RESEARCH

Assignment 5

Suresh Kumar Choudhary, Emilio Kuhlmann and Sofya Laskina

Full list of author information is available at the end of the article

Abstract

The goal of the project: Need to implement, estimate and compare the variance on different sample sizes using different sampling techniques on the given dataset ie. Simple random sampling without replacement and stratified sampling.

The main result of the project: We achieved a reasonably small sample variance by performing stratified sampling.

Personal key learnings: We learned the importance of Sampling and learned how and when to apply SRR, SRS and stratified sampling.

Estimated working time: 8

Project evaluation: 1

Number of words: 1189

1 Scientific background

Real world available data is practically unlimited. Since we lack infrastructure and means to record or even process all the data available to us on a topic of interest, we need to extract a smaller, more manageable portion of the data. This we can then use to process and gain insights from. Obviously we cannot simply extract this smaller portion, i.e. sample, in whatever way as this will give us biased data. Say we wanted to gain insights on how people will vote. If the sample is taken from only one, conservative neighbourhood we will not have data representative of the total population.

This is where random sampling comes in. SRR (simple random sampler with replacement) and SRS (simple random sampler without replacement) are two conceptually basic, unbiased ways of picking samples from a population. SRR simply picks k samples from the sample space each with the same probability, and SRS picks k samples that do not repeat. Drawing a parallel to the voting poll, SRR could ask the same person zero, once or several times how they're going to vote. SRS would ask a given person at most once about their voting preferences.

Stratified sampling takes this a step further. It divides population into groups, draws samples from each group and finally joins selected individuals from the groups into one large sample. This avoids small population groups of staying unsampled. Again drawing the parallel, stratified sampling in a real-world-population would be splitting the people into rural, sub-urban and urban population and then drawing from each group.

One further interesting application for sampling is to balance out unbalanced datasets. For instance, when performing breast cancer analysis only a small percentage of samples will be diseased ones. For machine learning algorithms to 'learn' decision functions a distribution of approximately half of one label and half of the other is advantageous. Here undersampling, oversampling and synthetic sampling algorithms can be used to balance out datasets. This, however, was not a part of the exercises.

2 Goal

In this exercise sheet we are aimed to learn how to meaningfully sample data in order to not only minimize the sample mean that we learned during the lecture but also the variance of sample mean. We also are supposed to analyze data first, to choose good subdividing parameter to perform stratified sampling.

3 Description of Data Set and Exploratory Data Analysis

Task 1 & 2: Download the dataset, learn about it and its individual variables

This week's dataset is the player database of the popular FIFA19 football videogame. The dataset is of a manageable if not 'toy example' size. Concretely it consists of 18207 rows and 88 columns. It could be said that the data consists of three distinct types of features.

- Real world Data. Essential information about a given player, such as name, age, nationality, club, wage, weak foot, etc...
- Performance values. In the FIFA game, video-game players can choose a team based on how 'good' football players are. FIFA stores how good a player is based on certain performance criteria ranging from 0-100, such as overall, headers, penalties, dribbling, speed, passing, etc...
- Internal values. These are links to images, such as flag, club logo, or a player's face. They are used by the game to nicely display the players, but are of little use outside of the game.

Also we found out, There are 76 columns which have missing values, top 5 missing values column can be found in figure 6. We found few more more things about data like countries having the most players, number of players by position, by preferred foot, by work rate, by body type etc. in the dataset, which can be seen in the figures 8, 9 and 7 respectively.

4 Discussion and comparison of the results of the data sampling

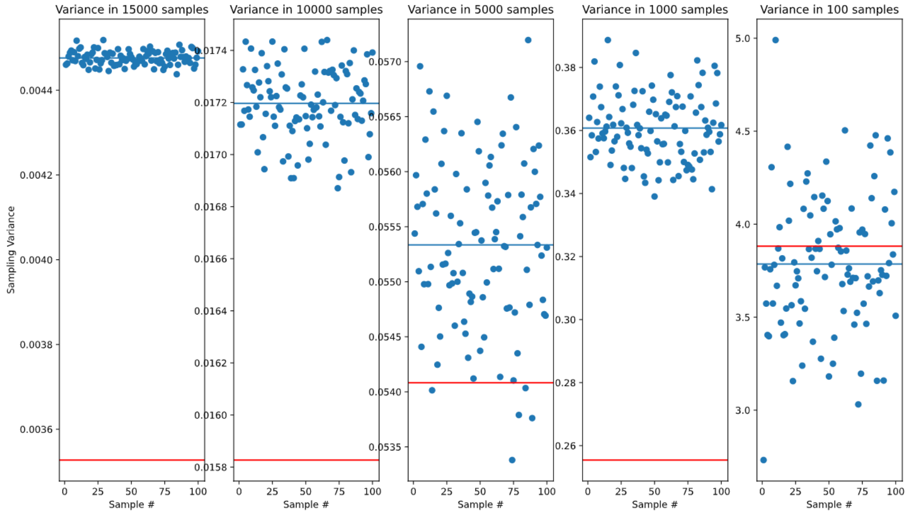
Task 3: SRS Sampling of the dataset for different sample sizes

In this part we sub-sampled our data to 15000, 10000, 5000, 1000 and 100 data points with SRS. Similarly to the lecture, for each size we performed 100 iteration to sample a data and computed mean for each of iterations. We then computed variance of mean of each sample. At the same time we were calculating the analytical value given by formula $Var(\hat{T}) = \frac{\sigma^2}{k}(1 - \frac{k-1}{N-1})$, where $Y_i \in \mathbb{R}$, $\hat{T} = \frac{1}{k}\sum_{i=1}^{k} y_i$ and σ^2 is variance of $\{y_1, \ldots, y_n\}$, which can be found in table 1. Figure 1 Shows the differences in the computed variances. Firstly, we observe that both variances grow with reducing size of the sample. Secondly, manual computation underestimates variance in bigger sample size in comparison to the analytical value. Only when sampling to 100 data points it outgrows the analytical one.

Table 1: SRS Sampling of the dataset for different Sample size with variance

Sample Size	Variance
15000	0.0045
10000	0.0172
5000	0.0553
1000	0.3608
100	3.7858

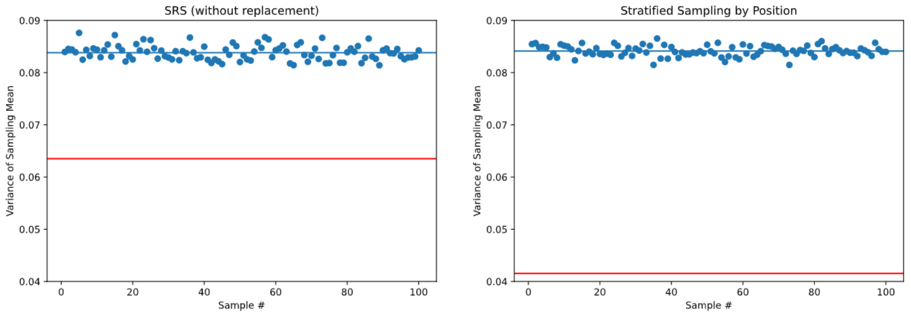
Figure 1: Comparison of calculated and analytical variance of sample mean for different sampling size performed with SRS. Red line denotes variance of mean calculated manually and blue dots denote variance of sample for each of 100 iterations calculated with a formula. Blue line is their mean and drown for convenience.



Task 4: For Finishing Score- Subdiving the Srata of Positions and comparing the variance with SRS Sampling

In this part we discovered, how the stratified sampling performs by performing similar computations as in previous step. This yielded following results(Figure 2). We again observe that manually estimated variance is smaller than analytical one. Interestingly, in SRS it is closer to the analytical Variance of sample mean. However, in both methods the difference is very small between the two variances(0.02 and 0.04 respectively). We also observe, that in both methods analytically estimated Variance is similar and fluctuates around 0.85, however, stratified sampling delivers such data points that variance is more dense around their mean value. In any case, the total Variance of 0.85 seems to be a good result. We think, the strata performs well here because the Finishing variable might be dependant on Position, since it describes the accuracy of shots from within the penalty area and there are definitely such positions, which account on the ability of a player to perform a shot from comparable distance during the normal game time.

Figure 2: Comparison of calculated and analytical variance of sample mean for SRS and Stratified sampling divided by category Position. Red line denotes variance of mean calculated manually and blue dots denote variance of sample for each of 100 iterations calculated with a formula. Blue line is their mean and drown for convenience.



Task 5: Performing Strata Sampling and compare with Simple Random Sampling - For Overall Score

Similar to the Part 3 we observe that variances grow with the decreasing size of the samples, which can be seen table 2. However, this time it is much smaller and is below 1(Figure 3). As for stratified sampling, it performed worse with the Overall variable(Figure 4). The manually calculated value was 0.04 and was much bigger than analytical one, so that it is not shown. Also we observe that the data points for analytical variance are much more dispersed around their mean. The reasons for that could be the choice of strata. We do not think that position on fields reflects general effectiveness of a player.

Table 2: SRS Sampling with overall variable for different Sample size with variance

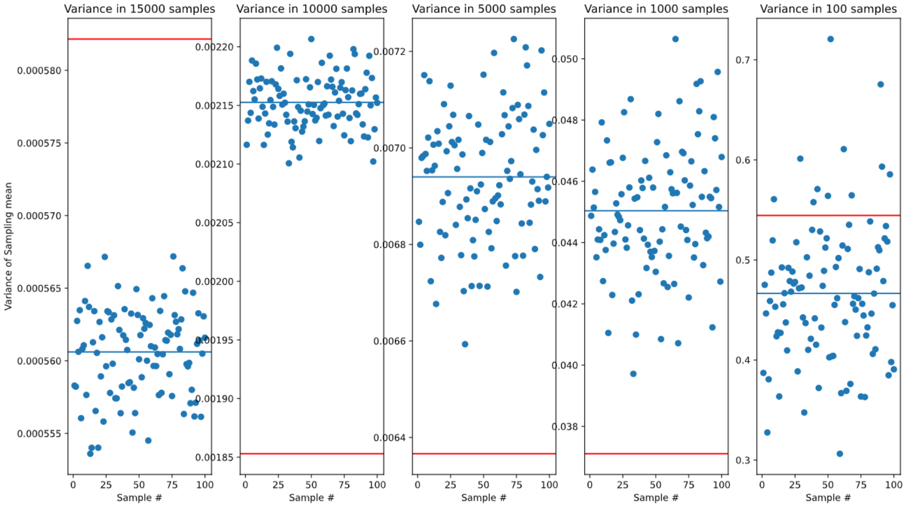
Sample Size	Variance
15000	0.00062
10000	0.0022
5000	0.0069
1000	0.045
100	0.4666

Task 6: Creating a new Variable Strata

We decided to try out to sub-divide our data based on variable LAM, since it is said that attacking midfielders are often in charge of penalty kicks[1]. Before performing that kind of stratification, we dropped all players that had an unknown value in this field(2085 rows). Also during the data process we realised that variables LAM, CAM and RAM are equal and hence, can all be used for this stratification. We observe that our sample has now even lower analytically-estimated variance(Figure 5) and all the variance data points are located tightly around their mean.

[1] https://the18.com/en/soccer-learning/soccer-positions-explained-names-numbers-and-roles

Figure 3: Comparison of calculated and analytical variance of sample mean for different sampling size performed with SRS for variable Overall. Red line denotes variance of mean calculated manually and blue dots denote variance of sample for each of 100 iterations calculated with a formula. Blue line is their mean and drown for convenience.



Appendix

In this project we divided our working effort as follows: Emilio wrote parts of reports. Suresh wrote some parts of reports, some parts of code and performed some analysis. Sofya wrote the code and some parts of the report.

Author details

References

Figure 4: Comparison of calculated and analytical variance of sample mean for SRS and Stratified sampling divided by category Position for Variable Overall. Red line denotes variance of mean calculated manually and blue dots denote variance of sample for each of 100 iterations calculated with a formula. Blue line is their mean and drown for convenience.

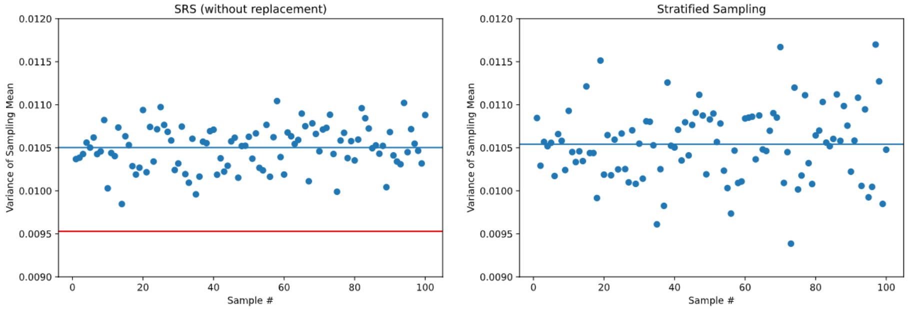


Figure 5: Comparison of calculated and analytical variance of sample mean for SRS, Stratified sampling divided by category Position and Stratified sampling by category LAM for Variable Finishing. Red line denotes variance of mean calculated manually and blue dots denote variance of sample for each of 100 iterations calculated with a formula. Blue line is their mean and drown for convenience.

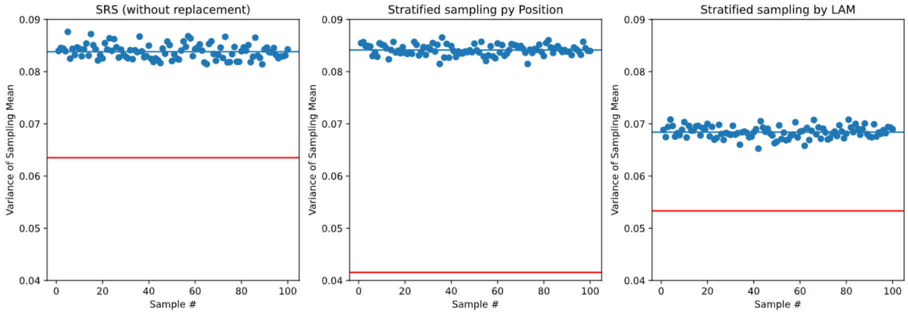
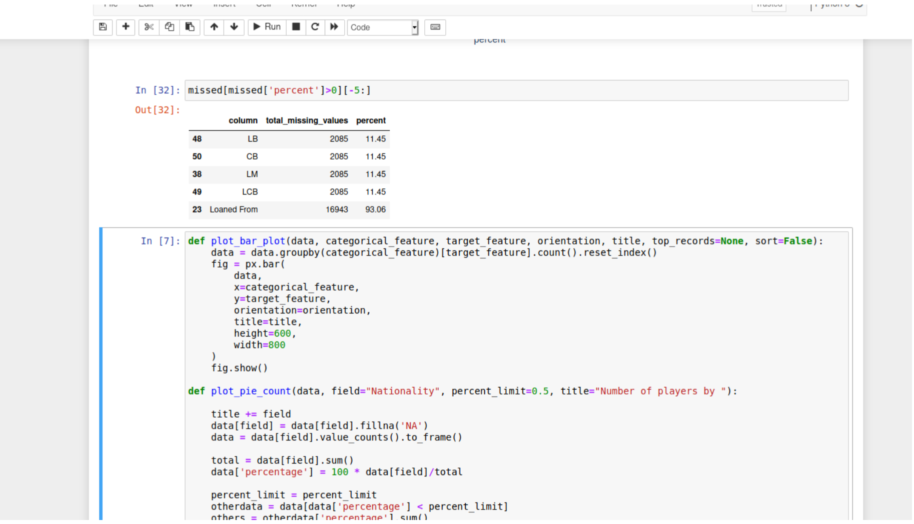


Figure 6: Top 5 columns with missing values

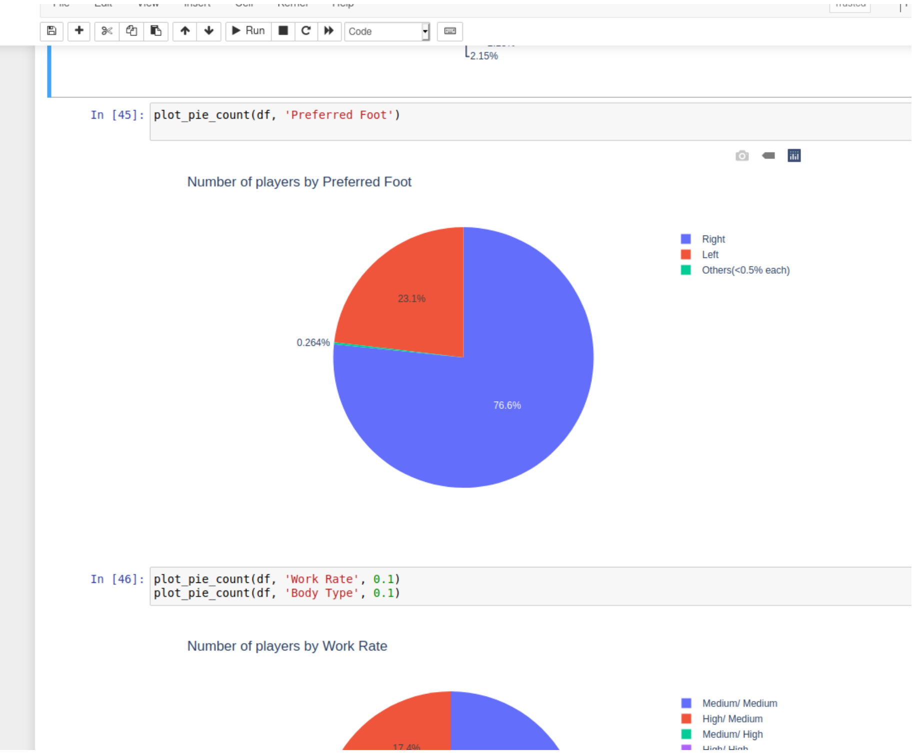


```
In [32]: missed[missed['percent']>0][-5:]
Out[32]:
```

	column	total_missing_values	percent
48	LB	2085	11.45
50	CB	2085	11.45
38	LM	2085	11.45
49	LCB	2085	11.45
23	Loaned From	16943	93.06

```
In [7]: def plot_bar_plot(data, categorical_feature, target_feature, orientation, title, top_records=None, sort=False):
            data = data.groupby(categorical_feature)[target_feature].count().reset_index()
            fig = px.bar(
                data,
                x=categorical_feature,
                y=target_feature,
                orientation=orientation,
                title=title,
                height=600,
                width=800
            )
            fig.show()

        def plot_pie_count(data, field="Nationality", percent_limit=0.5, title="Number of players by "):

            title += field
            data[field] = data[field].fillna('NA')
            data = data[field].value_counts().to_frame()

            total = data[field].sum()
            data['percentage'] = 100 * data[field]/total

            percent_limit = percent_limit
            otherdata = data[data['percentage'] < percent_limit]
            others = otherdata['percentage'].sum()
```

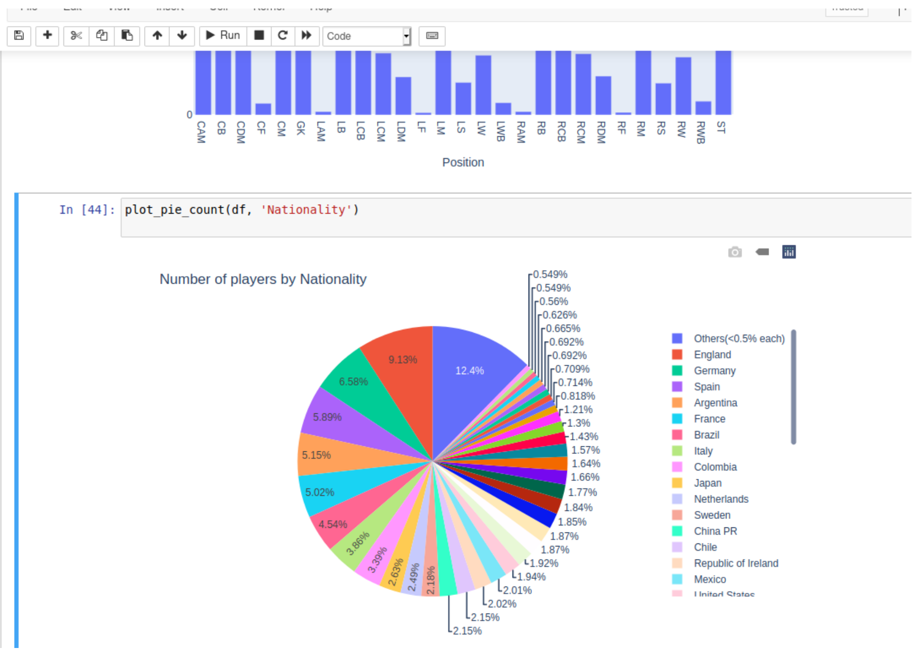
Figure 7: number of players by preffered foot



```
In [45]: plot_pie_count(df, 'Preferred Foot')
```

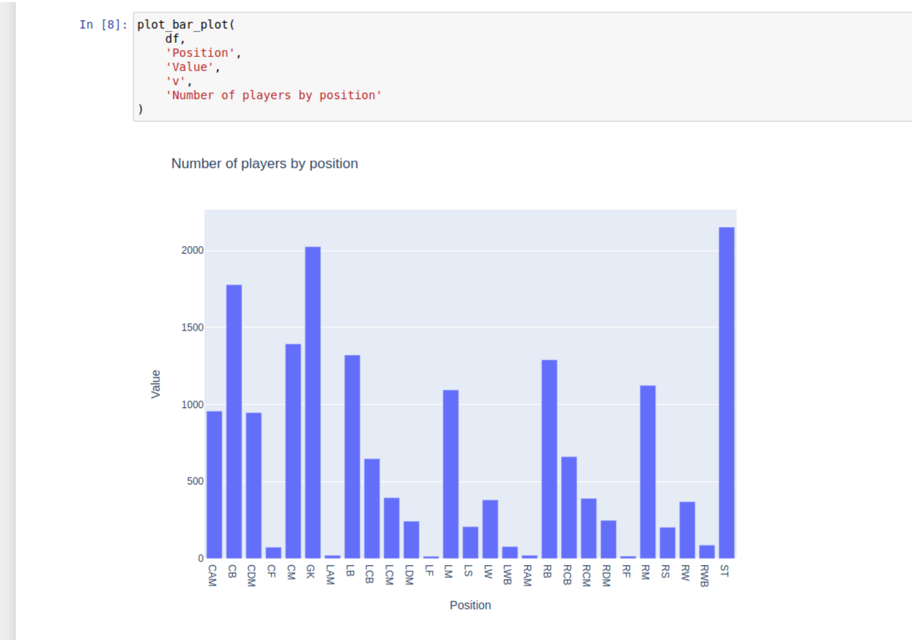
```
In [46]: plot_pie_count(df, 'Work Rate', 0.1)
         plot_pie_count(df, 'Body Type', 0.1)
```

Figure 8: number of players by countries



```
In [44]: plot_pie_count(df, 'Nationality')
```

Figure 9: number of players by position



```
In [8]: plot_bar_plot(
            df,
            'Position',
            'Value',
            'v',
            'Number of players by position'
        )
```