Paper

Simulate Data Distribution From Quantiles

Xia Ke Shan, CBC(Beijing) Credit Management Co. Ltd., Beijing, China
Matthew Kastin, I-Behavior, Inc., Louisville, Colorado
Arthur S. Tabachneck, Ph.D., AnalystFinder, Inc., Thornhill, Ontario Canada

ABSTRACT

Sometime, we only have data quantiles, but we want know how the real data looks like, what is data distribution ? Maybe it was Normal distribution or Uniform distribution. This paper is trying to simulate Uniform data according to its data quantiles (of course, you can use Normal distribution to simulate data or some other data distribution). Basic idea is using Table distribution to sample these quantiles with some probability (calculated by quantiles) and simulate data assuming it is Uniform distribution. It is a method called Smooth Bootstrap method (Rick Wicklin said).

INTRODUCTION

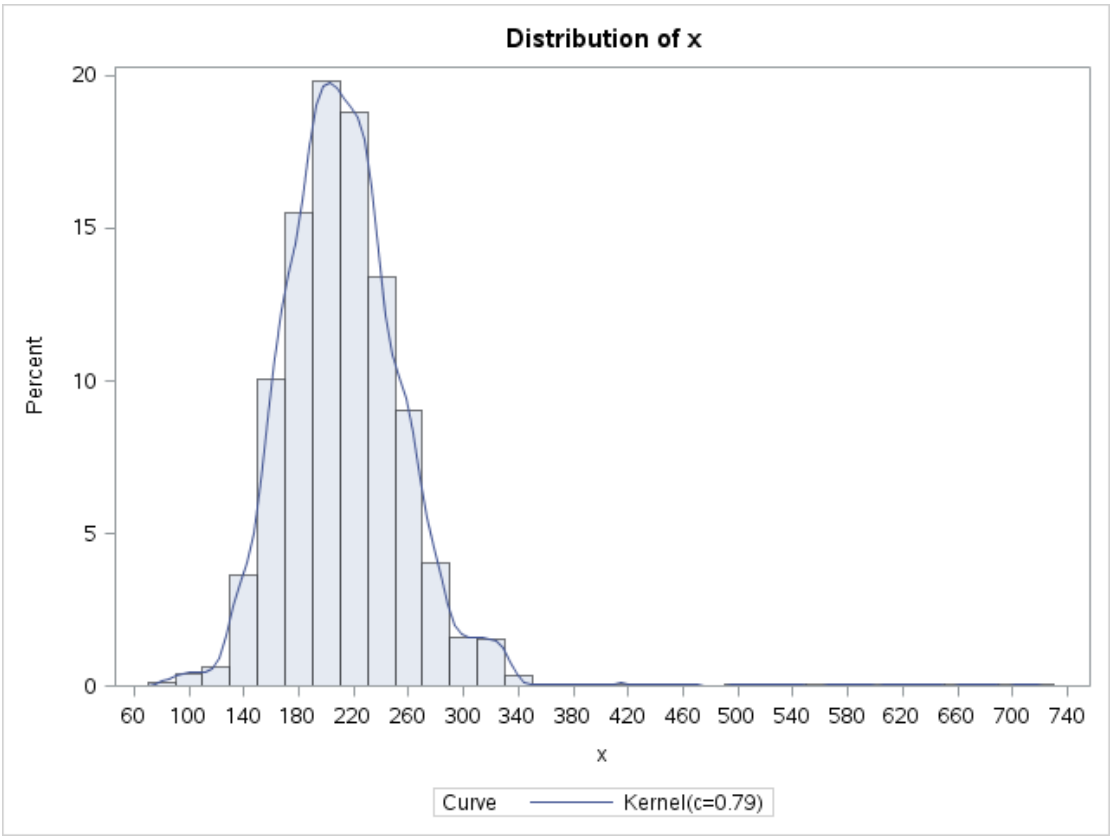
Actually the data is taken from Rick Wicklin's blog (http://blogs.sas.com/content/iml/2014/06/18/distribution-from-quantiles.html). He used a different way to simulate data.

DATA STEP CODE

This code is for SAS programmer who can understand how it is done easily.

```
data have;
infile cards truncover expandtabs;
input Quantile Estimate ;
prob=dif(Quantile);
width=dif(Estimate);
cards;
0    80
.01  128
.05  151
.10  162
.25  184
.50  209
.75  236
.90  265
.95  284
.99  333
 1   727
;
run;
```

```
proc sql;
select prob into : probs separated by ','
 from have
  where prob is not missing;

select Estimate  into : points separated by ','
 from have
  where Quantile ne 1;

select width into : widths separated by ','
 from have
  where width is not missing;

select count(*)-1 into : count from have;
quit;

%let n=10000; /* sample size */
data want;
call streaminit(1234);
array points(&count) _temporary_ (&points);
array widths(&count) _temporary_ (&widths);

do i=1 to &n;
 n=rand('table',&probs);
 x=points{n}+ceil(widths{n}*rand('uniform'));
 output;
end;
drop i;
run;
proc univariate data=want;
var x;
histogram x/kernel;
run;
```



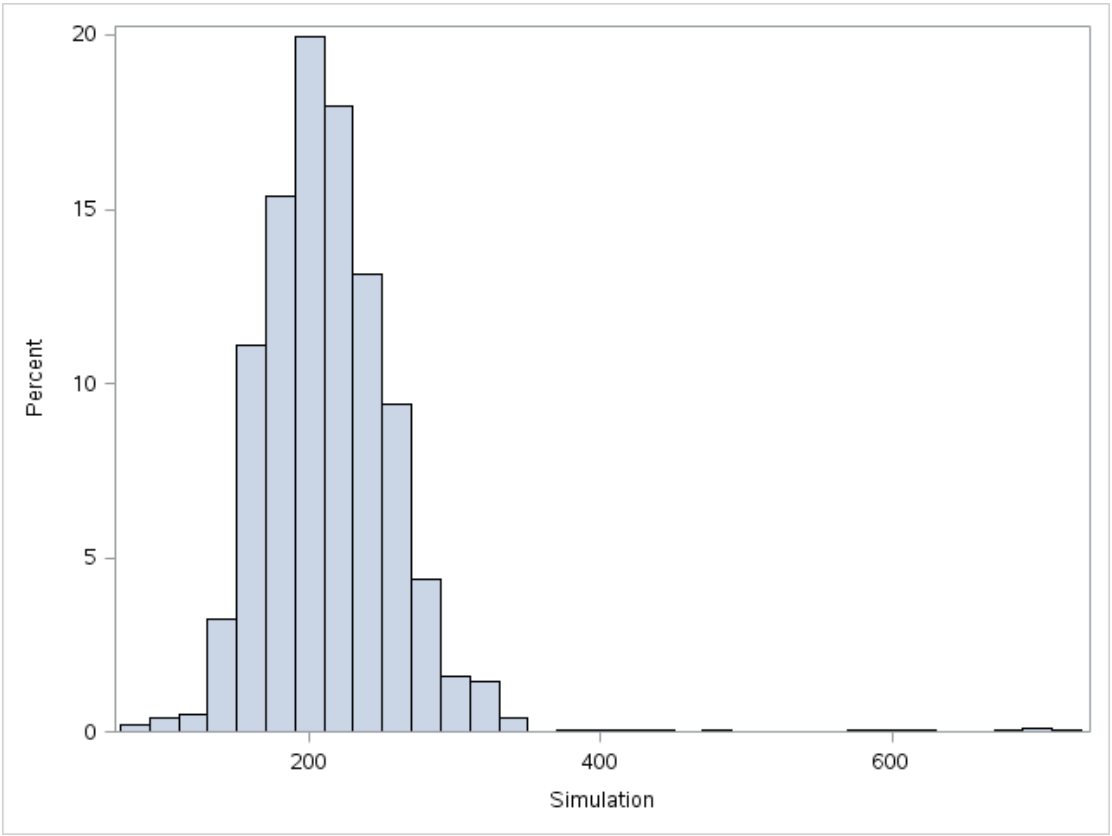
Quantiles (Definition 5)	
Level	Quantile
100% Max	727.0
99%	333.0
95%	284.5
90%	265.0
75% Q3	236.0
50% Median	209.0
25% Q1	184.0
10%	162.0
5%	150.0
1%	129.0
0% Min	81.0

IML CODE

This code is for IML programmer, since IML is a sharp tool for simulating data.

```
proc iml;
/*
quantiles of total cholesterol from NHANES study
http://blogs.sas.com/content/iml/2014/06/18/distribution-from-qua
ntiles.html
*/

Quantile = {0 , .01, .05, .10, .25, .50, .75, .90, .95, .99, 1};
Estimate = {80, 128, 151, 162, 184, 209, 236, 265, 284, 333, 727};

point=t(remove(Estimate,nrow(Quantile)));
width=t(remove(dif(Estimate),1));
prob=t(remove(dif(Quantile),1));

print Quantile Estimate point width prob;


/*Start to simulate uniform data*/
n=10000; /* sample size */
call randseed(123456789);

points=t(sample(point,n,'replace',prob));
widths=j(n,1);
do i=1 to nrow(point);
 widths[loc(points=point[i])]=width[i];
end;
eps=randfun(n,'uniform');

data=points+ceil(widths#eps); /* This is the simulate data */

/* Check the simulate data */
call histogram(data) label="Simulation" ;
call qntl(SimEst, data, Quantile);
print Quantile[F=percent6.] Estimate SimEst[F=5.];
quit;
```

Quantile	Estimate	point	width	prob
0	80	80	48	0.01
0.01	128	128	23	0.04
0.05	151	151	11	0.05
0.1	162	162	22	0.15
0.25	184	184	25	0.25
0.5	209	209	27	0.25
0.75	236	236	29	0.15
0.9	265	265	19	0.05
0.95	284	284	49	0.04
0.99	333	333	394	0.01
1	727			



Quantile	Estimate	SimEst
0%	80	81
1%	128	126
5%	151	152
10%	162	162
25%	184	183
50%	209	209
75%	236	236
90%	265	265
95%	284	283
99%	333	333
100%	727	727