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 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: $s = \sqrt{\Sigma(x_i - \bar{x})^2/(n-1)}$.
Standard Error of mean: $s_{\bar{x}} = s/\sqrt{n}$.

Standard error of median:
(1) be approximated by standard error of mean if sample is large.
(2) use an asymptotic formula $s_{median(x)} = \sqrt{\pi/2} \cdot s/\sqrt{n}$.
(3) use a bootstrap method: $\widehat{Var}_B(M_n) = \sqrt{\frac{1}{B-1}\sum_{l=1}^{B}\left(M_n^{*(l)} - M_n\right)^2}$ where $M_n^{*(l)}$ is the lth 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.

4. Mse 的公式？有什么用处？

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

Where y_i is the observed value and $\hat{y}_i$ 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？

5. 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

6. 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。