
How many grid points does one usually need?
Why does the choice of d>1 not make the area > 100%?
How does the data plot compare with the more usual dotplot?
Is the jitter useful?
Why is this program called a "primative estimator"? How could it be improved?
How do you make this program available to MINITAB?
How do you run this program in MINITAB?
Do I have to modify the program every time I want to change a parameter value?

Use on another data set: Birthtimes, or simulated Normal.

Assignment 3 Due Wed Sept 24: (Note: Keep reading Ch 3 to end – next topic)

Modify the "histo" program (or write one of your own) to incorporate all of the data in the estimation of the density at each grid-point. Your submission should include:

1. a copy of your program
2. an explanation of your modification
3. an example of its output on the birth-rate distribution

(My objective in this assignment is to ensure that you understand what is called the "kernel" method of density estimation, and also to ensure that you know how to run this kind of program (not merely a procedure from a stat package). Needless to say, I do not want you to use a ready-made procedure from a stat package.)