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.

```sas
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;
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-quantiles.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;
<table>
<thead>
<tr>
<th>Quantile</th>
<th>Estimate</th>
<th>SimEst</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>90</td>
<td>91</td>
</tr>
<tr>
<td>1%</td>
<td>126</td>
<td>126</td>
</tr>
<tr>
<td>5%</td>
<td>151</td>
<td>152</td>
</tr>
<tr>
<td>10%</td>
<td>162</td>
<td>162</td>
</tr>
<tr>
<td>25%</td>
<td>184</td>
<td>183</td>
</tr>
<tr>
<td>50%</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>75%</td>
<td>236</td>
<td>236</td>
</tr>
<tr>
<td>90%</td>
<td>285</td>
<td>285</td>
</tr>
<tr>
<td>95%</td>
<td>284</td>
<td>283</td>
</tr>
<tr>
<td>99%</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>100%</td>
<td>727</td>
<td>727</td>
</tr>
</tbody>
</table>