Experience 1

2021-12-25 22:56:39 [Phone + VO]

https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=833662&ctid=232566

1. 如何假设检验一个数字是否是从normal distribution里sample出来的

-- z-test. Suppose this number is k, the mean and variance of the normal distribution are mu and sigma^2, then the z-score is (k-mu) / sigma and lookup its value from a z-score table.

2. 如果normal distribution 只要大于一的数值，sample出来的数的分布是什么？-- truncated normal distribution。定义是什么？如何数值上sample？-- sample normal distribution and only keep numbers with values greater than one.

如果想直接sample呢？-- use cumulative density function of truncated normal distribution. Generate a number between 0 and 1 and map the number to corresponding x value of the truncated normal distribution. See the Python script attached. 有没有其他方法？-- monte carlo.

3. 如果做linear regression的时候有500个变量600个数据点会怎样？ -- overfitting. 为什么会over fit？-- the number of variables and the number of data points is comparable, therefore the parameters would fit to the errors in the datapoint. Why? --([ISLR](https://hastie.su.domains/ISLR2/ISLRv2_website.pdf) Page 234) Prediction Accuracy: However, if n is not much larger than p, then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training. 有什么解决办法？

-- L1 regularization. 解释一下什么是L1 regularization.

-- Dimension reduction

-- Subset selection

4. 想分析是否上过大学和收入有没有关系。于是采集了Mountain View 1000个人的data，建立一个linear regression model。这样做可以吗？ -- 问变量是binary的？是。-- 有问题。首先Mountain View 无法代表population。其次因为只有一个binary变量，linear regression 不合适。怎么改进？-- 用 t test. 还有什么改进方法？-- 增加其他变量。

5. Coding。Sample 100x100 binomial distribution. Normalize this matrix such that the sum of each column is one. – See the Python script attached.

Experience 2

2022-4-20 19:54:51 [Phone]

<https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=887161&ctid=234558>

1. mean and median, ste(mean), ste(median)

Standard deviation: .

Standard Error of mean: .

Standard error of median:

(1) be approximated by standard error of mean if sample is large.

(2) use an asymptotic formula .

(3) use a [bootstrap](https://faculty.washington.edu/yenchic/17Sp_403/Lec5-bootstrap.pdf) method: where is the th bootstrapped median.

2. 有两个variables：X1 and X2, fit regression; 同时，我们用X1+X2 和 X1-X2再fit一个regression。比较两个regression。

-Predicted results are identical. Case 1.

如果X1和X2 correlated，那么这两个regression有什么相同和不同， X1+X2和X1-X2还correlated吗？

-Predicted results are still identical. Case 2.

如果给这两个regression都加regularization，这两个regression有什么相同和不同

-Predicted results are not identical anymore. Case 3.

3.想判断go to college (predicting var) 和income (response var) 之间有没有因果关系。问在moutain view随机sample 1000个人做regression有什么问题吗？我说selection bias。然后他接着问，那这个linear regression的slope和真实population data fitted linear regression的slope比，有什么不同？

-See the answer in Experience 1 – Q4.

4. 简单的coding：给一个list 是roots of number 1 to 1000, 分别计算even and odd index元素的和

Experience 3

2022-3-22 11:52:59 [Phone]

<https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=871165&ctid=234558>

前段时间面了谷歌DS，以下是记得的题目。求加米看面经呀～～

- How to calculate standard error of mean and median

- 1000 observations, 900 features, linear regression.

    - What’s the problem here?

    - How to improve？

- Online survey red/blue/yellow, trigger is people who searched color, what’s the problem

- Coding: generate normal data

Experience 4

2022-2-11 16:15:36 [Phone]

<https://www.1point3acres.com/bbs/thread-849923-1-1.html>

1. 如何生成 normal random variable in a matrix (答：np.random.normal)；如何话histogram （答：plt.hist); 如何normalize the matrix so that each column sums to 1 （我是直接divide by np.mean(, axis=1))。
2. sample mean 和 median的定义；怎么估计 sample mean的 variance （用sample variance / sqrt(sample size))， 怎么估计sample median的variance （我回答了bootstrap 和 delta method，但是面试官好像比较熟悉boostrap，此处建议大家挑简单的说，毕竟面试不是掉书袋）。
3. 经典的 regression: outcome y~x1+x2, 然后 x3=x1+x2, x4 =x1-x2, 问跑y~x3+x4出来的model 和之前的比有什么区别：linear regression coefficient 和predicted outcome都要说，然后follow up是问 如果correlated会怎么变？ 如果加了一个l1 (lasso) 和 l2 （ridge) 的panelty会怎么变

Experience 5

2022-1-19 18:10:49

<https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=840836&ctid=234558>

1. 怎么计算mean和median？

2. 举例说明什么场景下应该使用mean，而不是median；什么场景下应该使用median，而不是mean。

* Both the mean and the median can be used to describe where the "center" of a dataset is located.
* It's best to use the mean when the distribution of the data values is symmetrical and there are no clear outliers.
* It's best to use the median when the the distribution of data values is skewed or when there are clear outliers.

1. Mse的公式？有什么用处？

Where is the observed value and is the predicted value.

The mean squared error (MSE) tells you how close *a regression line* is to *a set of points*. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.

4.        1000个observation, 900个feature, 会有什么问题？怎么解决？

1. 设计一个实验来验证某项培训是否有效：从一个群体中选出最好和最差的100人进行培训，培训完再测试，和之前的成绩做对比。这样做是否有问题？为什么？

1. 首先问他们是想验证什么 如果他们想验证训练分别对于成绩好和坏的同学的效果 那这样算是基本成立 如果想验证训练generally对于成绩的提升 那么这样有selection bias是不合理的 需要random sampling

2. 假设我们random sampling了 我认为还存在一个问题 就是没有control group 我感觉这个问题更多像是一个causal的问题 就是看成绩的提升是否是由于增加了训练 但如果单纯比较before and after成绩的变化又是不够的 因为没有考虑confounding (困惑 adj.) factor 也许是出现了别的因素导致了成绩的提升我们没考虑到呢 比如说 这时候相比ab testing 我觉得可以做一个difference in difference 找一组对照组 也进行一下前后的成绩差异 然后和实验组前后的成绩差异比较 对照组就是用来减去cofounder的作用

1. Coding: 给了一个list，计算所有奇数的平方根的和，以及所有偶数的平方根的和。

Experience 6

2021-10-31 04:11:15

<https://www.1point3acres.com/bbs/forum.php?mod=viewthread&tid=814542>

上来相互间短介绍后，就问当前工作的项目，项目里承担的工作，以及具体的技术细节。这个准备得不错，要尽量做到interactive，时不时问问面试官do you want to know x more or y?

之后就是case analysis. 问要设计网络问卷，调查最喜欢的颜色，有哪些需要考虑的。这个就要多和面试官clarify。调查目的是什么？发放调查的平台？等等。多把这些细节问出来，相关的技术考虑就比较明显了。

本来说要问些统计的问题，后来可能时间不多了，就直接和之后的coding合在一起。

coding部分就是做simulation.问有x面的骰子，每面被扔到的可能性是Px，都不相同。现在要扔N次，写一个程序，输入P1，P2...Px, 和N，就输出可能的扔到的骰子面。这个可以参考roulette wheel selection的做法。不过我当时并没有写出来，面试官人很好，提醒很多细节我还是写不出来，时间到了就互相道别结束面试。

import numpy as np

px = [0.1,0.2,0.65,0.05]

def gen(px, N):

cdf = np.cumsum(np.array(px))

res = []

for \_ in range(N):

p = np.random.random()

n = 0

while True:

if p <= cdf[n]:

res.append(n)

break

n += 1

return res

res = gen(px, 1000000)

import collections

cd = collections.Counter(res)

print(cd)

# Counter({2: 650416, 1: 199373, 0: 99962, 3: 50249})