Stats Project report

Dissecting night club profits: Pacha versus Green Valley

By : Callum Simpson

SN: b6030326

Intro

In this report, I will be comparing the profits for two-night clubs. The two nightclubs I will be looking at are Pacha Ibiza which is in Ibiza and Green Valley in Cambori´u. Whilst Green valley is larger (Capacity 6,000) than Pacha (Capacity 4,000) both draw in the similar party crowd.

This report will be using the profits converted to US$ (000s) from 30 randomly sampled nights from Pacha and Green Valley during 2018.

Numerical analysis

The first step that I decided to investigate was the profits of each night.

Doing a quick look at some numerical summaries we find the following.

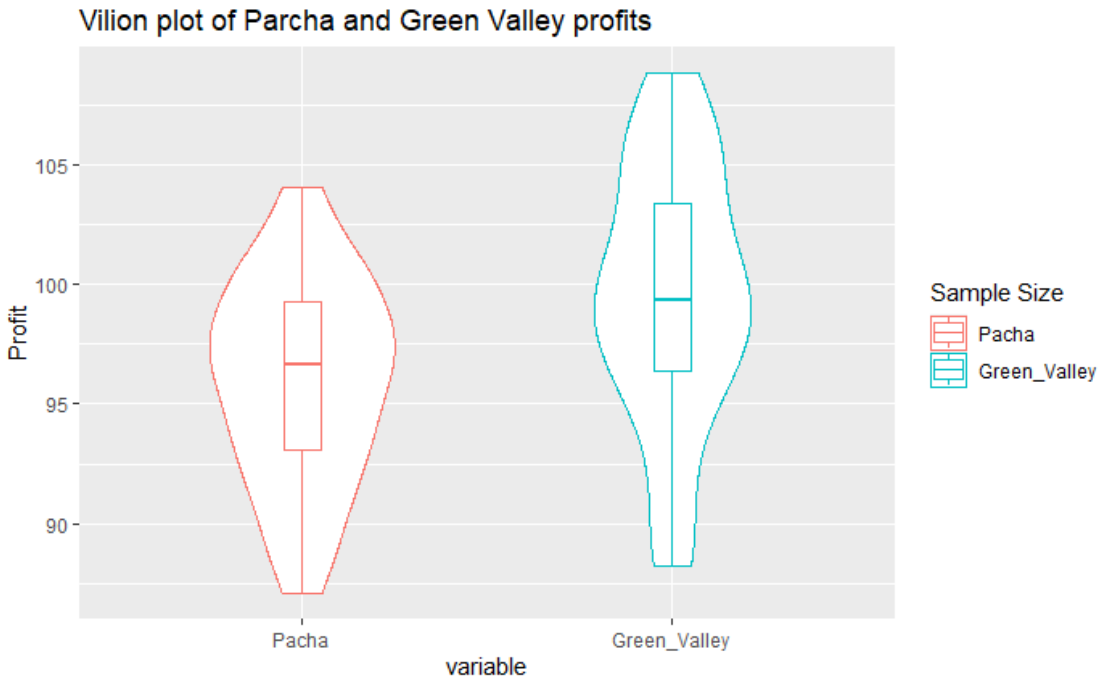
key <fctr>	mean <dbl>	sd <dbl>	max <dbl>	min <dbl>	range <dbl>	IQR <dbl>	length <int>
Pacha	95.95565	4.232797	104.0733	87.11056	16.96271	6.16722	30
Green_Valley	99.43406	5.454555	108.7992	88.21276	20.58649	6.97948	30

From this, we can determine the following.

- Both share similar max and mins stating that my sample probably didn't contain any outliers that could of skew results.
- In the sample given the Green Valley had the greater mean of 99.4 compared to Pacha mean of 96.00, However Green Valley had a greater standard deviation of 5.45 compared to Pacha 4.23. This suggests that the mean gotten for Pacha is a little bit more reliable than Greens as there seems to be more of a spread of result in Green. This can somewhat be backed up by the fact that Green Valley had a higher IQR range which suggests that it has a greater variety of results.
- What should be said is that Pacha means + 1SD is just a little bit over Green Valley's mean which suggests that given any night Green valley has a higher chance of generating the most profit. I.e roughly 68.26% of Pacha sample results fall under Greens sample mean.
-

Graphical analysis

To do a further comparison I decided that I wanted to graph out the sample's profits using a violin plot + boxplot so that I could get an idea of the density.



From this graph, we know that all values all within the 95% confidence intervals.

We also see that from the given sample...

There were a few times where Green Valley was able to produce a profit that was greater than the maximum profit that Pacha was able to achieve. We also see that Pacha has a high density of values that are above its mean and that Green Valley highest density area is below its mean. As these densities almost overlap it suggests that there is a somewhat “typical night club profit” and that Pacha falls under this threshold and Green Valley typically makes more.

Using the Boxplots, we see that Green Valley lower quartile is around Pacha mean. Again, showing that if we were to compare nightly profit it's expected that Green valley would make more profit.

We can say that given the distribution, our samples don't look normally distributed.

Overall, from just this peak at the data, I believe that for most nights it's likely that the profit made at Pacha is less than the profit made at Green Valley.

Confidence Interval

As I was given was a sample, I wanted to find ranges of plausible values for the average nightly profit at each club. As we don't know anything about the population, I calculated a confidence interval for the mean of a normal distribution using the samples standard deviation. Though the data normally must be normal distributed for this to work we can do this due to our sample size and the Central Limit Theorem we can do this (sample is greater or equal to 30)

I did a calculation for a 90% confidence interval and a 95% confidence interval.

90%

key <fctr>	confidence <dbl>	lower <dbl>	upper <dbl>
Pacha	90	94.64257	97.26874
Green_Valley	90	97.74197	101.12616

95%

key <fctr>	confidence <dbl>	lower <dbl>	upper <dbl>
Pacha	95	94.3751	97.5362
Green_Valley	95	97.3973	101.4708

The clear message that can be taken from this is that when looking at the 95% confidence interval the upper plausible value for Pacha profit value is just a little bit over Green Valleys lower plausible value. When two confidence intervals overlap, the difference between the two parameters can be significant or non-significant. Given the ranges the population mean for Green Valley is probably greater than Pacha population mean however as there is a slight overlap at a 95% confidence interval, we cannot say this with 100% certainty.

What interesting is that doing a test 90% confidence there is no overlap in ranges meaning going forward we will be careful to use the 95% confidence interval so that useful information like this won't be missed.

A Two-sample T-test got (a p-value of 0.022) Suggesting that the means of the two population are not equal.

Probability as a mean

Next, I wanted to see if what is the probability the night from Green Valley produced a higher profit than Pacha if a night was chosen independently at random from each sample. To solve this, I looped through each possible combination of nights and worked how many of these combinations had the Green Valley producing the most profit. Doing this I got the probability that a night from Green Valley produced more profit than Pacha was 69.22%. This is significant as it suggests that roughly 7/10 times Green Valley produces more profit each night.

A positive of this is that it's a quick method of getting a percentage that allows us to compare the two sample together, say if the probability was around 50% then we would know that both samples share a similar mean. You can also say that calculating the probability this way is reliable as we are using all possible combinations of nights meaning if we run the calculation again with the same samples, we would get the same percentage back.

Though a negative about this is that we are using a predicted probability from a sample to try and make a prediction about the probability. The sample we got isn't normally distributed meaning that it's possible that if we were to increase the amount of sample data to make the graph more normally distributed, we may a different percentage showing that our initial prediction is incorrect.

Hypothesis test

Financial analysts expect each club to make an average nightly profit of $100,000. I wanted to check If this is true.

To do this is created a hypothesis test to compare the mean gotten from the sample to this predicted mean to see how likely it was.

The Null hypothesis (H0) was $\mu = 100$ and the Alternative was (H1) $\mu > 100$. As my hypothesis is checking the probability that the mean is equal to 100 or greater than I used a greater than one tailed test. Using the sample means and standard deviation I was able to calculate a t value which I then went on to work out a p value.

$$t = (mean - 100) / (sqrt((sd*sd) / 30))$$

$$p = 1 - pt(t, 29))$$

These are the results,

key <fctr>	mean <dbl>	sd <dbl>	t <dbl>	p <dbl>
Pacha	95.95565	4.232797	-5.233373	0.9999934
Green_Valley	99.43406	5.454555	-0.568290	0.7128937

Using the following p interpretations.

p-value	Interpretation
$p \geq 0.1$	No evidence against H_0: do not reject H_0.
$0.05 \leq p < 0.1$	Slight evidence against H_0, but not enough to reject it.
$0.01 \leq p < 0.05$	Moderate evidence against H_0: reject it and go with H_1.
$0.001 \leq p < 0.01$	Strong evidence against H_0: reject it and go with H_1.
$p < 0.001$	Very strong evidence against H_0: reject it and go with H_1.

As we can see we achieved a high probability for both Pacha and Green Valley which both fall under the No evidence against H0 so do not reject H0. This means we don't go with the alternative that the mean exceeds $100,000.

This makes sense for Pacha as the 95% confidence interval predicted range was less than 100 but the range for Green valley range went slightly above and below $100,000 which is probably why we got a lower probability but not enough to reject the null hypothesis.

What could be said is that the t-value produced was more reliable for Pacha than it was for Green, Low t-values are indications of low reliability of the predictive power of that coefficient. Whereas +2 or less than – 2 is more reliable. Meaning the p value for Green Valley could be slightly off.

Next, I wanted to test the analysis other claim that the mean exceeds $105,000 on 10% of nights and the probability that my samples exceed that. I performed a hypothesis test with the Null hypothesis being H0 : p = 0.1 (probability is 10%) and the alternative hypothesis being H1: p >0.1 (the probability it exceeds 10%). As this is a population probability problem, I worked out the number of nights that above $105,000 number for both samples. Using a p binomial distribution to find the probability of getting said number of nights under the analysis claim I got the following.

Caluclation

$$1-\Pr(X \leq greaterThan - 1 \mid \theta=0.1)$$

R code

1 - pbinom(greaterThan - 1 , length , 0.1)

These are my results.

key <fctr>	length <int>	greaterThan <int>	p <dbl>
Pacha	30	0	0.95760884
Green_Valley	30	7	0.02582679

For the sample we got for Pacha we found the number of nights which made a profit greater than $105,000 was 0. With probability of this happening is 0.957 which can be interpreted as no evidence against the null hypotheses so don't reject it.

For the sample we got for Green Valley we found number of nights which made a profit greater than $105,000 was 7. If we are expecting 10% of a population to be above $105,000 the probability of this happening is 0.0258 which can be interpreted as moderate evidence against the null hypotheses so reject it a go with the alternative hypotheses.

From this we can say that the analysis prediction may be correct for Pacha as we don't reject the null hypotheses, but we know that the mean doesn't exceed $105,000 on more than 10% of nights. However, as we rejected the null hypotheses for Green valley, we know that it must make $105,000 on more than 10% of nights.

Conclusion

So, in conclusion, from the sample that I have been given it easy to say that the Green Valley club makes more money on average. Using a 95% confidence interval I found that the range of possible means for Green Valley is for the most part greater than the 95% confidence interval range for Pacha. Also, if you were to pick a random night for each night club then you have a roughly 7/10 chance that the money made by Green is greater than Pacha.

Whilst both gave a similar result for a hypotheses test if the mean was $100,000, we know green has a higher probability of making $105,000 a night.

This analysis could be improved with an increased sample size. It allows us to get a more detailed numerical summary which would allow us to have a more accurate understanding of the two nightclubs means. This is because as even though that the nights are random it's possible that the worst nights from one club is selected whilst the best of the other is chosen. The larger the sample the less likely this is. It would also allow us to identify outliers as well. This would give cleaner point to start our investigation from.

Increasing the sample size would also make the samples more normally distributed as central limit theorem states “as sample size increases the distribution of sample parameter (in this case, the mean) will start to approximate a normal distribution.”

Also, in hypothesis testing the P-values depend upon both the size and the precision of the sample size. When the sample size is large, results can reach statistical significance even when the effect is small and clinically unimportant. However, when the sample size is small it's possible that the results can fail to reach statistical significance meaning an incorrect judgment pay be made. So increasing the sample would allow us to use hypothesis with greater confidence.

The sample is only giving us the profit of each night which is alright for working out which makes more money but doesn't it us any insight into why that nightclub is making that much. We know that Green Valley is large than Pacha so that may be one of the reasons (Green can hold more people so has more people spending money) and that ticket prices vary for each night club during the year We can also sort of interoperate that “attracts a similar clubbing crowd of house and trance–loving revelers” may mean that both people form the same economic background, but we have no idea if that is true. It's possible that both clubs are getting the same number of people in but Green just has people who are willing to buy more. Having extra information on things the crowd size each night and ticket prices may allow to understand why Green was able to make more profit.