Dissecting Club Profits : Pacha and Green Valley

Haoran Duan[1]

School of Computing, Newcastle University, UK

Abstract. In this project, I try to present a simple and complete process of analyzing by using statistic foundation. Combine the idea from real life with the idea in statistic. And analyze the nightly profit in Pacha Club and Green Valley Club to give the manager some advice. The process contains the basic numerical and graphical knowledge, also the basic statistic methods, such as Inference and Hypothesis Test.

Keywords: Data Analysis · Statistic Foundation · Profits.

1 Introduction

With the development of technology, data collection is more accessible than before, and the method of analyzing, classifying and predicting become more and more sophisticated and precise. But the majority of them are always using computers to do the statistic on datasets. Pacha is a nightclub franchise with its headquarters in the Balearic island of Ibiza, Spain. The first Pacha club was opened in Sitges outside Barcelona in 1967. The Ibiza club, located in Ibiza Town, is the best known venue today, although the franchise also has clubs in Madrid, London, New York, Rio de Janeiro and Sydney. In the summer, the Ibiza club regularly plays host to top house DJs, including Steve Lawler, Pete Tong and David Guetta. Although (in terms of capacity) Pacha is by no means the largest clubs in Ibiza, it is open all year round, helping to make it one of the most profitable clubs on the island. This is, of course, aided by its global reputation as one of the worlds best night clubs: for years now, DJ Mag magazine has rated Pacha Ibiza in its top ten clubs in the world. Managing a global night club franchise is big business. At peak times, party-goers can pay in excess of 75 Euro for a basic entry ticket to one of Pacha Ibizas main events, although those wanting a more VIP treatment can expect to pay up to ten times that amount. Entrance to the Green Valley club in Camboriu can range from 150 to 750 Brazilian Real. Although the clubs pay huge fees to the top performers and high salaries to their resident DJs, extremely healthy profits are often made.

2 Related Work

2.1 Statistic

The data in our life are infinite. Data are the collection of numbers or symbols. We can't see anything with just looking at the data. So we will look at the

average, the tendency, the classification and so on. Statistics is used to find the properties of different data with useful information from the data, and it can just use the part of data(sample) from large data (population), to see its properties, to infer, etc.

2.2 Club Dataset

Data contain profits sampled independently at random from Pacha and Green Valley from 2017. These have been converted to dollars for ease of comparison.

2.3 Data Analyzing Tool : R

With the development of data science,R and python become the two most popular tools. For statistic area, R is the most convenient programming language, and it can be quickly mastered without worrying about the different data format or many requirements of programming skills. So in this project, R was used as analyzing tool.

3 Method

3.1 Observe the data simply before analyzing

Start with observing, cleaning and preprocessing of data are very important to understand the data and analyze it deeply later. Since they are useful not only for the professional researcher to analyze and predict, but also for other people to understand. In statistic, it is easy to use some numerical and graphical summerise to observe and preprocess our data. In this project, I start with using line diagram to be familiar with the trend of the data. Then I find the mean and the variance, the maximum and the minimum. Also, get the minimum, lower quartile, median, upper quartile and maximum (MQMQM) from Boxplots for each sample. Finally, I use the standard deviation(SD) and the coefficient of variation(CV) to show the dispersion of the two samples.

Coefficient of variation For the choice of CV and SD, the mean between the samples are different, so it is better to use CV. The CV does not change if the data are rescaled. And in my opinion, the reason that the CoV is better is that it is more likely normalized.

3.2 Distribution

After understanding the data, we can plot and observe the distribution of them. In the wild, the data collected from real life and the things happened in our life are always compatible with some distribution in the statistic. If the distribution of the data or the things can be confirmed, it will be easy for people to get the probability of what will happen in the future, and make appropriate forecasts

and decisions. As it is continuous of the club data, for example, it may follow the expectational or normal distribution. My method is to plot the histogram for both samples, and then use the Q-Q plot to observe. Also, after the distributions were confirmed, the probability density function(PDF) and cumulative distribution function(CDF) can be easily calculated. So, the probability of events about them can be generally and easily calculated.

3.3 Forecast ? Inference !

After getting the distribution, we can use the PDF or the CDF to calculate the probability whatever will happen. I will use the method of inference to evaluate a range of plausible vaule of each club, and how confidence these will be. In the area of data analyzing, for instance, if the data has the label and predictor or classifier in machine learning, it will be trained many times to get the best predictor or classifier. Then it can be used in real life without overfitting or underfitting. As for statistic, I think the aim is to find the best estimator which is to find the excellent parameters in our estimator, and the estimator comes from observing and choosing from the distribution of data. So it can be tested by using many batches of samples to test, and use the standard like Maximum Likelihood Method(MLE). The method means that the most significant value or probability caused by the parameter will be regarded as the best parameter, because people can not find the precise best true parameter without any bias.

3.4 Deviation Gives Confidence

After making some inference, we can consider if we have the confidence to announce this inference, so in the statistic, I think giving some deviation to have a range of inference make us confidence. Confidence interval will be very useful. Because the 'range inference' must better than the 'Single Value', the weird things can't be avoided at all. And in this project, the sample size is 100 which bigger than 30, so I will try to find the different confidence interval for each club.

3.5 Compare samples

The comparison is a very useful way to get information. Different population, different samples have different rules. And using comparison will give each object chance to be better. In life, we always compare many things and choose the better one. But if we don't have any idea of how to compare, we don't know enough details. So making some hypothesis is always a good way to start. If we can find which hypothesis is correct, so the content of this hypothesis is what we need, and it is the big probability things. Take the two samples as the example. If we want to compare the two sample, I think the majority is to test if it is a suitable method. So after getting every information in samples, the simplest way is to find a probability which shows how many times the profits in one club higher than another club. But in real life, there will be more and more uncertainty

and the samples just the samples. For example, the profits can be influenced by festive, but we choose the sample randomly, we don't know if the profits of each night are come from days in festive or not. If we want to use them to estimate, to predict, to analyze, it is necessary to evaluate the probability. And the key point is to control the uncertainty variables, only make one difference which is used to compare.

3.6 Solve Practical Question

The statistic is everywhere in human's life. As a data scientists, machine learning programmer or a statistic scientist, the target of learning statistic is to make it as a tool, and the data is raw material. After drilling the all useful and necessary information and the distribution or the probability about the data, it is time to give the company what the employers want to know and give suggestions. In this project, I try to analyze if employer can expect the profit of two clubs can get a specific level.

4 Experiments

4.1 Understand Data

For the first attempt, it is not apparent to get any regular information from line chart in this project.

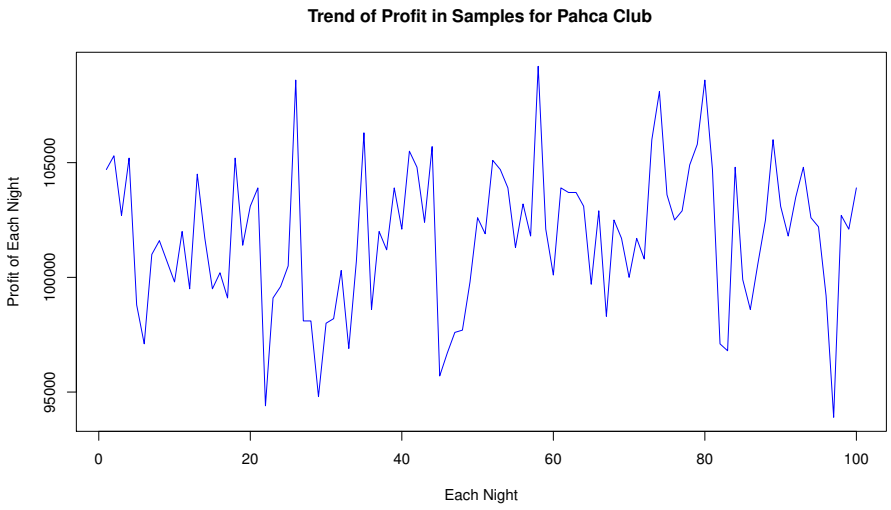

Fig. 1. Trend of Profit in Samples for Pacha Club.

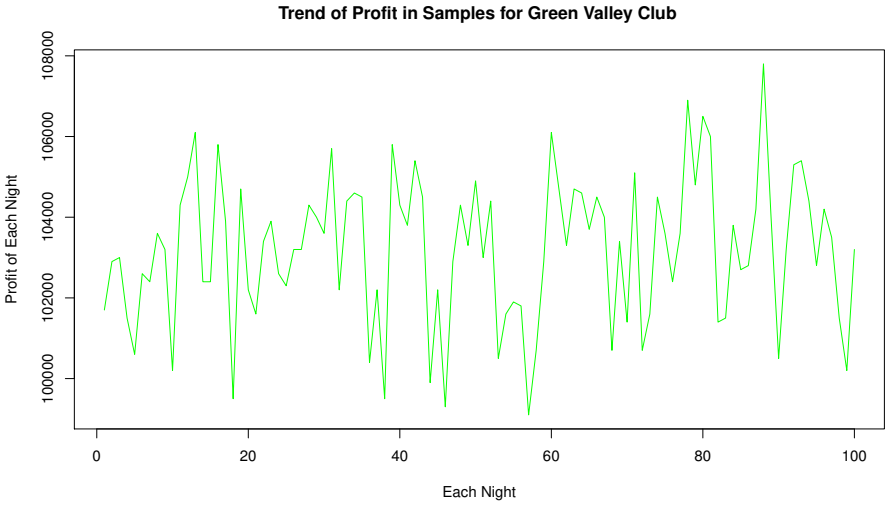

Fig. 2. Trend of Profit in Samples for Green Valley Club.

Then Plot the Box Graph and get the information

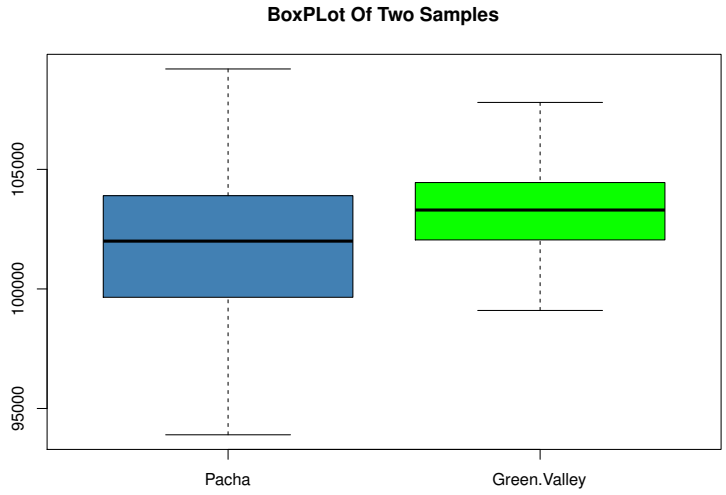

Fig. 3. Use the Box Plot First

Basic Information As the Table 1 appears, the most significant differences between the Pacha club and Green Valley club are the variance, the standard deviation and the coefficient of variance. And it can simply make some inference here, the profits in Pacha clubs each night are more unstable.

Table 1. The Numerical Summaries of Profits

Heading level	Values of Pacha	Green Valley
Mean	101754	103193
Variance	9878065	3270759
Minimum	93900	99100
Lower Quartile	99675	102125
Median	102000	103300
Upper Quartile	103900	104425
Maximum	109200	107800
SD	3142.939	1808.524
COV	0.03088762	0.01752564

4.2 Distribution of the Two samples

Pacha Club Plot the histogram with the density line and the normal distribution curve, and the Q-Q Plot

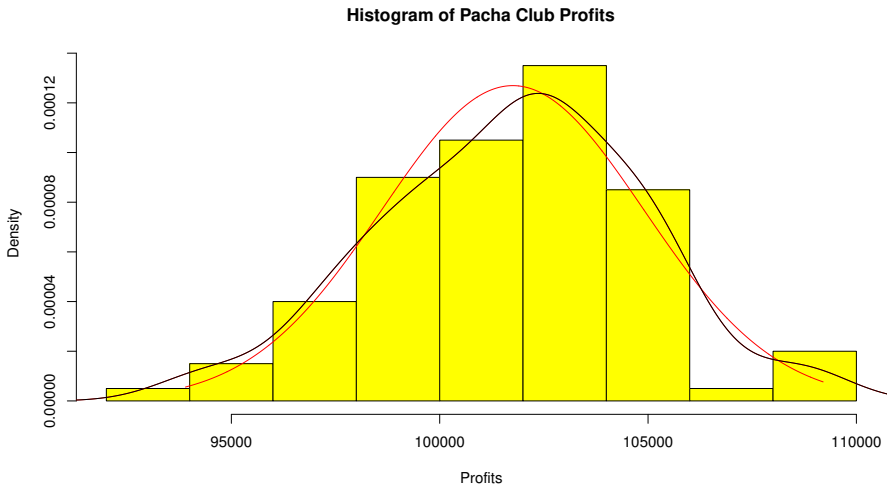

Fig. 4. Histogram of Pacha Club Profits.

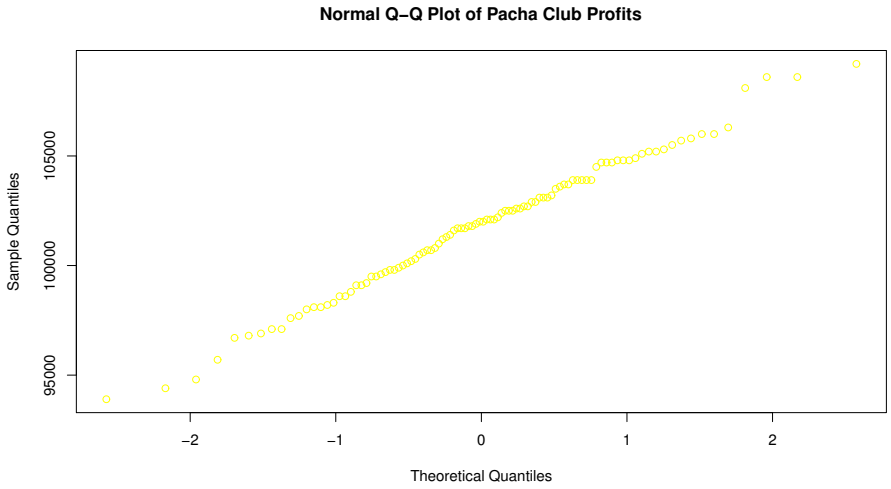

Fig. 5. Normal Q-Q Plot of Pacha Club Profits.

Green Valley Club Plot the histogram with the density line and the normal distribution curve, and the Q-Q Plot

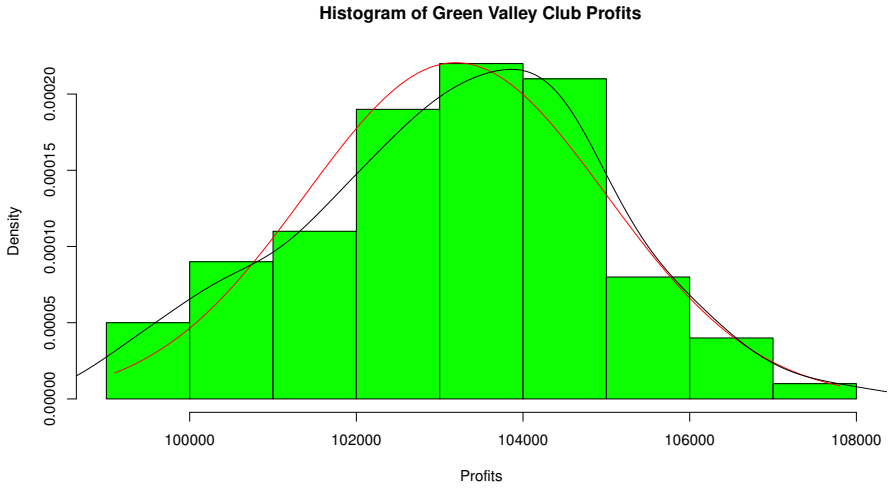

Fig. 6. Histogram of Green Valley Club Profits.

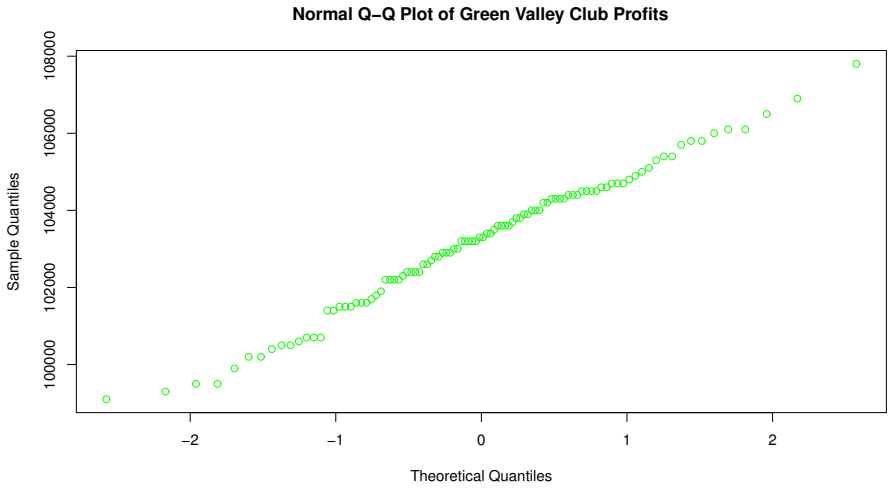

Fig. 7. Normal Q-Q Plot of Green Valley Club Profits.

So,from Fig3 to Fig6, it can be confirmed that the two samples are compatible with Normal Distribution and we can get the same formula for each like :

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Use Distribution After getting the distribution of the samples, the probability of any event about the profits can be calculated easily.

4.3 Inference

Although I got the distribution of the data,the two samples, the variance or the expectation of populations is a problem for the two samples. As the samples have the size more than 30, the variances and expectations can be calculated by point estimates. Another preciser way is to use the **T-Distribution** to get the confidence interval in this situation, which is a bit like Normal Distribution graphically:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/100}} \sim t_{n-1} \tag{2}$$

Table 2. The Confidence Interval of Profits

Each Club	95% Confidence	98% Confidence
Pacha Club	$101130 < \mu < 102377.6$	$101099.9 < \mu < 102408.1$
Green Valley	$102834 < \mu < 103551.9$	$102816.6 < \mu < 103569.4$

Comparing the confidence interval of two samples in Table 2, It shows that both 95 % and 98 % confidence interval(Call it CI just here), the profits of Green Valley Clubs' are always at the right of the Pacha Clubs', so the Green Valley is always bigger than Pacha, such as the Mean and even the Expectation of population. Also, if we check the comparison using the 98 % confidence interval specifically, any events with the probability of 2% , I think these events can be set as a 'Will Not Happen in a Single Experiment' events. So that, we can truly think the Green Valley's profits are higher than Pacha's profits. Also, the average nightly profit of Green Valley Clubs is higher than Pacha Club's.

4.4 Comparison in Profits

Binomial distribution Make an assumption, 'if the Green Valley Club has the higher profit than Pacha Club', this events follow the Binomial distribution because of the answer is YES or NO, and then get the probability from Binomial distribution. But in this case, the night is chosen randomly and the Mean and Variance of population are unknown, so it can not work.

Use Hypothesis Testing A night is chosen at random from each sample, so the format in our data is not match. One row for two clubs is not the same night probably. They should be independent. And as the target to get the probability that the night from Green Valley Club produced a higher profit than Pacha without variance and mean we can use the hypothesis testing to help us.

$$H_0 = GV \leq PA \tag{3}$$

Make a Hypothesis and we can try the Galley Valley Club does not have the higher profit than Pacha. So we will know the probability of the night from Green Valley Club produced a higher profit than Pacha as:

$$P_{(GV>PA)} = 1 - P_{(GV \leq PA)} \tag{4}$$

p-Value According to my own understanding, when the Null Hypothesis is True, the *p-Value* is the probability of the occurrence of the result of the observation from a sample we make [1] [2]. So in this case, we just easily think :

$$p\text{ - }Value = P_{(GV>PA)} = 1 - P_{(GV \leq PA)} \tag{5}$$

Also ,the size of each club samples is 100 which can be 'Large Enough', and the Distribution T-Test is a bit same as Normal Distribution.

Use Two Sample T-Test we get the ***p-Value = 0.0001095*** So, from the deducing above. the probability that the night from Green Valley Club produced a higher profit than Pacha is:

$$P_{(GV>PA)} = 1 - P_{(GV \leq PA)} = 1 - p\text{-}Value \tag{6}$$

So we can get the probability that the night from Green Valley Club produced a higher profit than Pacha is ***0.9998905***.

4.5 Practical Solution

1. Make a nightly profit of $ 100,000 First we should confirm the **Null Hypothesis** and **Alternative Hypothesis**

$$\begin{aligned} H_0 &: Profit = 100,000 \\ H_1 &: Profit > 100,000 \end{aligned} \tag{7}$$

Then find the ***Mean*** and ***Variance***, but for some situation, we do not know the population mean and variance like this practical demand. So we can use unknown variance Binominal-Test.And for this situation, the two samples are divided, so we choose to use the one-sample t-test.Next, try to calculate the ***p-Value*** by the method we used above :

$$\begin{aligned} Pacha\_Club &: p - value < 2.2e-16 \\ Green\_Valley\_Club &: p - Value = 1.046e-7 \end{aligned} \tag{8}$$

Both club has the ***p-Value*** less than 0.001, then we can check the Table 3

Very Strong Evidence Against H0 : Reject it and go with H1 means that **we have strong evidence to suggest two clubs absolutely perform better than expected about nightly profit is $100,000.**

Table 3. Rule-of-Thumb For The Interpretation of P-Value

P Value	Interpretation
$p \geq 0.1$	No evidence against H0: do not reject H0
$0.05 \leq p \leq 0.1$	Slight evidence against H0, but not enough to reject it.
$0.01 \leq p \leq 0.05$	Moderate evidence against H0: reject it and go with H1.
$0.001 \leq p \leq 0.01$	Strong evidence against H0: reject it and go with H1.
$p < 0.01$	Very strong evidence against H0: reject it and go with H1.

2. Make profit only exceed $105,000 on 5% of night Confirm the hypothesis,and according to the data, and the practical demand, it can be set as

a Binominal Distribution. Because the hypothesis is 'is' or 'Better'. Next we should try to calculate the ***p-Value*** by using

$$\begin{aligned} H_0 : P_{(profit=105{,}000)} &= 5\% \\ H_1 : P_{(profit=105{,}000)} &> 5\% \end{aligned} \tag{9}$$

$$\begin{aligned} Pacha\_Club : p - Value &= 0.9999 \\ Green\_Valley\_Club : p - Value &= 0.9941 \end{aligned} \tag{10}$$

So we can check the Table 3 above

No Evidence against H0: Do Not Reject H0 means that **For both club, we do not have evidence to suggest the clubs on average perform better than before.So the number of night which has profit $105000 are always less than 5%. And simply check our data in Excel, there is no more than 1 night get the profit as $105,000.So the method works.**

5 Discussing and Future work

5.1 Pacha and Green Vallery

After the whole process of analyzing, we can know Pacha Club develop more unstable, and always got the lower profit to Green Valley Club. We can predict the nightly profits of both of them are about $100,000 to $105,000. So, the manager can make some chaneges from these analyzing.

5.2 Analyze With Statistic

Analyzing with the statistic is one of the most accurate and powerful ways. In this project, I try to use the example data to simulate what I think the process of analyzing with the statistic is. It is absolutely useful to use some numerical or graphical summaries to understand the data, to extract the valuable information. Before analyzing, we can get the distribution or the pattern of the data we have, and it leads to the right way for later analyzing. And during the analyzing, the probability is not the true value, so using the probability to describe the data and to predict, make it better for scientists to handle the deviation. But if we want to try the statistic in small samples, datasets or events, to keep the distribution or the information correct is the most important. So that, choosing data or sample will be the more difficult things in the statistic. If we are wrong at the beginning, nothing will be successful. I find it will be faster to understand something like machine learning or data analyzing from statistic than starting from the single area concept.

References

1. Ruxton, G.D., 2006. The unequal variance t-test is an underused alternative to Student's t-test and the MannWhitney U test. Behavioral Ecology, 17(4), pp.688-690.
2. Derrick, B., Toher, D. and White, P., 2016. Why Welchs test is Type I error robust. The Quantitative Methods in Psychology, 12(1), pp.30-38.