MAT 265
Assignment 7
Ilia Nikiforov

Sample size and representativeness

Abstract

In this report, I repeatedly draw samples of increasing size to demonstrate that higher sample size results in a more representative sample. I use all checkouts in 2018 as my population and use item type distribution as the main parameter to evaluate how well samples represent original data. I was able to achieve acceptable results with sample sizes of 50000 and 100000 (1.5% and 3% of the population).

Report

I will use the data from 2018. Query 1 will get the counts of all item types checked out in 2018 and select top 15 item types (to make comparison easier). The following Queries (2,3,4,5,6,7,8,9) will repeat the same process, but instead of using all available data will draw samples with increasing sample size (100,500,1000,2000,5000,10000,20000,50000,100000) by manipulating the limit() command.

QUERY 1

```
select

itemType,

count(*) as counts

from spl_2016.inraw

where year(cout)=2018

group by 1

order by 2 desc

limit 15
```

QUERIES 2,3,4,5,6,7,8,9

```
select

itemType,

count(*) as counts

from (

select *

from spl_2016.inraw x

where year(cout)=2018

order by rand(123)
```

limit 100) y

group by 1

order by 2 desc

limit 15

RESULT

The results are stored in the csv files with corresponding names (itemtypes_all_data and itemtypes_sample_N). The following table visualizes the proportion of different item types from each draw ("counts" corresponds to the original, full data table, "counts_N" correspond to each of the samples). For reference, the total number of counts in the original data is around 3.3M. That means, that the highest sample (100k) is about 3% of the population. Missing values are highlighted with red.

itemType	counts	counts_100	counts_500	counts_1000	counts_2000	counts_5000	counts_10k	counts_20k	counts_50k	counts_100k
acbk	31.937193	32.000000	30.400000	31.400000	31.700000	33.173269	32.966483	32.507879	32.420046	32.135816
jcbk	30.951233	36.000000	30.800000	30.500000	31.800000	30.612245	30.795398	30.651859	30.718889	30.930961
acdvd	21.707574	15.000000	22.200000	21.500000	21.500000	21.328531	21.160580	21.791986	21.496618	21.514275
accd	7.235069	8.000000	8.800000	8.000000	7.950000	7.262905	7.563782	7.619191	7.367010	7.321198
jcdvd	3.214894	3.000000	3.400000	3.100000	2.800000	2.841136	2.841421	2.836560	3.150142	3.160244
pkbknh	3.204468	5.000000	3.200000	4.000000	2.850000	3.101240	3.001501	2.936615	3.152143	3.213281
jccd	1.046956	1.000000	1.200000	1.000000	0.950000	0.920368	0.970485	0.920506	0.974663	0.999710
ucfold	0.206024	nan	nan	nan	nan	0.180072	0.240120	0.225124	0.206140	0.198141
acmus	0.122859	nan	nan	nan	0.100000	0.080032	0.080040	0.095052	0.106072	0.127090
aceq	0.106951	nan	nan	nan	0.050000	0.180072	0.120060	0.130072	0.138094	0.131093
bcbk	0.100630	nan	nan	0.200000	0.100000	0.120048	0.090045	0.100055	0.098067	0.098070
dcillb	0.075345	nan	nan	0.100000	0.050000	0.040016	0.060030	0.070039	0.080054	0.083059
jckit	0.038556	nan	nan	0.100000	0.100000	0.100040	0.060030	0.065036	0.046031	0.041029
areqnh	0.036250	nan	nan	nan	nan	nan	nan	0.030017	0.030020	0.031022
acfold	0.015998	nan	nan	nan	nan	nan	nan	nan	0.016011	0.015011
bcdvd	nan	nan	nan	0.100000	0.050000	0.040016	0.020010	nan	nan	nan
acrec	nan	nan	nan	nan	nan	0.020008	nan	nan	nan	nan
areq	nan	nan	nan	nan	nan	nan	0.030015	0.020011	nan	nan

We can see several things:

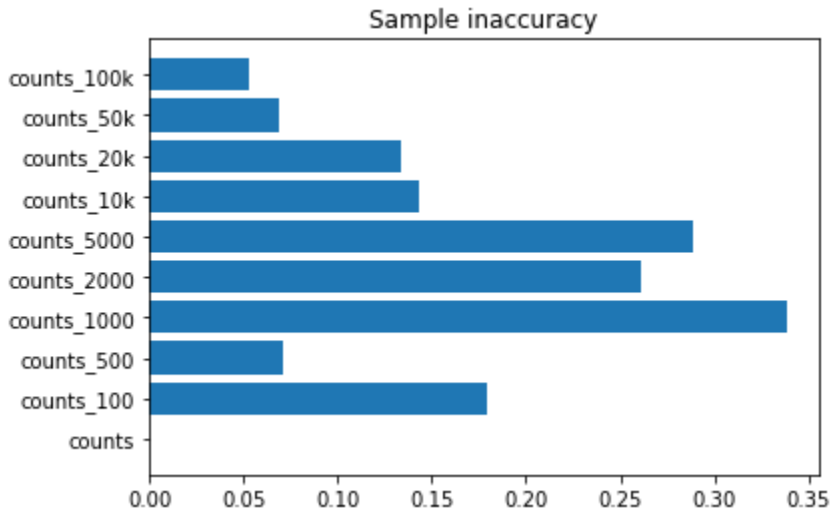
1) Low sample sizes tend to miss the range of item types present in the original data. They either exclude certain types (see counts_100) or also include certain types that are not in the original data (see counts_5000).
2) The higher the sample size, the more accurate this category representation is. We start to see an accurate situation at sample size of 50k (1.5% of the population) and 100k (3% of the population).

Next, I would like to see how closely the proportions of different item types in my samples match the proportions of the original data. While this table does give some idea, I will divide all columns by the "counts" column, effectively centering the results on the original data. The result is displayed below:

itemType	counts	counts_100	counts_500	counts_1000	counts_2000	counts_5000	counts_10k	counts_20k	counts_50k	counts_100k
acbk	1.000000	1.002000	0.952000	0.983000	0.993000	1.039000	1.032000	1.018000	1.015000	1.006000
jcbk	1.000000	1.163000	0.995000	0.985000	1.027000	0.989000	0.995000	0.990000	0.992000	0.999000
acdvd	1.000000	0.691000	1.023000	0.990000	0.990000	0.983000	0.975000	1.004000	0.990000	0.991000
accd	1.000000	1.106000	1.216000	1.106000	1.099000	1.004000	1.045000	1.053000	1.018000	1.012000
jcdvd	1.000000	0.933000	1.058000	0.964000	0.871000	0.884000	0.884000	0.882000	0.980000	0.983000
pkbknh	1.000000	1.560000	0.999000	1.248000	0.889000	0.968000	0.937000	0.916000	0.984000	1.003000
jccd	1.000000	0.955000	1.146000	0.955000	0.907000	0.879000	0.927000	0.879000	0.931000	0.955000
ucfold	1.000000	nan	nan	nan	nan	0.874000	1.165000	1.093000	1.001000	0.962000
acmus	1.000000	nan	nan	nan	0.814000	0.651000	0.651000	0.774000	0.863000	1.034000
aceq	1.000000	nan	nan	nan	0.468000	1.684000	1.123000	1.216000	1.291000	1.226000
bcbk	1.000000	nan	nan	1.987000	0.994000	1.193000	0.895000	0.994000	0.975000	0.975000
dcillb	1.000000	nan	nan	1.327000	0.664000	0.531000	0.797000	0.930000	1.063000	1.102000
jckit	1.000000	nan	nan	2.594000	2.594000	2.595000	1.557000	1.687000	1.194000	1.064000
areqnh	1.000000	nan	nan	nan	nan	nan	nan	0.828000	0.828000	0.856000
acfold	1.000000	nan	nan	nan	nan	nan	nan	nan	1.001000	0.938000
bcdvd	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
acrec	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
areq	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Note that missing values from the original data were propagated into the sample columns as the result of this operation (which is normal).

This table shows the mismatches between sample item type distributions and population item type distribution. The main criterion for accuracy here is to have all numbers as close to 1 as possible. For the next step, I will get an average inaccuracy value for each sample size (by subtracting 1 from each column, taking the absolute value of the results and averaging the resulting vectors) :



We have to keep in mind that this graph visualizes sample inaccuracy only for those item types that are present in each sample. Clearly, the level of inaccuracy goes down as sample size increases (reaching the

lowest value of 5.3% for the sample size of 100000). Interestingly, there is a fairly small inaccuracy for one of the smallest sample size – 500. We have to be mindful, however, that this sample only has 7 out of 15 item categories.

That being said, the best sample size selection should balance three things in this case: representativeness of different item types, accuracy of the sample distribution, and the lowest possible sample size for computational efficiency. 50k and 100k samples are the only ones able to represent all categories, so they will make the short list. The choice between these two is the matter of trade off between inaccuracy (6.9% vs 5.2%) and size (1.5% and 3% of the total 3.3M population, respectively).

Conclusion

This report shows that increasing sample size will lead to decreased sample error. Typically in statistics, a sample size of 10% is considered the best. However, since the population dataset is very big in my case (3.3M rows), I was able to achieve reasonable results with far smaller samples (1.5% and 3%, which is, however, still a lot, given size of the data)