- Methods to narrow down confidence interval
  Increase sample;
  Reduce variability: difficult, can adjust the way collect data, like paired design to compare two groups;
  One-side confidence interval;
  Lower the confidence interval
- 扔一个筛子三次，出现的数字=return，设计一个策略使 return 最高
  https://www.zhihu.com/question/30470526
  https://www.weiming.info/zhuti/JobHunting/31687701/
- Density, cumulative,
- 写 code 生成二项分布的随机数
  rbinom(n, size, p)
- How to test multinomial distribution?
  Chi-square goodness of fit,
  df = k-1,
  assumption: independent, random sample, <10%*pop, each case only contribute to one cell, each particular scenario have at least 5 expected cases
  assume normal if n/k>=2 and n^2/k>=10
- UserVisit(user_id, page_id, date), 问某一天有多少用户访问某个页面超过平均值
  SELECT count (*) as countHigher, Table.user_id from table Join
  (select (AVG(t) as AVGT, q from table group by q) b on a.q=b.q
  https://stackoverflow.com/questions/31052969/sql-count-values-higher-than-average-for-a-group
- What is least square regression
  Least squares is a statistical method used to determine a line of best fit by minimizing the sum of squares created by a mathematical function
- Derive formula for least squares
- How to test regression model is a good fit
  First, r-squared, it effectively explains how much of variability in the dependent parameter is explained by the independent variable, good fit 0.2~0.8
  Second, t statistic to test slope different from zero
  Third, error terms are normally distributed
  Fourth, domain of the predictor is within the range of sample data
  Model evaluation metric for regression: MAE, MSE, RMSE
- Linear regression: high bias, low variance
- 现在假设你收集了一组数据 A，和一组数据点 B，两者都 measure 同一个变量，但是是在不同 condition 下采集的。问现在手里有一个新的数据点 X，问 X 属于 A 组还是 B 组?
  提供一个 Bayes 的思路。A 的 prior probability 是 P(A) = len(A) / [len(A) + len(B)], 类似 B 的 prior 是 P(B) = 1 - P (A).
  之后知道 x 的就是简单的算一下 conditional probability: P(A | x) = P(x | A) * P(A) / Constant. P(B | x) = P(x | B) * P(B) / Constant. Constant 用来 normalize probability。

If P(A | x) > P(B | x)， classify 为 A；otherwise, classify 为 B。

其实就是 naive bayes 加了一个 prior。

P(x | A) 看你的 assumption，比如画个 histogram 然后看 normal density 的 fit，比较好的就用 sample estimate distribution 的 parameter。prior 的话可以想象 A 是有病的 case，B 是没病的 case，这样的话 len(A)的长度就是有多少有病的 case。

- 有 x1, x2 两个 variables, y 是 target，用 linear regression 来 fit 这个 model。其中 x1 和 x2 是 correlated。那么请问，第一个 model y，x1, x2 和 第二个 model y，（x1+x2）以及第三个 model y，（x1-x2）的 coefficient 是否相同。为什么，怎么证明？用这三个 model 分别 train data 然后 predict data，testing result 会有什么不同？解释一下 MSE 和 variance 的关系，解释下 bias-variance tradeoff.

```
library(MASS)   # allows you to generate correlated data
set.seed(4314)   # makes this example exactly replicable

# generate sets of 2 correlated variables w/ means=0 & SDs=1
X0 = mvrnorm(n=20,   mu=c(0,0), Sigma=rbind(c(1.00, 0.70),   # r=.70
                          c(0.70, 1.00)) )
X0_sum = X0[,1] + X0[,2]
X0_diff = X0[,1] - X0[,2]
y0 = 5 + 0.6*X0[,1] + 0.4*X0[,2] + rnorm(20)   # y is a function of both
m1 <- lm(y0~X0[,1]+X0[,2])
m2 <- lm(y0~X0_sum)
m3 <- lm(y0~X0_diff)
```

```
> coef(summary(m1))
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  5.20281012  0.2546347 20.4324493 2.109130e-13
X0[, 1]     -0.07711621  0.3496694 -0.2205404 8.280772e-01
X0[, 2]      0.73747212  0.3113873  2.3683438 2.998253e-02
> coef(summary(m2))
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 5.1499540  0.2575546 19.995581 9.669046e-14
X0_sum      0.3589687  0.1435241  2.501104 2.225687e-02
> coef(summary(m3))
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  5.443774  0.2596477 20.966001 4.260022e-14
X0_diff     -0.511149  0.3301522 -1.548222 1.389710e-01
```

```
X1 = mvrnorm(n=20,  mu=c(0,0), Sigma=rbind(c(1.00, 0.87),   # r=.87
                          c(0.87, 1.00)) )
X1_sum = X1[,1] + X1[,2]
X1_diff = X1[,1] - X1[,2]
y1 <- predict(m1, as.data.frame(X1))
y2 <- predict(m2, as.data.frame(X1_sum))
y3 <- predict(m3, as.data.frame(X1_diff))
df <- data.frame(pts=1:20, y1=y1, y2=y2, y3=y3)
library(reshape2)
dt <- melt(df, id.vars = 'pts')
library(ggplot2)
```

ggplot(data = dt, aes(pts, value, color=variable)) + geom_line()
Intercept almost no change, coefficient differ a lot, coef_sum is about half of half of the first model, coef_diff absolute is about the sum of the first model slopes, coef_se_sum is half of the first one each, coef_se_diff is the same magnitude as the first one
Variance just measures the dispersion of the values; the MSE indicates how different the values of the estimator and the actual values of the parameters are. The MSE is a *comparison* of the estimator and the true parameter, as it were. That's the difference.
If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.
This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

- 想一个 case， 用 median 衡量 variables 比较正确而 mean 会 misleading； 想一个 case， 用 mean 会正确，而 median misleading
  Whenever a graph falls on a normal distribution, using the mean is a good choice. But if your data has extreme scores (such as the difference between a millionaire and someone making 30,000 a year), you will need to look at median, because you'll find a much more representative number for your sample.
- 已知 salary 的均值 63k, 10percentile 45k, 90percentile 87k， 问 65percentile？
  懵了个圈，然后问 right skewed distribution 有什么？给几个数据，怎么测试是不是 right skewed 的？然后似乎看到我学过金融，问 brownian motion 的 increment 符合正太分布的背后原因？？？
  Right skewed, take lognorm distribution, then mean, variance to estimate
  Right skewed distribution: F dist, chi-sq dist
- 如果 X 服从 Possion (lambda); 条件概率 P(X = 1 | X > 0) = ?

- 案例分析，假设一个公司在 Youtube 上想做一个广告宣传，现在有其中一个城市从 1 月 15-2 月 15 的销售结果，在 2 月 15 投放广告 campaign 之后，一个月的销售额的 plot 图。大概是一个波动很大的图像。问如何判断这个 campaign 的效果显著，提升了销量。如果不能，请问设计怎样的实验才能验证提高了销量？ 楼主答了一个 Difference in difference; 应该是对了，因为之后再向面试官问他们所做的内容的时候，他也提到了这个方法在实际工作中的运用。 楼主前两个题答得磕磕绊绊，感觉这轮能过，大概是因为第三题 运气好碰上了面试官想要的答案。
  https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation

- 第一轮，一个硕士学校的校友 ABC, 简短寒暄。之后开始问题：
  1）假如 X, Y, Z 三个变量 分别都服从正态分布，如何检验 X, Y, Z jointly 服从正态分布？
  The Kolmogorov-Smirnov **test** (K-S) and Shapiro-Wilk (S-W) **test** are designed to **test** normality by comparing your data to a **normal distribution** with the same mean and standard deviation of

your sample. If the **test** is NOT significant, then the data are **normal**, so any value above .05 indicates normality.

X+Y 还是正态，**要求这 X 和 Y 必须是 jointly normal 的。两个相互独立的正态是这种情况的一个特例。**
比如，X, Y 是 jointly normal 的，则， X+Y ~ N(EX+EY, var(X) + var(Y) + 2cov(X,Y)) 。 如果 X,Y independent, 则 cov(X,Y)=0。
一个常见的，非 jointly normal 的两个正态随机变量加起来不是正态的。
X ~ N(EX, var(X)), 是一个正态随机变量。
令 Y= m * X. 其中，m 有 1/2 概率为 1，1/2 概率为-1，m 独立于 X.
可以证明， Y 的分布也是正态的。
但是 X+Y = (1+m) *X 不是正态分布，因为其会在 0 点有一个概率为 1/2 的聚集。

R package MVN to test multivariate normal distribution, Mardia's, Royston's and Henze-Zirkler's tests

- 2）假设一个 data generator 不停 generate range [0,1]的数。有人声称这个 generator 生成数据是随机生成的。。如何检验是不是随机的?
  use a frequency test to illustrate the general principles of the tests. Most, if not all, random number generators attempt to generate numbers that are equally likely. Such a distribution of values is called "uniform" because all possible values are equally or uniformly likely. Frequency tests examine whether the frequency of different random numbers is consistent with the subsequences that would be produced by a uniform distribution.

  Kolmogorov–Smirnov test (K–S **test** or**KS test**) is a nonparametric **test** of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S **test**), or to compare two samples (two-sample K–S **test**).

  ```
  y <- runif(30)
  ks.test(y, 'punif')
  ```

  Chi-sq test
- 3）老题，第二拍卖定价。假设 A 和 B 竞拍。A 和 B 竞价高者获得物品，但是是按价格较低者的报价获得物品。对于 A 来说，每次知道自己的报价，但是只有赢得物品拍卖时候，才能观测到对手的报价。如果报价低于对方的时候，是不知道的。。。假设作为 Player A, 知道 Player B 的竞拍价钱服从一个 exponential distribution with parameter lambda.. 然后 A 和 B 假设竞拍了 100 次，作为 A，你能观测到的就是这 100 次你每次的出价，每次是否获胜。以及获胜时，所需付出的价格（即 Player B 的出价）。如何通过这些 data 和 Player B 的 竞拍策略服从 exponential distribution 来 估计 Player B 的参数 lambda...

  bayesian Nash equilibrium action question

自己想的是用 MLE，不知道对不对。假设 A 和 B 的 bidding 分别为 x，y。当 A > B 时，density for B: lambda * exp(- lambda y)。当 A < B 时，probability for B: exp(-lambda x)。然后 likelihood is their product and find the lambda that maximizes the likelihood. 求大神指点。

Mean: λ^{-1} (= β)
CDF: $1 - e^{-\lambda x}$
Variance: λ^{-2} (= β^2)
PDF: $\lambda e^{-\lambda x}$
Parameters: $\lambda > 0$ rate, or inverse scale
Median: $\lambda^{-1}\ln(2)$
Quantile: $-\ln(1 - F) / \lambda$

- 第二轮，一个中国大哥 主要面概率
有 1 到 N N 个数，从中随机抽取第 1 个数，第 2 个数无放回。如果第 2 个数小于 1 第 1 个数，就停止。如果第 2 个数大于第 1 个数，就可以继续抽取，直到抽出的数 x(n) 小于上一轮抽到的数 x(n-1). 假设这个随机生成数列为 X，求 X 的长度的期望。

```
myFun <- function (N) {
 arr <- c(1:N)
 p <- c()
 x <- sample(arr, 1)
 p <- c(p, x)
 x <- sample(arr, 1)
 while (x >= tail(p,1)){
  p <- c(p, x)
  x <- sample(arr, 1)
 }
 p <- c(p,x)
 return (length(p))
}

mean(replicate(1000, myFun(100)))
```

- 第二题，OLS 的基本假设条件，如果 data 所有点都多了一份拷贝，那么估计参数，估计参数方差会如何变化？
The regression model is linear in the coefs and error term;
Error term has a population mean of zero
All independent variables are uncorrelated with the error term
Observations of the error term are uncorrelated with each other
Error term has constant variance (no heteroscedasticity)
No independent variable is a perfect linear function of the other explanatory variables
Error term is normally distributed

- 第三题， 辛普森悖论是什么，为什么会出现这个问题?
Simpson's paradox occurs when groups of data show one particular trend, but this trend is

reversed when the groups are combined together. Understanding and identifying this paradox is important for correctly interpreting data.

- 第三轮，被韩国大哥坑。因为这一轮突然换成 video 面试，然后出现了技术问题，导致只面了 25 分钟。这轮本来是一个 coding 题目。但是我实在看不懂这个题目。。说假设从一个大 data population 里面做 subsampling, 然后我可以生成一个关于 subsample 的 confidence intervals. 做 100 次 subsample 可以生成 100 个 confidence intervals...假设现在我能观测到这 100 个 confidence intervals, 如何估计整个 data population 的 confidence interval。。。20 分钟太短，题目又太晦涩难懂，导致这轮做的非常的不舒服。。。

  第四轮，一个 open business problem...机器学习题目。如何设计 Google 在第三方网站投放网页。应该展示什么广告，类似于做一个广告推荐系统。。

- x 正态分布 n(u, 1), 已知一个数据点 x1. null hypothesis u = 0. 用什么 test？若已知 x1 > 1, 对测试有什么影响?
  Likelihood ratio test??
- android 升级后的用户比没升级的用户搜索更多，为什么不能说升级导致搜索变多？ 如果要验证，怎么设计实验?
  user 会主动升级可能本身就说明他们是更 active 的 users 或者 tech savy。所以本来就是 biased 比较。同意比较 new users。也可以做 A/B test 但是有一组不给新的版本。另一组用给升级的选择。
- 总共有 n steps， 每次只能爬 1 或 2 steps， 有多少种方法
  Recursive method:

```
def findStep( n) :
   if (n == 1 or n == 0) :
      return 1
   elif (n == 2) :
      return 2

   else :
      return findStep(n - 2) + findStep(n - 1)


# Driver code
n = 4
print(findStep(n))
```

- 扔筛子如果是 1-4 则保留结果，5-6 重新扔，得到的结果是什么就是什么，求期望和 variation
  $D(\xi)=E(\xi^2)-[E(\xi)]^2$
- linear regression, causation 和 correlation 的关系，怎么证明是哪个?

Regression: manipulate predictor variable to predict dependent variable
Correlation quantifies the degree to which two variables are related. Correlation does not fit a line through the data points. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.

You can only have weaker or stronger *evidence* of causality.

1. When it comes to **correlation**, there is a relationship between the variables. **Regression**, on the other hand, puts emphasis on how one variable affects the other.
2. **Correlation** does not capture causality, while **regression** is founded upon it.
3. **Correlation** between x and y is the same as the one between y and x. Contrary, a **regression** of x and y, and y and x, yields completely different results.
4. Lastly, the graphical representation of a **correlation** is a single point. Whereas, a **linear regression** is visualized by a line.

- 湾区 income 和全美国 income 的对比，猜测会有什么不同，能否代表全美国的情况？如果想建模分析湾区的 income, 应该考虑哪些 feature
- 是个 SQL 题，因为不难，我忘记题目了。大概就是 groupby, count, sum,case when 就可以解决的
- case study,现在假如做个试验，旧手机和新手机的某个 app 使用量，发现新手机上的打开和使用次数远远大于同时间内的旧手机的打开量, 可否得出新手机推动了这个 APP 的使用次数，如果能为什么？如果不能，为什么？如果让你做这个实验，你怎么做。答案应该是不能。
- case study, 假如某个产品在 12 月做出了改变，然后在 1 月 1 号-10 号之间产品的点击量增长了很多，能否说明这个改变可以增加用户点击量?
- What is GP, different with MVN; Gaussian Process 和 multivariate normal
  Practically they are the same within any finite interval. Aka within any finite time interval, GP is just a MVN with observation at any particular time point being normally distributed.
  MVN is distribution over vectors while GP are distribution over functions
  https://www.quora.com/What-is-a-Gaussian-process
- Mobile apps, monitor user engagement every week, what metrics, how to measure/calculate
- Regression on multiple predictors with OLS, removed all predictor if p >0.05 and the resulting model doesn't make sense, how do you explain to the product manage
  In a model, if we believe some variables should be kept, however, those variables are not significant accoring the model output. Explain why and what we should do?
- How to explain mean/mode/median to 8 years old
- linear regression assumption.
- Conditional mean and conditional variance
  https://newonlinecourses.science.psu.edu/stat414/node/116/
- Regularized regression
  When experience multicollinearity, over-fitting, and model too complex to interpret, an alternative to OLS is to use regularized regression (penalized models to shrinkage methods) to control parameter estimate.
  http://uc-r.github.io/regularized_regression
- 然后有问 OLS 的 Coefficient SE 是什么意思

Shows the standard errors of each coefficient estimate

- 传统的 paired t-test 和 2 sample t-test 的差别
  An Independent T-Test is used to examine if there is a difference in the means of two DIFFERENT groups(i.e Independent variable: Group 1: Police Officers/ Group 2: Soldiers.
  Paired Samples T-Test: To test if there is a difference in the levels of aggression between the members of only ONE group, or different but dependent group
- 问了一下做过的 Statistical Project 延申的问题包括：p-value 含义，z-score 的计算, model validation metrics 之类的
- 假设有一个 dataset D, 有 m rows 和 n columns. 还有一个 dataset D-, 是 D 所有 observations duplicate。问如果在 D 和 D-上分别建了一个 regression model，这两个 regression models 有什么不同。
  答案是：regressIon 相同，因为 coefficients 相同，但每个 coefficient 的 CI 有变化。CI = mean +/- 1.96 sd., 因为 n 变大了，sd 减小了，所以 CI 的 range 变小了。
  还是这个题，用 D-做 regression violate regression 的哪个 assumption, 答案是 observation iid.
- coding 题。
  A B C
  1 2 3
  4 5 6
  7 8 9
  。
  。
  。
  任选一种语言，read in this csv, calculate row variance 和 column variance, 然后如果 A, B, C 里面有很多 0, 怎样计算 A, B, C 除零外的 column variance 和 row variance.
- 捕捉标记题，具体数字记不清了，用字母代替，比如第一次捕到 a 只熊，标记之后放回，第二次捕了 b 只其中 c 只有标记， 问一共多少熊，follow up question 是如何用一个更统计的方法估计，面试官给的 hint 是把第二次捕捉看成 b 次 bernoulli trial, 所以合起来就是一个 binomial distribution。
  听着耳熟，就搜了一下...https://en.wikipedia.org/wiki/Mark_and_recapture 不过根据面试官的 hint，binomial 的话,感觉可以构建一个 confidence interval，第二次抓的越多，interval 越小。如果抓的动物多了还可以用 normal approximation，置信区间 P +- Z* sqrt(p*(1-p)/n)
  Hypergeometric distribution

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

- 停车位题。比如你要去上班，公司在一条街道的最底端，你从街道的最上端向公司驶去，在街道的一侧有连续的 N 个停车位，问用怎样的策略可以停车后到公司的步行距离最近。中间一步步问了几个小问题， 比如每个停车位之间是否独立，如何检测停车位之间的相关

性。最后假设停车位的 availability 是 i.i.d.的， 然后用的策略是停在第一个 available 的车位，问停车后步行到公司距离的期望。（看成一个 negative binomial distribution 就好了）最后还问了除了停在第一个 available 车位和一直开到街道最底端公司门口找车位以外还有没有别的策略。

感觉可以先开到 1/3 处，统计一下空位概率设为 P。然后再往前找，这时候感觉这个问题和第一题有些类似，假设还有 100 个车位，知道 P=0.1，那也就是说平均有 10 个空车位，这时候也可以算出一个置信区间，比如 95%的概率空车位的数量在（4.12, 15.88），那么我保险一些，我就第 5 个空位停下，可是可能离得远。这时候距离和有空位的概率是一个 tradeoff，好了需要高手出现指导一下如何解决这个 tradeoff！

- dimension reduction 怎么做，答 regularization 比如 lasso 或者 ridge，问 nonparametric 有什么方法吗，答 pca 或者 cca，紧接着问 pca 的细节，答 spectral decomposition，再问了 cca 的细节。之后问了 lasso 和 ridge 的区别。有 outlier 怎么办。lasso 和 ridge 对 outlier 的影响是什么。

  Drop columns with too many missing values, drop columns with low variance, decision tree (ultimate solution to tackle missing values, outliers and identify significant variables), random forest (in built feature to select subset, but bias to variables have more # of distinct values), remove high correlation, backward feature elimination, forward feature selection, factor analysis (exploratory factor analysis, confirmatory factor analysis), PCA
  PCA, ICA, CCA: https://www.cs.cmu.edu/~tom/10701_sp11/recitations/Recitation_11.pdf

  lasso methods work by restricting the sum of the absolute value of the coefficients to being less than a set value, so the impact of outliers is significantly curtailed.
  Why use lasso (meant to deal with outliers AND model selection) rather than ridge regression or some other method meant to only deal with outliers? What's got you so worried about outliers in the first place? If it's in the data, you shouldn't ignore it.
  So i'm actually using elastic net. I didn't think ridge handled outliers though and just took care of correlation between predictors. I have a lot of predictors and I'm looking at interactions so I'm using lasso for variable selection.
- 还有一个 ab testing 的问题，是说如果有两组不一样颜色的 button，想比较 ctr 的区别，怎么设计实验，怎么进行比较。就是简单的 two sample t test
- 怎么 estimate bivariate normal 的 parameters；如果 data 里面只有 y<x 的情况下，要怎么 estimate parameters。如何 rank ratings 给很多 video
  https://stackoverflow.com/questions/37294131/how-do-i-estimate-the-parameters-of-a-bivariate-normal-distribution-in-r-from-re
- time series 怎么判断 AR 和 MA 的 order；如何看几个工厂的 defective rates 有何区别
- 一道简单 sql 题，一道 product：如何比较两种 query rankings 的结果
- 一组数据里面有[1s, 0s, NAs]，NA 表示 value 未知。estimate 1 的 proportion 以及总共有多少个。之前面经里面出现过。NA 的 proportion 是多少最好。

  theoretically, 25 to 30% is the maximum missing values are allowed, beyond which we might want to drop the variable from analysis.

- 一个 dataframe，用 groupby 和 max, min 基本可以解决
- 如何比较两个 right-skewed 的 distribution
  generalized Wilcoxon test, also called the BrunnereMunzel test, should be used
  mean or median
  https://www.jclinepi.com/article/S0895-4356(10)00257-X/pdf
- You're running an experiment to test the redesign of google.com. We've made a change in how search results are ranked. Construct a few metrics to gauge the success/failure of the experiment.
- They collected a sample of 50 users from each algorithm. The number of users that said they were satisfied was: 45 in the new algorithm and 40 in old algorithm. How would you help that team to interpret these results? Which questions would you ask that team
- Given a data set with two variables: a boolean "was link clicked" and an integer "font size" (that only takes on a few different values) how would you specify a logistic regression model to model the relationship between the two variables
- Google serves searches from a large number of data centers distributed across the world. You're interested in analyzing the distribution of searches per user. You have a workstation that is connected to all of these data centers, but it doesn't have enough storage to download all the logs from all the data centers and communication between it and the data centers is relatively slow so should be minimized. You are able to run arbitrary code on each data center and on your workstation.
- Part 1: how do you calculate the mean number of searches per user across all of Google?
  Part 2: how do you calculate the variance of the number of searches per user across all of Google? (no clue what he wants to ask)
- Explain logistic regression, 用 R 怎么写

  怎么测试 LR？哪些统计数据可以看？ 怎么用英文解释他们

  Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
  https://datascienceplus.com/perform-logistic-regression-in-r/
- ks 的 statistics 和 ols 的 beta 方差公式
- ads optimization 问题最后以 ridge 项的 bayesian updating 结束
- What are the assumptions of Ridge and LASSO Regression
  The basic thing to remember about Ridge and Lasso is that they are both parametric methods. What this means is that for them to be applicable, a specific model has to be postulated, usually a linear one.

  The major advantage of these methods compared to OLS is that they can handle multicollinearity, i.e. a predictor matrix with rank less than the number of its columns.

  Another thing to remember is that neither Ridge nor Lasso actually respond well to outlying observations. This may be seen most easily for the case of an orthonormal predictor matrix as

then the estimators may be written as (unbounded) functions of the notoriously non-robust OLS estimator. Therefore, much like the OLS estimator, Ridge and Lasso should be used with caution in non-clean datasets.

- stratified sampling
- ab testing, optimization, regularization, order of series, sql, time series, hypothesis testing, bonferroni, fdr, repeated measurement, linear regression
- R codeing (dplyr::group_by+summarize)
- Ridge v.s. Lasso
- 看 data distribution 想要怎么做 t test
- Logistic regression
- Causal inference: propensity score matching, why is randomized experiment the gold standard
  （这部分应该是根据我的简历和面试官的经历）
- Probability (unfair coin to get 50% probability)

  https://en.wikipedia.org/wiki/Fair_coin#Fair_results_from_a_biased_coin

  If a cheat has altered a coin to prefer one side over another (a biased coin), the coin can still be used for fair results by changing the game slightly. John von Neumann gave the following procedure:[4]

  1. Toss the coin twice.
  2. If the results match, start over, forgetting both results.
  3. If the results differ, use the first result, forgetting the second.
- R coding (use CDF to sample a random variable; what if only need to sample 99 percentile or above)

  笨办法：sample from Uniform[0,1]，如果不在 99 percentile 以上就重新 sample，然后用 inverse CDF 得到需要的 random variable；
  聪明办法：直接 sample from Uniform[0.99,1]。
- 给你一个均匀骰子，最多可以投 N 次，问你采用什么样的策略使得你最后一次投掷的点数最大。（为啥我第一反应是一个女生这辈子会遇到 50 个男生，然后选第几个那道题，1/e）
- 一个骰子，如何判断它是不是均匀的。

  Chi-sq test

https://rpg.stackexchange.com/questions/70802/how-can-i-test-whether-a-die-is-fair

- 一个 hotel，最多有 N 个 room，客人 make reservation 后有可能不来，给你一个 tolerance rate，设计一个 model 并给出在可以容忍的范围内，最多可以发出多少个 reservation。（N+1, N+2, ...）（我居然一开始说正态分布，经她提醒才知道二项分布就完事了。）

  The question: A hotel has 230 rooms and uses an overbooking policy. The probability that a customer cancels or does not show up at the hotel is 0.12. Find the maximum number of rooms the hotel can book and still be 85% sure everyone who turns up will have room.

  https://math.stackexchange.com/questions/1700456/binomial-distribution-question-holiday-resort-overbooking

  the approximation to X~B(n,p) is given by the normal distribution N(np,np(1−p))

  The standard normal is

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

$$Z = \frac{230 - 0.88n}{0.32\sqrt{n}}$$

  Then you need the 15% upper tail of Z which is ≈ 1.04. (You want to be 85% sure $100 - 85 = 15$)

  Now you solve this for n to find the number of rooms you are looking for

$$1.04 = \frac{230 - 0.88n}{0.32\sqrt{n}}$$

- 一共四轮，把记得的题目写一写吧，写以常用分布 generate density 为 f 的 r.v. 的 code，用 bootstrap 计算 two sample mean difference 的 confidence interval，需 genenrate 同样分布但>c 的时候 code 要怎么改，白板写。 logistic regression 及 test。各种 a/b test，每轮都有，用什么 metric，怎么 test， 怎么 design experiment。 很多面经提的捕捉－标记－放回－再捕捉估计熊数目的题，注意这题要考虑总数不够大，用无放回来算。介绍自己的一个 project，怎样估计手机电池还能用多久，多久需要更新一次 model（这样的 open ended question 也很多，怎么样估计 xxx，需要什么 feature 什么 model 需要自己想）。有一个给人做 survey 的 app，怎么预测这个人是不是 bad user (每次都瞎填）。问怎么 Measure 一个地区 不同人种 mixed / seperated 程度。

- 描述一个 data analysis 的 project. 期间面试官会问一些问题
- data manipulation, r 或者 python 都可以。面试官当场用邮件发给我一个 csv file, 内容是不同国家几种商品不同 vendor 的价格， 有以下几列
  country | type | vendor1 | vendor2 | vendor3 | vendor4
  india | 1 | 13.5 | 14 | 15 | 14.5
  要求先算 vendor1-- vendor4 的 overall median. 然后 group by 国家和 type 再算 vendor1 -- vendor4 的 median.
- 还是上面那个数据，怎么分析 4 个 vendor 之间价格是不是有区别
  ANOVA
- single linear regression with interaction item，怎么处理 heterogeneous variance in error quadratic term?
  If group level heterogeneous variance, treat as random factor in mixed model fit
  https://quantdev.ssri.psu.edu/sites/qdev/files/ILD_Ch06_2017_MLMwithHeterogeneousVariance.html
- Youtube 教育视频 打分。 2000 个视频，100 个人打分 1-5. 每个人看 100 个视频，每个视频有 5 个人打分。 问怎么整合打分为一个 metrics。
- 50 个人各扔硬币 100 次，求 head 次数最大值的期望，CI， 如果是 biased 怎么办。
  http://www.math.ucsd.edu/~gptesler/283/slides/longrep_f13-handout.pdf

  ```
  z <- replicate(n = 50, sum(sample(c(0,1), replace = T, prob = c(0.5, 0.5), size = 100)))
  mean(z)
  sd(z)
  mean(z) + c(-1,1)*1.96*(sd(z)/sqrt(100))
  ```

- 2 个候选人票数， A:60， B:40 可以有什么结论
  z 检验, one sample proportion test
- time series，如何自动找到 season 的周期。楼主说 detrend 之后画图看波动周期，但面试官说要自动 detect。楼主当时表示不会，后来查了一下，可能可以写个 code，自动找 local max，然后看两个连续 max 之间的距离。
  做 Spectral density estimation 然后找 max
- 说有一个连续函数 f，画出来就是一个 curve，问什么方法求 curve 下的面积。但是函数 f 不知道，但是有另外一个函数 g(x, y)。g(x, y) = 1 if f(x)<y; g(x,y) = 0 if f(x) >=y。
  楼主先说，把 x 轴上 f 的取值范围[a,b]，等分 100 份，每个小区间上找中点 x0，用二分法找 f(x0) 的值，区间的 length*f(x0) 估计这个区间下 curve 下的面积，然后都加起来估计整个 curve 的面积。面试官问还有其他方法吗？ 回答，做 simulation。已知取值范围：x~[a,b], f(x)~[c,d]. Generate 随机数 uniformally（x,y)，发到函数 g 里，数多少 g(x,y) =0.
- 很多统计问题，什么是 t test，z test， 怎么用。比较两个 group 的 ctr，用 z test 还是 t test？statstics 的公式是什么，怎么确定 sample size。
- 问一个硬币连续扔出 0 或 1 后，出现第一个 1 或 0 的投掷次数期望是多少

$$\mathbb{E}(first\ 0\ or\ 1) = \sum_{i=1}^{\infty}(i \cdot \mathbb{P}(i_{th}\ time\ appears\ 0\ or\ 1)) = \sum_{i=1}^{\infty}(i \cdot 0.5^i) = 2$$

Geometric distribution
P(x=k)=p^(k-1)*p
Mean=1/p
Var = (1-p)/p^2

第一题的要点我觉得是考察 geometric distribution 具有 "memoryless"的特点，不一定正确希望和大家讨论。
goal 是求 E(# of flip to get first 1 after two consecutive 0) 或者 E(# of flip to get first 0 after two consecutive 1), 根据对称性，这两个期望是一样的。
E(# of flip to get first 1 after two consecutive 0) = E(# of flip to get first 0 after two consecutive 1)
我们以第一个期望为例，由于 geometric distribution 的"memoryless"的特点:
P(X>s | X>t) = P(X>s-t), 把 t 当做 get two consecutive 0 的时间点，s 为我们得到 get first 1 after two consecutive 0 的时间点，根据等式得到 get first 1 after two consecutive 0 的 distribution 和 get first 1 的 distribution 是一样的，而 get first 1 就是经典的 geometric distribution 了，所以 E(# of flip to get first 1 after two consecutive 0) = E(get first 1) = 2. 同理可以算 E(# of flip to get first 0 after two consecutive 1)也是 2.

- data analytic，说给了一个 dataset 是有关 youtube 的视频观看量的，有 user_id(上传者 id)，video_id, #views（被看了多少次）, tags（分类）, days（几天前上传的）,然后要建立一个 priority 的 system 来给新上传的视频排序，因为处理能力有限，所以要先上传优先级高的视频，比如 ladygaga 的新歌什么的，问怎么可以排序
  我说可以按上传者的 total #views 来排序，然后谁在前谁的 video 先被处理，教授说那万一有个人传了几千个视频比嘎嘎小姐一个视频还多呢，我说那可以用（total views/ # videos)算平均观看量，教授说那大明湖畔的 tags 你还记得吗，我说可以用一个 dict 来存 tags 的 popular 程度，就是 total views of certain tag/video frequency，然后也做一个指标。教授又说那天数呢，有个祖传的视频在我们这几年了积攒了好几百万的点击量和新视频能一样吗，我说那就限制的天数或者找到视频第一天的播放量来比较，他说你觉得怎么能找到第一天的播放量呢。我说这个应该是 possion 分布吧，他说为什么，如何建立模型，我只好说我只会套公式，此时估计教授内心是崩溃的。他说好吧，最后问一下如果建立了这个新的排序系统那如何来验证和以前的比较是否更好呢，我说可以找两个 sample 然后 randamize 的放一些视频，一个按照旧方法来排序处理，一个按新方法来，这个应该是 paired t test 吧，然后看是不是 significant different。他问为啥你觉得是 t test，我就随便说了下 ttest 的 assumption，估计教授已经全面了解了我渣渣的统计学知识， 就放过我了
- 是经过朋友推荐给谷歌的 recruitor 的。Recruitor 一开始说谷歌的 Data Scientist 有两个 track，让我选一个。Quantitative Analyst 对技术要求高一点，Product Analyst 更注重 stakeholder management.我一开始选 PA，问卷也填好了， 结果 recruitor 看完问卷觉得我更

合适 QA，我就说好啊。于是约了店面。店面问了很仔细的 linear regression 问题，假设是什么，怎么 check，不符合怎么办。还有我最得意的 project，具体做了什么，技术细节。还问了一个 distribution 的问题，我没答出来。以为就挂了。结果还给了我 onsite。四个面试，有问到圆里面的点 simulation 怎么做[就是说只给你一个随机均匀分布，怎么以此模拟随机抽取单位圆里面的点。]，还有 optimization 问题[给定一个总数（N），要分成 n 个区间，要你给出最佳的分割点，当然他也不给你说什么是最佳，就你得一步步跟他就具体的情景讨论需要优化的的是什么指标。]，问到我的项目，还有机器学习的一些基本概念，以及给定两个集合知道他们各自的平均数，对于他们的合并集的平均数有什么猜想，怎么验证。面下来我也不知道他们到底要的是什么，有些题我 struggle 了蛮久。

```
t = 2*pi*random()
u = random()+random()
r = if u>1 then 2-u else u
[r*cos(t), r*sin(t)]
```

- Here it is in Mathematica.

```
f[] := Block[{u, t, r},
  u = Random[] + Random[];
  t = Random[] 2 Pi;
  r = If[u > 1, 2 - u, u];
  {r Cos[t], r Sin[t]}
]

ListPlot[Table[f[], {10000}], AspectRatio -> Automatic]
```

How to generate a random point within a circle of radius *R*:

```
r = R * sqrt(random())
theta = random() * 2 * PI
```

(Assuming `random()` gives a value between 0 and 1 uniformly)

If you want to convert this to Cartesian coordinates, you can do

```
x = centerX + r * cos(theta)
y = centerY + r * sin(theta)
```

https://stackoverflow.com/questions/5837572/generate-a-random-point-within-a-circle-uniformly

- 如果有个 software engineer 跑一个 model，本来预期 significant 的 一个 feature 结果显示不显著。是什么原因，怎么解释
  follow-up, 如何选择 feature
  collinearity
  sample size too small
  overfit?

Cross-validation to choose feature

- 一个 dataset, with two columns. 第一列是 click, 0 或者 1， 第二列是 cost, 是连续值。average cost per click = sum of cost / sum of click. 问如何给出 average cost per click 的置信区间 （hint：没有公式可算)
  Bootstrap sampling
  Taylor expansion is an alternative: http://www.stat.cmu.edu/~hseltman/files/ratio.pdf

- 问了好多遍, 我后来理解下来是这样，不清楚对不对：100 bundle of purchases, 有三种 models of cars. 知道每个 bundle 的 total price, 和每种车的数量，问如何估计 price for each model.
- 第一轮：国人小哥，1）面条题 (碗里有 N 个面条，随机拿起两头接上，直到没有头可接，最后碗里面条圈的数量的平均值)；2）关于 simpson 悖论的理解；3）一组很基本的 logistic regression 问题；4）如果 training data 和 test data 的 feature 顺序不一样（但是不知道如何不一样），应该如何让 model 的错误最小化【面试官：修改 loss function】
  第二轮：美国黑人，1）聊实习项目，问得比较细；2）NBA 季后赛打到七轮的概率，如果主客场胜率不同呢；3）如何判断两个 classification model 的优劣，很 open 的讨论
  第三轮：不像中国人的东亚人，1）pandas 简单题，简单得我都没记住；2）用 pandas 从 frequency table 求中位数；3）求一个只有图像没有公式的函数的积分，置信区间；4）卖飞机票，价格固定，有的乘客可能不来，但是如果乘客多了不够坐需要赔钱，优化卖票数量。
  第四轮：白人，东欧的？1）聊实习项目，问得比较粗略；2）关于搜索引擎广告的 A/B test case study （检验一个广告在两个国家效果是否一样，让我列举有用的 metric，我说了几个，然后他选了一个 metric，假设了一些数据，让我做 hypothesis test，再问有什么潜在的问题，我说 variance 可能太大，然后问我可能导致 variance 大的原因，如何应对）；3）有两个数组，把他们合并后的中位数和各自的中位数有什么关系
- How do we messure user engagement for google maps app? What metrics? How to collect data and what data we need to collect? If we found there is no change in A/B testing, what is reason?
- Google map, 现在新加了一个 feature，如何 evaluate？如何设计实验（AB）
  如果给所有人都加了 feature，怎么 evaluate？（比较 feature 前后）
  前后对比，人数增加 5%， 好么？
  本来市场就会增加，如何 quantify？ 如何同时 quantify（我这里提到了用 regression，反正是过了)
- ==========P4 G DS done==========
- 听谷歌音乐中驾驶音乐是否降低驾驶速度的题目
  细节都是面试官现场编的，已知 1000 万用户，听过驾驶音乐的大概 3%左右，要求构造统计模型估计影响