You Chen Portfolio

by You Chen

Submission date: 05-Jun-2019 09:05PM (UTC+1200)
Submission ID: 1140231073
File name: YOU_CHEN_PORTFOLIO.pdf (404.2K)
Word count: 2091
Character count: 10024

YOU CHEN PORTFOLIO

YOU CHEN

12th May 2019

Portofolio: An analysis of Business Instagram Performance

I am running my online handbag store within New Zealand. I created an Instagram account to promote my store and share the pictures on it regularly, and this project presents a statistical analysis of my business Instagram account performance.

The data is based on the time period from May 2018-May 2019.The dataset has 23 rows and 4 columns. Each row corresponds to a style. The first column shows each style's name, and the second column "Likes" presents how many likes for the image of the style, and the third column "Profile visit" shows the number of visits on Instagram Page after seeing the post. The four column "Reach" refers to how many people who have seen the post.

```
#import the data
portfolio <-read.csv("../data/portfolio.csv")
portfolio
```

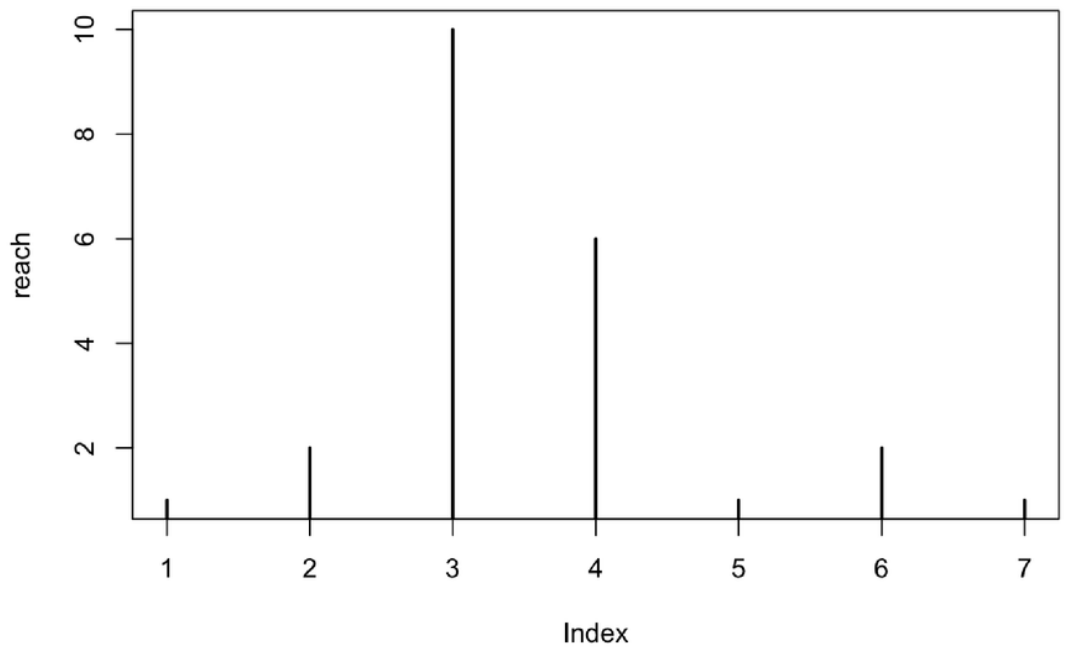
```
##                       Style.Name Likes Profile.Visit Reach
## 1                   Bamboo Basket    61             4   517
## 2                     Bamboo Tote    44             2   307
## 3                Black Bucket Bag    56             0   484
## 4                Brown Bucket Bag   102            14   492
## 5               Zipper Bucket Bag   110            17   896
## 6               Black Padlock Bag    85            10   450
## 7               Brown Padlock Bag    47             2   190
## 8         Black Texture Crossbody    49             5   372
## 9         Brown Texture Crossbody    79            22   495
## 10        Small Texture Crossbody    78             7   693
## 11                     Black Tote   109            86  1382
## 12                     Brown Tote   170            17   720
## 13        Black Vintage Crossbody    55             6   599
## 14        Cream Vintage Crossbody    92            12   640
## 15           Mocha Knot Crossbody   105            12  1003
## 16            Wine Knot Crossbody    89             8   649
## 17 Green Circle Handle Crossbody    87             9   417
## 18  Grey Circle Handle Crossbody    63             9   551
## 19                       Mini Bag    41             2   484
## 20                    Mini Clutch    80             8   640
## 21                   Pleat Clutch   134            55  1076
## 22              Crystal phone Bag    44             4   425
## 23                   Vegan Wallet   104            35   653
```

As I know the average reach rate is 529 from my business account, so I would like to know if average reach rate of the sample differ from the known mean?

Null Hypothesis: H0:μ=529. The average reach rate is 529.

Becasue I would like to know whether the sample is greater or less than the average, so I use two-sided test, and I have seperated into 7 ranges to see if the distribution is close to noraml distribution.

```
reach <- c('0-199'=1,'200-399'=2,'400-599'=10,'600-799'=6,'800-999'=1,'1000-1199'=2,
'1200-1399'=1)
plot(reach,type="h",lwd=2)
```



We can find the plot is close to normal distribution.Therefore, t.test is conducted as below:

```
t.test(portfolio$Reach,mu=529)
```

```
##
##  One Sample t-test
##
## data:  portfolio$Reach
## t = 1.538, df = 22, p-value = 0.1383
## alternative hypothesis: true mean is not equal to 529
## 95 percent confidence interval:
##  499.1873 729.9431
## sample estimates:
## mean of x
##  614.5652
```

We can see that p-value(0.1383) exceeds 5%, so we fail to reject the null hypothesis, and infer there is no evidence for the average reach rate of the sample differ from population mean 529.

“Likes” for styles

It looks like some of the styles have more “LikeS” than the others. For example, the Brown Tote has 170 likes while the Mini Bag only has 41 likes, so I wonder if it is the case or variablity explain the difference.

```
o <- portfolio$Likes  #'o' for 'observation'
o
```

```
##  [1]  61  44  56 102 110  85  47  49  79  78 109 170  55  92 105  89  87
## [18]  63  41  80 134  44 104
```

Now the null is that each item has the same probability:

```
e <- rep(mean(o),length(o))   #'e' for expectation
e
```

```
##  [1] 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304
##  [8] 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304
## [15] 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304 81.91304
## [22] 81.91304 81.91304
```

We use Pearson’s chi-square to compare the observation with expectation.

```
B <- sum((o-e)^2/e)
B
```

```
## [1] 276.828
```

```
pchisq(B,df=length(o)-1,lower.tail = FALSE)
```

```
## [1] 5.914313e-46
```

Therefore, p value is extremely small (5.914313e-46). It shows some styles are indeed have more “Likes than the others.

“Porfile Visit” for styles

Meanwhile, it looks like some of the styles have more “Profile Visit” than the others such as Black Tote which image reach 86 profile visits, while Black Bucket Bag image doesn’t bring any profile visits. So I wonder if it is the case or variability explain the difference.

```
o <- portfolio$Profile.Visit  #'o' for 'observation'
o
```

```
##  [1]  4  2  0 14 17 10  2  5 22  7 86 17  6 12 12  8  9  9  2  8 55  4 35
```

Now the null is that each item has the same probability:

```
e <- rep(mean(o),length(o))   #'e' for expectation
e
```

```
##  [1] 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348
##  [8] 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348
## [15] 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348 15.04348
## [22] 15.04348 15.04348
```

We use Pearson's chi-square to compare the observation with expectation.

```
B <- sum((o-e)^2/e)
B
```

```
## [1] 567.0867
```

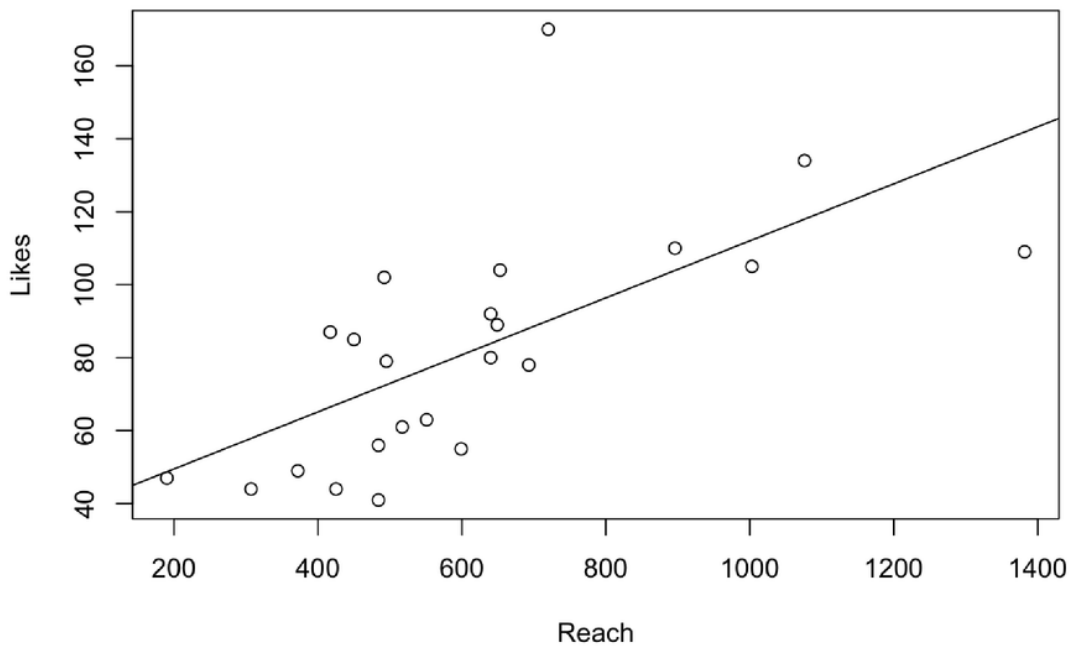
```
pchisq(B,df=length(o)-1,lower.tail = FALSE)
```

```
## [1] 6.928735e-106
```

Therefore, it shows the some styles are indeed have more profile Visits than the others.

Next I am wondering if a certain style has high number of "Reach" which explain the high number of "Likes". #if more "Reach" leads to more "Likes"?

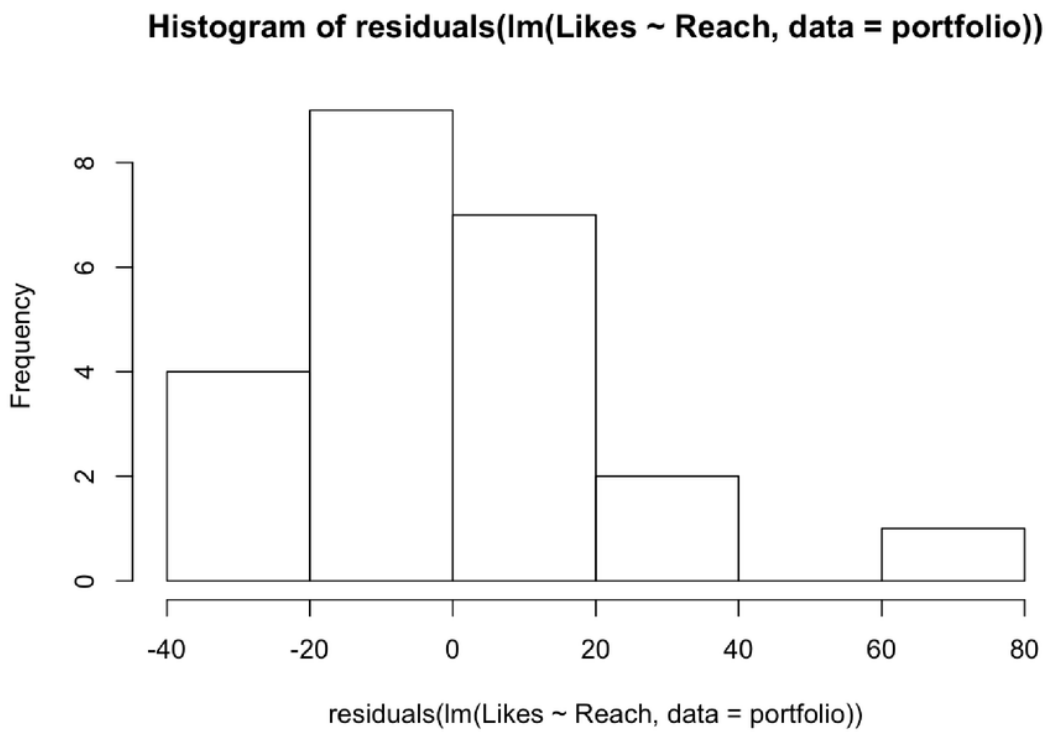
```
#So I plot the likes" against "Reach" to draw a scattergraph.
plot(Likes~Reach,data=portfolio)
abline(lm(Likes~Reach,data=portfolio))
```




```
#From the scattergrah, it seems it is on a upward trend,so more higher likes has high
er reach.Is the regression line significant??
summary(lm(Likes~Reach,data=portfolio))
```

```
##
## Call:
## lm(formula = Likes ~ Reach, data = portfolio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.886 -13.951  -3.901  12.023  79.847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.88634   13.33093   2.542 0.018973 *
## Reach        0.07815    0.01997   3.914 0.000798 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.99 on 21 degrees of freedom
## Multiple R-squared:  0.4218, Adjusted R-squared:  0.3943
## F-statistic: 15.32 on 1 and 21 DF,  p-value: 0.0007979
```

```
hist(residuals(lm(Likes~Reach,data=portfolio)))
```



The P value is 0.0007979<0.05(one side test), so there is strong evidence that higher “likes” has more “Reach”. The below histogram shows that the residuals are roughly Gaussian, in line with the assumption of linear modelling.

We might wonder whether a high number of reach rate for the image on the Instagram also means a higher likes rate? For this, we use Fisher’s exact test. “high number of likes” means the style which has more than 80 likes, and “high reach rate” means the image reaches more than 600.

Our null would be high “Likes” is idependent of high number of “Reach”.

```
p <- table(HighReach=portfolio$Reach>600,HighLikes=portfolio$Likes>80)
p
```

```
##           HighLikes
## HighReach FALSE TRUE
##     FALSE    10    3
##     TRUE      2    8
```

We can discover 13 out of 23 styles don’t have high number of reach rate, only 3 styles have high reach rate. 10 out of 23 styles have more than 80 likes, and 8 styles hit high reach.

As we want to know whether the high number reach would cause a higher or lower like, so it is two sides Fisher’s exact test to test our null.

```
fisher.test(p)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  p
## p-value = 0.01228
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    1.34348 172.78045
## sample estimates:
## odds ratio
##   11.52638
```

Thus the p value is 0.01228 < 0.05, so there is strong evidence we can reject the null. It seems the high number of likes also indicate a high rate of reach.

Next, I am wondering if the it is because a certain style has high number of reach which explain the high number of profile Visit?

Is a higher number of reaches correspond to higher profile visit?

Therefore, I try a null of number of profile visit being proportional to number of reaches.

```
e <- portfolio$Reach*sum(portfolio$Profile.Visit)/sum(portfolio$Reach)
e
```

```
##  [1] 12.655253  7.514821 11.847471 12.043297 21.932508 11.015210  4.650867
##  [8]  9.105907 12.116732 16.963424 33.828935 17.624337 14.662469 15.666077
## [15] 24.551680 15.886381 10.207428 13.487513 11.847471 15.666077 26.338592
## [22] 10.403254 15.984294
```

```
#The observations are:
o <- portfolio$Profile.Visit
o
```

```
##  [1]  4  2  0 14 17 10  2  5 22  7 86 17  6 12 12  8  9  9  2  8 55  4 35
```

```
B <- sum((o-e)^2/e)
B
```

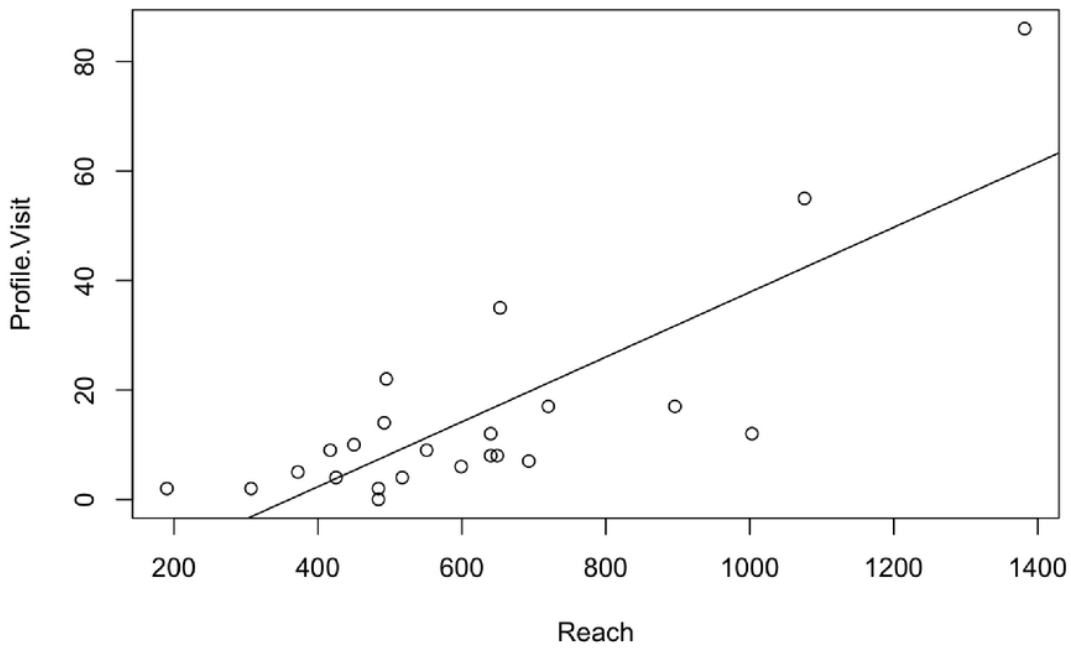
```
## [1] 208.7233
```

```
pchisq(B,df=length(e)-1,lower.tail = FALSE)
```

```
## [1] 2.214272e-32
```

Therefore, the p value is lower than 0.05 which indicated we can reject the null.

```
#So I plot the "Reach" against "profile visit" to draw a scattergraph.
plot(Profile.Visit~Reach,data=portfolio)
abline(lm(Profile.Visit~Reach,data=portfolio))
```



From the scattergrah, it seems it is on a upward trend,so more higher likes has higher profile visit.Is the regression line significant? We start the Null Hypothesis:H0:β=0. There is no relationship between profile visit and reach rate.

```
summary(lm(Profile.Visit~Reach,data=portfolio))
```

```
##
## Call:
## lm(formula = Profile.Visit ~ Reach, data = portfolio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.053  -7.715  -2.278   5.938  25.497
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.361044   6.414302  -3.330  0.00318 **
## Reach         0.059236   0.009607   6.166 4.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.02 on 21 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6272
## F-statistic: 38.02 on 1 and 21 DF,  p-value: 4.066e-06
```

The P value is 4.066e-06<0.05(one side test), so there is strong evidence that higher likes has more profile visit.

The histogram shows that the residuals are close to Gaussian, in line withe the assumption of linear modelling.

Conclusion

The dataset was analysed using various statistical methods such as t.test, chi-square, fish-test, linear regression.To conclude, there is no evidence that the average reach rate of the sample differ from population mean 529, and the results show that some styles are indeed have more profile Visits and likes than the others.

The chi-square, fish tests prove that there are relationship between “Reach” and “Profile Visit”, and also the “Reach” and “likes” these two linear regression models.Compared the R square, the model for “reach~profile Visit” is 0.6272, while the model for “Reach~Like” is 0.3943, which means the model for “reach~profile visit” is more fit than the model “Reach~Like”, and can explain better that high reach can brings to high profile visit.

You Chen Portfolio

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9