VIA_Stats

Yijun Wan (yw3038)

6/21/2019

1. Assume Via has a demand prediction algorithm for predicting hourly demand, 24 hours in advance. Answer the following questions theoretically:

a) How could you compare the performance of this algorithm across cities of varying size? Propose a way to predict performance of the demand algorithm in a city we're considering operating in.

1) I will calculate the absolute value of difference between predict demand and true demand, divided by operation size in that city. Basically, I am considering that there is some variation between the true demand and the predict demand, but the whether the variation is acceptable or not should be decided by the operation size. For example, a variation of 5 thousand may indicate good performance in large cities, while bad performance in small cities.
2) If we can find a similar city for that our target (in aspects like city size, population, city location, residents' habit etc..) where we have a operation, then we can use algorithms performance in that city to estimate algorithm performance in our target city. If we find no similar cities, we need to find factors which have impact on algorithm performance (such as operation size, population, private vehicle ownership, public transportation etc). Train predictive models by making those factors into features and the performance of algorithms in different cities be the dependent variable. Then we can predict the performance by the model.

b) Adatascientistbuildsanewdemandpredictionalgorithm,butitismoreresource intensive than the current one - we'll only switch if we're sure the new one is better. How would you determine if the new algorithm is worth it?

Assume we use profit to measure the performance of the algorithms and we set up a bar of changing algorithms. I will compare two numbers to see if the algorithm is worth it. First, I will calculate how much more profit gained by the new algorithm. Second, I will calculate the extra effort to deploy the new algorithms, including using more expensive software &hardware, and cost of change and potential impact on our current business.

2. A report claims that between 22.4% and 43.9% of Via rides have excellent music. You can assume that "excellent" is a binary decision at the ride level. What do you think the sample size was?

Let p be the rate of the 'excellent music'. Since a confidence interval of estimated rate is given, the sample size depends on confidence level $1 - \alpha$ Assume in an dataset of size n, x is observed to be 'excellent music'. The distribution of x is a binomial distribution with parameters n and p.
x ~ B(n, p) What we want is the ditribution of good music rate, estimated by

$$\hat{P} = \frac{x}{n}$$

and

$$\hat{P} \pm Z_{\alpha/2}(\frac{\hat{P}(1-\hat{P})}{n})^{\frac{1}{2}}$$

. We already know $\hat{P}$, if the α is set, then n can be calculated.

```r
find_sample_size = function(alpha){
 p_hat = (0.224+0.439)/2
 z_score = qnorm(1-alpha/2, mean = 0, sd = 1, log.p = FALSE)
 x = ((p_hat-0.224)/z_score)**2
 y = p_hat*(1-p_hat)/x
 return(y)
}
```

```r
for(alpha in c(0.10, 0.05, 0.03, 0.01))
  cat(c('The sample size for confidence level ', 1-alpha, 'is at least', find_sample_size(alpha),'.\n'))
```

```
## The sample size for confidence level  0.9 is at least 51.8826952851757 .
## The sample size for confidence level  0.95 is at least 73.6655096568245 .
## The sample size for confidence level  0.97 is at least 90.3074664293911 .
## The sample size for confidence level  0.99 is at least 127.233705331311 .
```

As we see, as the confidence level going up, the responding sample size is getting large.