**Probability**

Knowledge

1. Bayes Theorem

Interview Question

# 1. Suppose you toss a fair coin 400 times. What is the probability that you get at least 220 heads? Round your answer to the nearest percent.

# A: This is a central limit theorem question. The trick is to view each toss as a random variable that returns 1 if a head is tossed and 0 if a tail is tossed. Then each such random variable has expected value 1/2 and variance 1/4. So your Z-variable (for using the central limit theorem) will be -

# So we've reduced the question to asking what's the probability that Z takes a value bigger than 2. Recall that on the standard normal, the probability that z takes values between -2 and 2 is about 95%, so the probability that it takes values less than 2 is about 97.5% (it's actually more like 97.7% but just estimating). So the probability that we are bigger than 2 is a little less than 2.5%, which after rounding to the nearest percent gives us 2%.

2. Suppose you toss a fair coin 1000 times. How to test if the coin is fair or not? What to do if only toss 10 times? (Google)

A: 1. Calculate *p*-value using normal approximation. Use two-tail test. (see Question 1)

2. Calculate exact probability (binomial distribution).

2. The probability of a bus passing through a certain intersection in a time window of 20 min is 0.9. What is the probability of the same bus passing through the same intersection in 5 min? (Microsoft - Yammer)

A1: 1 – (1 – X)4 = 0.9; X ≈ 0.4377.

A2: Poisson distribution, λ – number of buses pass through the intersection in 1 min.

P(X <= 20min) = 1 – e-20λ = 0.9

P(X <= 5min) = 1 – e-5λ =

3. Probability Riddle Loaded Revolver

Gretchen will place two bullets into a six-chambered revolver in successive order. She will spin the chamber, close it, and take one shot. If Henry is still alive, she will then either take another shot, or spin the chamber again before shooting. Spin or not spin after an empty shot? (WalmartLabs)

A: spin – 1/4 chance got shot; not spin – 1/3 chance got shot.

4. Rockets are launched until the first successful launching has taken place. if this does not occur within 5 attempts, the experiment is halted and the equipment inspected. suppose that there is a constant probability of 0.8 of having a successful launching and that successive attempts are independent. Assume that the cost of the first launching is K dollars while subsequent launching cost K/3 dollars. whenever a successful launching take place, a certain amount of information is obtained which may be expressed as financial gain of, say 'C' dollars. if 'T' is the net cost of this experiment, find the probability distribution of T?

A: P(T = K - C) = 0.8

P(T = 4K/3 - C) = 0.2 \* 0.8 = 0.16

P(T = 5K/3 - C) = 0.22 \* 0.8 = 0.032

P(T = 6K/3 - C) = 0.23 \* 0.8 = 0.0064

P(T = 7K/3 - C) = 0.24 \* 0.8 = 0.00128

P(T = 7K/3) = 0.25 = 0.00032

5. A candidate is selected for interview for 3 posts. The number of candidates for the first, second, third posts are 3, 4, 2 respectively. What is the probability of his getting at least one post?

A: P(at least one post) = 1 – P(no post) = 1 – 2 / 3 \* 3 / 4 \* 1 / 2 = 3 / 4

6. Find the expectation value of number of times you need to pick numbers to find a number smaller than the number you pick from a hat containing 1 to n.

Case 1: if you replace

Case 2: if you don't replace

A: If getting a number smaller than the number that is picked first is considered a success, otherwise a failure, then these events become Bernoulli trials. The question asks to find the expectation value of number of times you need to pick numbers. The answer to the question is how many trials before a success which is provided by geometric distribution, whose mean, E[X] = 1/p, where p is the probability of success.

Case 1: Let m be the number picked, then probability of choosing a smaller number is m-1/n. The expected value is therefore 1/(m-1)/n = n/(m-1).

Case 2: ???

7. There are 6 pairs of black socks and 6 pairs of white socks.What is the probability to pick a pair of black or white socks when 2 socks are selected randomly in darkness?

A: P = 1 – 12 \* 12 / (24 \* 23 / 2) = 11/23.

8. There are three persons A, B, C. What is the probability of hitting the target twice when 2 persons are selected at random.

A: P = (P(A) \* P(B) + P(B) \* P(C) + P(C) \* P(A)) / 3

9. [What is the probability of 5 people with different ages sitting in ascending or descending order at a round table?](http://www.careercup.com/question?id=16425697)

A: (5+5)/5! = 1/12.

10. In a pocket calculator, a person is randomly typing a 8 digit number. What is the probability that the number looks the same even if the calculator is turned upside down?

A: if we fill first 4 locations with any of 7 numbers {0,1,2,5,6,8,9}. Rest of the 4 locations in a valid number will be filled by(in corrosponding digit) {0,1,2,5,9,8,6}. But we can't use 0 as first digit.

P = 6 \* 73 / 9 \* 107.

11. Having an infinite stream of numbers write a function to take an element with equal probability for each.

A: Reservoir sampling.

12. We toss a fair coin n time. A k-streak of flips is said to occur starting at toss i, if the outcome of all the k flips starting from i th flip is the same. For example, for the sequence HTTTHH, there is a 2-streak occurring at 2 nd toss, there is a 2-streak occurring at 3rd toss, and there is a 2-streak occurring at 5th toss. Here the total number of 2-streaks is 3 in the sequence HTTTHH. What is the expected number of k-streaks which you will see in n tosses of a fair coin?

A: ???

similar question -

<http://www.askamathematician.com/2010/07/q-whats-the-chance-of-getting-a-run-of-k-successes-in-n-bernoulli-trials-why-use-approximations-when-the-exact-answer-is-known/>

13. Input: 4 jars and 50 balls of different colors (Red, Green, Yellow, Blue) where each jar can contain a maximum of 100 balls. Problem: When a user draws a red ball he loses his money while if he draws a ball of some other color his money is doubled. Arrange the balls in such a way that the user has highest probability to lose.

A: r1; r1; r48g50; y50b50; P(red) = 0.25 + 0.25 + 0.25 \* 48 / 98 = 0.62.

14. Given a normal dice and a dice with blank faces, fill in the blank dice with numbers from 0-6 so that the probability of each number coming up, when you roll the two dice together, is equal.

A: 36 combinations for possible outcome from 1 to 12. Need 3 cases for each number. Mark 3 faces with 0 and the other 3 with 6 so each number from 1-12 has the same probability of 1/12.

15. Five people are to be seated randomly around a circular table. What is the probability of two of them sitting next to each other?

A: 5 people can sit around a circular table in 4! ways. Consider these two people as one unit so 3 people + 1 Couple can sit around a round table in 3! ways. But the couple can exchange seats in 2 ways. So the total probability is 2 \* 3! / 4! = 1/2.

16. [What is the probability of a random generator generating 10 consecutive numbers in ascending order (assume it is a perfect random generator)?](http://www.careercup.com/question?id=4519850)

A: (N – 9) / N10.

17. You play a dice rolling game, you have two choices: 1. Roll the dice once and get rewarded the amount of $ equal to the outcome number (e.g, $3 for number "3") and stop the game; 2. You can reject the first reward according to its outcome and roll the dice the second time and get rewarded in the same way and stop the game. Which strategy should you choose to maximize your reward? (i.e. what outcomes of the first roll should make you play the second game?) What is the statistical expectation of reward if you choose that strategy?

A: If first roll less than 3.5 (expectation for one roll), then take the second chance.

E = 0.5×3.5 + 0.5×5 = 4.25

18. There are 1000 balls in a bag, of which 900 are black and 100 are white. I randomly draw 100 balls from the bag. What is the probability that the 101st ball will be black?

a)9/10 b)More than 9/10 but less than 1 c)Less than 9/10but more than 0 d)0 e)1.

A?: 9/10.

19. McDonald’s cuts a deal with Michael Jordan. They will print a large (and equal) number of each of twelve “action photo” trading cards showing highlights of Jordan’s basketball career. If you get a single randomly-selected card each time you eat at McDonald’s, what is the expected number of meals would you need to buy in order to collect all 12 cards? (Google)

A: Use geometric distribution to calculate the expected number of meals for each card and then get the sum.

1. How many meals will it take to get your first card? Obviously, 1.

2. What is the expected number of meals, after getting your first card, that it will take to get a different card?Probability is 11/12; expected number of meals is 12/11.

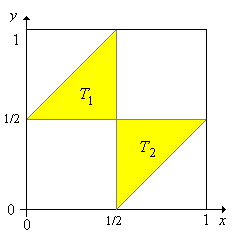
3. After getting two types of cards, what is the expected number of additional meals it will take to get your third type of card? Probability is 10/12; expected number of meals is 12/10.

4. And so on. The expected number of meals needed to get a full collection is therefore 1 + 12/11 + 12/10 + 12/9 + 12/8 + 12/7 + 12/6 + 12/5 + 12/4 + 12/3 + 12/2 + 12/1 = 37.24.

20. A straight stick is broken at random in two places chosen independently and uniformly along the length of the stick. What is the probability that the pieces can be arranged to form a triangle?

A: *T*1 = {(*x*,*y*)∈*S*: *y*>1/2, *x*< 1/2, *y*−*x*< 1/2}∪*T*2 = {(*x*,*y*)∈*S*: *x*> 1/2, *y*< 1/2, *x*−*y*< 1/2}

P = 1/4.



21. Given a coordinates, and two points A and B. How many ways to go from A to B? You can only move up or right. For example, from (1, 1) to (5, 7), one possible way is 1,1 -> 2, 1… 5, 1 -> 5,2 -> ..5, 7.

A:

22. 9个不同礼物，分给4个小孩，每个小孩至少要分到一个礼物，问有多少种分法？

A: Assume the answer is f(9,4), then, we have

f(9,4) = 4^9 - c(4,1)×f(9,3) - c(4,2) ×f(9,2) - c(4,3)×f(9,1), where f(9,1)=1.

23. 两个人扔dice, A开始扔,谁扔到6,谁赢,问A赢得概率.

A: P = = =

(sum of infinite geometric series)

24. Monty Hall problem: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

A: Don’t switch – 1/3 to win; switch – 2/3 to win.

**Statistics**

Knowledge

# 1. Central Limit Theorem

# Suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution.

# 2. Confidence interval?

# The confidence level describes the uncertainty associated with a *sampling method*. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter. There are four steps to constructing a confidence interval.

* Identify a sample statistic. Choose the statistic (e.g. sample mean, sample proportion) that you will use to estimate a population parameter.
* Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.

Margin of error = Critical value \* Standard deviation of statistic   
Margin of error = Critical value \* Standard error of statistic

For guidance, see how to compute the margin of error.

* Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.

Confidence interval = sample statistic + Margin of error

# 3. Critical Value

A: The **critical value** is a factor used to compute the margin of error. This section describes how to find the critical value, when the sampling distribution of the statistic is normal or nearly normal.

The central limit theorem states that the sampling distribution of a statistic will be normal or nearly normal, if any of the following conditions apply.

* The population distribution is normal.
* The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
* The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
* The sample size is greater than 40, without outliers.

When one of these conditions is satisfied, the critical value can be expressed as a t score or as a z score. To find the critical value, follow these steps.

* Compute alpha (α): α = 1 - (confidence level/100)
* Find the critical probability (p\*): p\* = 1 - α/2
* To express the critical value as a z score, find the z score having a cumulative probability equal to the critical probability (p\*).
* To express the critical value as a t score, follow these steps.
  + Find the degrees of freedom (DF). When estimating a mean score or a proportion from a single sample, DF is equal to the sample size minus one. For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.
  + The critical t score (t\*) is the t score having degrees of freedom equal to DF and a cumulative probability equal to the critical probability (p\*).

Should you express the critical value as a t score or as a z score? There are several ways to answer this question. As a practical matter, when the sample size is large (greater than 40), it doesn't make much difference. Both approaches yield similar results. Strictly speaking, when the population standard deviation is unknown or when the sample size is small, the t score is preferred. Nevertheless, many introductory statistics texts use the z score exclusively. On this web site, we provide sample problems that illustrate both approaches.

4. *p*-value

The *p*-value or calculated probability is the estimated probability of rejecting the null hypothesis (H0) of a study question when that null hypothesis is true. In other words, the *p*-value may be considered the probability of finding the observed, or more extreme, results when the null hypothesis is true – the definition of ‘extreme’ depends on how the hypothesis is being tested.

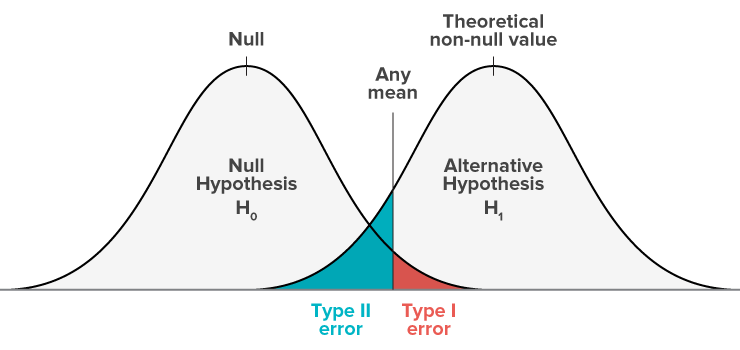
5. Independent and identically distributed

In probability theory and statistics, a sequence or other collection of random variables is independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent.

6. Type I and II error

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is the failure to reject a false null hypothesis (a "false negative"). More simply stated, a type I error is detecting an effect that is not present, while a type II error is failing to detect an effect that is present.

The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true. It is denoted by the Greek letter *α* (alpha) and is also called the alpha level. The rate of the type II error is denoted by the Greek letter *β* (beta) and related to the power of a test (which equals 1−*β*).



7. Power

The power or sensitivity of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true. It can be equivalently thought of as the probability of correctly accepting the alternative hypothesis (H1) when it is true – that is, the ability of a test to detect an effect, if the effect actually exists.

8. Sensitivity & specificity

Sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate.

*TPR* = *TP* / *P* = *TP* / (*TP* + *FN*)

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate.

*SPC* = *TN* / N = *TN* / (*TN* + *FP*)

9. 68–95–99.7 rule

In statistics, the so-called 68–95–99.7 rule is a shorthand used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of one, two and three standard deviations, respectively; more accurately, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.

10. Student's *t*-distribution

In probability and statistics, Student's *t*-distribution (or simply the *t*-distribution) is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. Whereas a normal distribution describes a full population, *t*-distributions describe samples drawn from a full population; accordingly, the *t*-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution.

11. Chi-Square test.

<http://stattrek.com/chi-square-test/independence.aspx>

Interview Question

1. A field with unknown number of rabbits. Catch 100 rabbits and put a label on each of them. A few days later, catch 300 rabbits and found 60 with labels. Estimate how many rabbits are there?

2. Given 10 coins with 1 unfair coin and 9 fair coins. The unfair coin has & #8532; prob. to be head. Now random select 1 coin and throw it 3 times. You observe head, head, tail. What’s the probability that the selected coin is the unfair one?

A: Bayes formula.

3.一个store有10排队，每个队只有一个收银员，然后每个队伍有10个人在排队。另一个store只有1排队，有10个收银员，有100人在排队，问你会选择哪个队伍来排，假设每个收银员办理一个客户的时间和variance一样。 那两个Store的expected waiting time是一样的 但是variance不一样 主要是分析mean-variance

A?: CLT说n个independent variable的variance converge to 1/n\*单个variable, 那10列variance小？

4. 如何detect outliers，以及用box plot的时候，为啥用h = 1.5IQR，系数为啥是1.5？

5. 如果data不服从正态分布怎么办？

A?: Transformation (Box-Cox).

6. 如果去一个卖油画的店子，里面有100幅油画，你买10幅，然后店主会以100幅油画的平均价作为每幅画的单价卖给你，一路浏览过去，决定买，或是不买，前提是你是行家，你知道每幅画值多少，如何能让你的利益最大化。

A?: 你不知道average price，所以我的回答是通过前五十幅画算出mean，和confidenceinterval，来选择后面的50幅画。

7: Power analysis and sample size calculation

A: <http://www.statsoft.com/textbook/power-analysis>

|  |  |  |  |
| --- | --- | --- | --- |
|  | | State of the World | |
| H0 | Ha |
| Decision | H0 | Correct Acceptance | Type II Error β |
| Ha | Type I Error α | Correct Rejection |

8. 44.3% vs 47.2% is it significant?

# 9. 几个random variable的相加相乘，得到的新rv的mean, variance，等

# E(X + Y) = E(X) + E(Y) and E(X - Y) = E(X) - E(Y)

# independent: E(XY) = E(X)E(Y)

# Var(aX + bY) = a2Var(X) + b2Var(Y) + 2abCov(X, Y)

# independent: Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)

# Var(XY)=E(X2Y2)−E(XY)2=var(X)var(Y)+var(X)E(Y)2+var(Y)E(X)2

# Var(X) = E[(X-E[X])2] = E[X2] – E[X]2

10. Tossing a coin ten times resulted in 8 heads and 2 tails. How would you analyze whether a coin is fair? What is the p-value? Is there any other method to test the fairness? In addition, more coins are added to this experiment. Now you have 10 coins. You toss each coin 10 times (100 tosses in total) and observe results. Would you modify your approach to the way you test the fairness of coins? (Google, others)

A: Set up a Hypothesis testing. H0: μ = 0.5; H1: μ <> 0.5.

p-value = P(0, 1,2, 8, 9 or 10 heads) = 112/1024 ≈ 0.109

1. If we have more (i.e. 10) coins, because each coin is independent, so I can judge the fairness of each coin use the same method above.

2. Assuming all ten coins are of the same mint, we find P(average of head-counts for the 10 coins is extreme). To do that we need to know the distribution of this average. It can be calculated explicitly with a bit of probability. In fact, the sum of (independent) Binomials with the same probability of success is another Binomial.

11. Given a lot of data, I want to random sample 1% of them. How to do it efficiently?

A: ?

12. 0到1之间任意取一个点是uniform distribution，那么任意取两个点，他们之间的距离是什么分布？

A?: Beta(1,2), 可以generalize到n个点，X(n) - X(1) ~ Beta(n-1, 2).

13. 怎么从uniform成生norm distribution?

A: Box-Muller transform.

**Generalized Linear Model**

Knowledge

1. Model selection:

If you have good *a priori* reasons to believe a variable should be in a model then simply include it, unless the evidence against it is very strong. If the main point of a model is prediction, you might not care too much about which independent variables are included, as long as the model “fits well”. But if the purpose of your model is to see which variables are important, then much attention needs to be paid to model selection.

2. OLS assumptions:

a. Correct specification – linear functional form is correctly specified

b. Strict exogeneity –

c. No linear dependence (among ***X***)

d. Spherical errors –

i. homoscedasticity –

ii. no autocorrelation –

e. Normality –

3. AIC vs BIC

AIC = 2k – 2ln(L)

BIC = kln(n) – 2ln(L)

AIC tries to select the model that most adequately describes an unknown, high dimensional reality. This means that reality is never in the set of candidate models that are being considered. On the contrary, BIC tries to find the TRUE model among the set of candidates. I find it quite odd the assumption that reality is instantiated in one of the model that the researchers built along the way. This is a real issue for BIC.

Nevertheless, there are a lot of researchers who say BIC is better than AIC, using model recovery simulations as an argument. These simulations consist of generating data from models A and B, and then fitting both datasets with the two models. Overfitting occurs when the wrong model fits the data better than the generating. The point of these simulations is to see how well AIC and BIC correct these overfits. Usually, the results point to the fact that AIC is too liberal and still frequently prefers a more complex, wrong model over a simpler, true model. At first glance these simulations seem to be really good arguments, but the problem with them is that they are meaningless for AIC. As I said before, AIC does not consider that any of the candidate models being tested is actually true. According to AIC, all models are approximations to reality, and reality should never have a low dimensionality. At least lower than some of the candidate models.

Interview Question

1. What is GLM?

2. What is the purpose of link function in GLM? (Amazon)

A: A.J. Dobson pointed out the following things in [her book](http://rads.stackoverflow.com/amzn/click/1584889500):

1. Linear regression assumes that the response variable is normally distributed. Generalized linear models can have response variables with distributions other than the Normal distribution– they may even be categorical rather than continuous. Thus they may not range from −∞ to +∞.
2. Relationship between the response and explanatory variables need not be of the simple linear form.

This is why we need the link function as a component of the generalized linear model. It links the mean of the dependent variable *Yi*, which is *E*(*Yi*)=*μi* to the linear term *xTiβ* in such a way that the range of the non-linearly transformed mean *g*(*μi*) ranges from −∞ to +∞. Thus you can actually form a linear equation *g*(*μi*)=*xTiβ* and use an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters.

3. Which regression methods are you familiar? How to evaluate regression result?

4. How do you evaluate regression? For example, in this particular case:

item click-through-rate  predicted rate

1       0.04        0.06

2       0.68        0.78

3       0.27        0.19

4       0.52        0.57

…

5. What to do with nonlinear regression effects?

A: Generalized additive models (GAM) or polynomial regression.

6. What is the risk of polynomial regression?

A: A drawback of polynomial bases is that the basis functions are "non-local", meaning that the fitted value of *y* at a given value *x* = *x*0 depends strongly on data values with *x* far from *x*0. In modern statistics, polynomial basis-functions are used along with new basis functions, such as splines, radial basis functions, and wavelets. These families of basis functions offer a more parsimonious fit for many types of data.

7. How to select order for polynomial regression?

A: Cross validation or RSS(m)/(n – m – 1) is a minimum or when there is no significant decrease in its value as the degree of polynomial is increased. RSS(m) is residual sum of squares for the mth order polynomial; n is number of data points; m is order of polynomial (so m+1 is the number of constants of the model).

8. 如果用手头上的50个sample得出来了模型，怎么能确定这个模型能generalized to all the population

A?: Validate with new data, compare with theoretical/empirical/simulation results, or carry out cross validation.

9. 5个变量的模型和50个变量的模型都很fit，通过一些指标，你知道50个变量的模型给的结果更好，你会选择用哪个模型？

A: Prefer model with 5 variables if the result is acceptable.

10. What is the consequence of multicollinearity?

A: 1. computer algorithm may not be able to get an approximate inverse of XTX, or it may be numerically inaccurate

2. estimate of regression coefficient is less precise

3. redundant -> overfitting

11. 以每周为一个时间段，每小时取一个data， 一共取50周。 如果用每周data的mean作为observation来做model，会有什么问题？

A: 1. narrow CI for new predictions

2. time series

10. Regression toward the mean

A: <http://en.wikipedia.org/wiki/Regression_toward_the_mean>

11. What criteria do you typically use to judge the goodness of fit of a linear model?

A: Coefficient of determination, denoted *R*2, is a number that indicates how well data fit a statistical model. (SSTO = SSR + SSE, R2 = 1 – SSE/SSTO)

Lack-of-fit sum of squares, is one of the components of a partition of the sum of squares in an analysis of variance, used in the numerator in an F-test of the null hypothesis that says that a proposed model fits well. (SSLF = SSE – SSPE)

12. What is regularization in regression? Why do regularization? How to do regularization?

A: In statistics and machine learning, regularization methods are used for model selection, in particular to prevent overfitting by penalizing models with extreme parameter values. The most common variants in machine learning are *L*1 and *L*2 regularization, which can be added to learning algorithms that minimize a loss function E(*X*, *Y*) by instead minimizing E(*X*, *Y*) + α‖*w*‖, where *w* is the model's weight vector, ‖·‖ is either the *L*1 norm or the squared *L*2 norm, and α is a free parameter that needs to be tuned empirically (typically by cross-validation; see hyperparameter optimization). This method applies to many models. When applied in linear regression, the resulting models are termed ridge regression or lasso, but regularization is also employed in (binary and multiclass) logistic regression, neural nets, support vector machines, conditional random fields and some matrix decomposition methods. *L*2 regularization may also be called "weight decay", in particular in the setting of neural nets.

13. When will a linear regression outperforms a logistic regression given the response variable is categorical (binary case or not?)?

A?: It depends on the underlying error distribution. Linear regression will beat logistic regression if the error distribution is normal and the response variable is just categorized from continuous variable.

14. How to convert a variable to dummy variable? What to do if there are a lot dummy conversions?

A: ???

15. A new search algorithm was developed for Youtube search. Customer feedback for how the new method is compared to the original one on particular searches are collected. It is in the form of “very good”, “good”, “same”, “bad”, and “very bad”. How to determine if the new one is better or not? For each search, there is also an internal score assigned to either new or old method. If we build a linear model which has the scores (new/old method) as independent variables and the feedback as dependent variable, what’s the potential problem with the model?

A: 1. Assign scores to the satisfaction level (vg – 2, g – 1, s – 0, b – -1, vb – -2). Conduct a hypothesis test to see if the mean is equal to 0 or not.

2. Independent variable is categorical. It would be better to use multinomial logistic regression or other classification methods to build the model.

**Logistic Regression**

Knowledge

1. Formula:

or

2. Odds, log odds and odds ratio:

odds = , log odds = , odds ratio =

odds = , log odds = , odds ratio =

Interview Question

1. What is logistic regression?

2. How to estimate the coefficients?

A: MLE.

3. How to calculate MLE?

A: Newton-Raphson, IRLS (solve score equation).

The idea indeed is MLE. Method for solving Logistic Regression (independent responses) IS MLE, and the problem is solving the score equation. When you use Fisher Scoring (close to Newton-Raphson except that Fisher Scoring uses expected second derivative instead of real second derivative), it is Equivalent to IRLS in some form. So IRLS is a way of solving the MLE score equation. Refer to Nelder, Wedderburn 1972. However, when responses are not independent, MLE is not available and you'll need to use GEE.

4. Regression Model diagnostics:

Issues with using linear regression instead of logistic regression on binary response variable?

1. It doesn’t make sense to model Y as a linear function of the parameters because Y has only two values. You just can’t make a line out of that (at least not one that fits the data well).

2. The predicted values can be any positive or negative number, not just 0 or 1.

3. The values of 0 and 1 are arbitrary. The important part is not to predict the numerical value of Y, but the probability that success or failure occurs, and the extent to which that probability depends on the predictor variables.

Logistic Regression uses a different method for estimating the parameters, which gives better results–better meaning unbiased, with lower variances.

5. How to choose link function in logistic regression? (Amazon)

A: Logistic has slightly flatter tails, i.e. probit curve approaches the axes more quickly than logit curve. Logit has better interpretation than probit. Logistic regression can be interpreted as modeling log odds. Usually people start the modeling with logit. You could use likelihood value to decide logit or probit.

Logit function:

* Leads to simpler mathematics due to complexity of the standard normal CDF
* It is easier to interpret (log odds)

Probit / Cloglog:

* Theoretical considerations
* Influences by disciplinary tradition (economists favor probit ; toxicologists favor logit)
* Underlying characteristics of the data (cloglog works best with extremely skewed distributions)

6. Why do we take all the trouble doing the transformation from probability to log odds?

A: It is usually difficult to model a variable which has restricted range, such as probability. This transformation is an attempt to get around the restricted range problem. It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity. Another reason is that among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret.

**Time Series**

Knowledge

1. Time Series Decomposition:

We shall think of the time series *yt* as comprising three components: a seasonal component, a trend-cycle component (containing both trend and cycle), and a remainder component (containing anything else in the time series). For example, if we assume an additive model, then we can write

*yt* = *St* + *Tt* + *Et*,

where *yt* is the data at period *t*, *St* is the seasonal component at period *t*, *Tt* is the trend-cycle component at period *t* and *Et* is the remainder (or irregular or error) component at period *t*. Alternatively, a multiplicative model would be written as

*yt* = *St* × *Tt* × *Et*.

The additive model is most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend-cycle does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative model is more appropriate. With economic time series, multiplicative models are common.

An alternative to using a multiplicative model is to first transform the data until the variation in the series appears to be stable over time, and then use an additive model. When a log transformation has been used, this is equivalent to using a multiplicative decomposition because

*yt* = *St* × *Tt* × *Et* is equivalent to log*yt* = log*St* + log*Tt* + log*Et*.

Sometimes, the trend-cycle component is simply called the “trend” component, even though it may contain cyclic behavior as well.

Interview Question

1. Additive or multiplicative model for time series analysis? (Uber)

A: The additive model is most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend-cycle does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative model is more appropriate. (See more details in Time Series Knowledge section.)

**Data Mining**

Knowledge

Interview Question

1. Pros/cons of decision tree and logistic regression. (Amazon)

A: Logistic regression:

* Lots of ways to regularize the model
* Easily update the model to take in new data
* Low variance and less prone to overfitting
* Assumptions

Decision tree:

* Easy to interpret and explain (maybe)
* Non-parametric (don’t need to worry about outliers or whether the data is linearly separable)
* Easily overfit

2. What is Gaussian mixture model? What is the latent variable when using EM algorithm?

A: Multiple Gaussian distributions represent different clusters. Latent variable is the cluster assignment.

3. What’s the difference between classification and regression?

A: The main difference between the classification tree and the regression tree is their dependent variable. For the classification tree, the dependent variables are categorical, while the regression tree has numerical dependent variables. Those of the classification tree also have a set amount of unordered values, while those of the regression tree have either discrete yet ordered values or indiscrete values. A regression tree is constructed with the purpose of fitting a regression system to each determinant branch in a way that the expected output value comes up. On the other hand, a classification tree branches out as determined by a dependent variable derived from the previous node.

4. How to evaluate classification result? What if the results are in probability mode?

5. If I want to build a classifier, but the data is very unbalanced. I have a few positive samples but a lot of negative samples. What should I do?

**Decision Trees / Random Forests**

Knowledge

Interview Question

1. Describe how decision tree works. How to build a decision tree from data? (Amazon)

A: Tree-like model, separate into two or more sets by some variables.

Stopping criteria:

* Number of cases in the node is less than some pre-specified limit.
* Purity of the node is more than some pre-specified limit.
* Depth of the node is more than some pre-specified limit.
* Predictor values for all records are identical - in which no rule could be generated to split them.

A related problem is deciding on the optimum tree size. While stopping criteria are a relatively crude method of stopping tree growth, early studies showed that they tended to degrade the tree's performance. An alternative approach to stopping growth is to allow the tree to grow and then prune it back to an optimum size.

2. What is “random”? What are “forests”? (Uber)

A: Random - Random forests use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. Typically, for a dataset with features, features are used in each split. Also, the training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.

Forests - Random forests operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set.

3. The importance() function in RandomForest package gives two values for each variable: %IncMSE and IncNodePurity. Is there simple interpretations for these 2 values? (Uber)

A: The first one can be 'interpreted' as follows: if a predictor is important in your current model, then assigning other values for that predictor randomly but 'realistically' (i.e.: permuting this predictor's values over your dataset), should have a negative influence on prediction, i.e.: using the same model to predict from data that is the same except for the one variable, should give worse predictions. So, you take a predictive measure (MSE) with the original dataset and then with the 'permuted' dataset, and you compare them somehow. One way, particularly since we expect the original MSE to always be smaller, the difference can be taken. Finally, for making the values comparable over variables, these are scaled. For the second one: at each split, you can calculate how much this split reduces node impurity (for regression trees, indeed, the difference between RSS before and after the split). This is summed over all splits for that variable, over all trees.

**Naïve Bayes**

Knowledge

Interview Question

1. What is Naïve Bayes and Bayes formula?

2. What’s the formula for Naive Bayesian classifier? What’s the assumption in the formula? What kind of data is Naive Bayesian good at? What is not?

**Analytics**

Knowledge

Interview Question

1. Given Amazon data, how to predict which users are going to be top shoppers in this holiday season.

2. We want to add a new feature to our product. How to determine if people like it? A/B testing. How to do A/B testing? How many ways? Pros and cons?

3. 5MM sellers are using Selling on Amazon (SOA). Find the best targets for FBA (a new type of services for seller other than SOA).

4. Given the customer demographic data. Can you arrive at the probability of finding a house owner in the low income group? How would estimate your confidence in the above probability, i.e. how would you arrive at the 95% confidence intervals? Suppose you are given the customer demographic data and whether they paid the loan successfully or not. Can you use the data and decide if future customers should be given loan or not? How would you check the performance of the model?

5. How to determine if one advertisement is performing better than the other?

6. Credit risk analysis – how to identify suspicious fraud?

7. Amazon sells a lot items. If the supplier for one item changes the price, how to decide if accept the change or not?

A:

**Experiment Design**

Knowledge

Interview Question

1. Design an experiment to test whether 5% increase in inventory volume is beneficial.

A: Sampling and fixed effect model.

2. T-test may not be powerful enough. Any way to improve?

A: Fixed effect model.

3. Bank marketing – two bonus offers (A & B) applied in two cities respectively, how to compare which is better?

**Business**

Knowledge

Interview Question

1. What’s the difference between forecast demand and forecast sales?

A: Need historical data for sales. Sales can somehow represent demand if not out of stock. Demand is also affected by pricing and competition. If the good is out of stock, sales will be 0 while demand is not related. In this case, demand can be predicted by other metrics, such as website traffic or stats collected from test market or surveys. Sales forecast could be done via time series analysis.

2. 一个仓库, 可以sell product in regular price,也可以sale in lower price,怎么决定多少产品sell,多少产品sale之类的.

A?: 这个应该是Revenue Management中Capacity-Controlled Discount的Model吧？用两个price算出最优预留capacity对应的ratio再根据demand的分布得到最优capacity…

**R**

**SQL**

Knowledge

Interview Question

1. Difference between where and having?

2. What are the differences among row\_number, rank and dense\_rank? (Amazon)

A: ROW\_NUMBER: Returns the sequential number of a row within a partition of a result set, starting at 1 for the first row in each partition. (MS)

RANK1: As an Aggregate function, it returns the rank of a row within a group of rows.

RANK2: As an Analytic function, it returns the rank of each row of a query with respective to the other rows. (Oracle, same as ROW\_NUMBER in MS)

DENSE\_RANK: Similar as RANK, but only consecutive values.

**SAS**

**Behavior**

1. How did you collaborate?

2. Example of innovation.

3. How did you learn something new?

4. How to tell if a relationship is correlation or causation?