QUANT BIBLE

MIT Sloan Business Club

Contents

1 Introduction

I started this guide sometime in my junior fall during an interview season for quantitative finance that I found super-challenging. As interviews wrapped up, I thought it would be a good idea to really go back to the basics and examine the fundamentals of what goes into interviewing for quant. That idea ended up turning into a question bank, and then a long write-up spanning a lot of the core concepts that apply to quantitative finance. Here, there's sections on probability and statistics, data science and regressions, quant research cases (with contributions from Kyri Chen), market making (written by Ravi Raghavan, Guang Cui, and Evan Vogelbaum), and an expansive question bank (with contributions by Evan and Ravi).

One big reason I made this guide was to democratize quantitative finance as a career for the SBC community. Quant finance definitely has a reputation as the kind of industry that's only for geniuses, that you can only break into if you're part of the intellectual "in"-group. As a result, this ends up being true to some extent as a self-fulfilling prophecy. In my opinion, though, as long as you build up a good amount of familiarity with the math and CS concepts behind quant, and you have a lot of energy and enthusiasm toward the field, quant finance is definitely within your reach. MIT is a great place to start building a path toward quant finance because not only is MIT one of the main colleges that quant recruits from, but there is a clear courseroad to building quant-related technical knowledge from math and CS courses here.

There's a list of these classes a bit later on in this intro; I recommend using this bible as a supplement as you go through the course-road in your semesters at MIT, and then as a primary piece of reading material as you finish quant-related coursework and dive more directly into quant interview prep (around sophomore spring or summer, probably).

Quant finance is more than an industry for big math brains to work in secrecy and make boatloads of money in; I think that for the right person, it's one of the most intellectually stimulating and exciting fields of work out there. Quant finance is a great intersection of math, computer science, and economics, in that you get to use the advanced concepts you learn in MIT math and CS classes, almost on the same level as a UROP student or SWE/ML engineer, but in the context of solving financial problems and puzzles. In a quant job, the financial markets become a prime playing ground for MIT technical knowledge. They're a microcosm of literally everything that happens in current events and the real world, boiled down to numbers and data, and they change and evolve literally every day and every hour, forcing you to constantly adapt, stay informed, and learn new, cutting-edge technical skills to keep up with the industry. The kind of creative problem-solving that goes into quantitative finance work has led to many innovations over the past few decades; for example, the McDonald's chicken nugget only exists today because Ray Dalio, founder of Bridgewater, helped McDonald's and their suppliers develop a new synthetic future for chicken that would protect them against risk and make chicken nuggets a viable product (`https://www.cnbc.com/2018/05/03/how-ray-dalio-helped-launch-mcdonalds-chicken-mcnugget.html`). In addition, quant finance can be a highly ethical way of working in the finance industry. The culture of philanthropy at places such as Jane Street is strong, and since quant finance pays so well, you can devote a large portion of your paycheck towards philanthropic causes, making quant one of the most money-efficient ways to use your earnings potential as an MIT grad towards social good.

If quant finance interests you in any way, feel free to delve into this bible; I hope you get something interesting or useful out of it!

1.1 List of Places to Apply

- Jane Street
  - Quant Trading, Quantitative Research, SWE
- Citadel & Citadel Securities
  - Citadel Securities - Systematic Trading (more CS) or Semi-Systematic Trading (more math), Fundamental Analyst
  - Citadel - Trading (Global Fixed Income, etc), Quantitative Research, SWE
- The D. E. Shaw Group
  - Prop Trading, Quantitative Analyst, SWE/Quant Developer
- Two Sigma
  - Quantitative Research, SWE
- Hudson River Trading
  - Algo Developer, SWE
- Jump Trading
- SIG
- Optiver
- Akuna
- J.P. Morgan
  - Quantitative Research Extern/Intern for MIT
- Bridgewater
- QuantCo
- DRW
- IMC Trading
- Five Rings
- AQR
- Virtu Financial
- Tower Research
- Seven Eight Capital
- TransMarket Group
- Wolverine Trading
- Old Mission Capital
- Point72 (Cubist)
- Belvedere Trading
- Group One
- Flow Traders

1.2 Other Resources

There's a very wide variety of resources you can use to prepare for quantitative finance interviews, from MIT classes to interview prep books to online listings and forums. In addition, quant interviews can cover a really wide range of topics and pretty much anything in the realm of math-y problem solving, probability and combinatorics, CS/SWE concepts, data science, even machine learning, etc, can be asked. It can definitely get pretty daunting to prepare for this kind of interview, and that's why most of these internships are intended for summer after junior year and maybe sophomore year. Especially if you didn't do a lot of math/CS extracurriculars in high school, it makes sense to spend a few semesters diving into these topics throughout your academic college career to build the foundation and intuition necessary for these interviews. For that reason, I'll start off with some of the prime MIT classes for quant-related material:

Core Classes

- 18.600 (Probability and Random Variables)
- 18.06 (Linear Algebra)
- 14.32 (Econometrics)
- 6.042 (Discrete Math)
- 6.006/6.046 (Algorithms stuff)
- 6.034/6.036 (Machine Learning)
- 18.650 (Statistics)

Extra Classes

- 18.615 (Stochastic Processes, basically the main field of mathematical finance)
- 6.867 (Graduate-level Machine Learning)
- 6.437/6.438 (Inference, basically the advanced theory behind stats/data science/machine learning)
- 18.211 (Combinatorics, advanced version)

If you get through these classes and really enjoy the material then quantitative finance makes sense as a career option. It's important to feel fluent in and have a deep understanding of things from these classes. In quant interviews, you'll encounter a lot of the same concepts from these classes but recontextualized for quant finance instead, so you'll want to be able to stay on your feet and apply what you've learned in some unfamiliar, out-of-comfort-zone ways. To really hone your ability to tackle these brainteasers, CS questions, and "case studies," you can explore learning resources for quant-related topics outside of MIT classes:

Books

- *Thinking Fast and Slow* by Daniel Kahneman (a more fun, relaxed read on the psychology of how we think, relevant to trader thinking styles)
- *Heard on the Street* by Timothy Crack (one of the main books for finance interview practice in general)
- *Elements of Statistical Learning* by Trevor Hastie, etc. (essential data science/quant research book)
- *Quant Job Interview Questions and Answers* by Mark Joshi
- *A Practical Guide to Quantitative Finance Interviews* by Xinfeng Zhou
- *Fifty Challenging Problems in Probability with Solutions* by Frederick Mosteller
- *Cracking the Coding Interview* (quant jobs are increasingly placing emphasis on data structures, algorithms, etc. so this is important)

Extra Books

- *Art of Problem Solving - Intro to Counting and Probability and Intermediate Counting and Probability* (these are some of the main books for high school math competition prep on these topics)
- *Option and Volatility Pricing* by Natenburg (important options book in the quant industry; some places such as Optiver teach directly from this book)
- *Options, Futures, and Other Derivatives* by John Hull

Websites

- Glassdoor. Look up individual companies and internships and you'll find question postings by past interviewers.
- The Puzzle Toad: `https://www.cs.cmu.edu/puzzle/`
- Wall Street Oasis. Some quant interview help, but also general advice and discussion about finance careers.
- LeetCode. Some trader/QR roles will give coding challenges (Two Sigma, HRT, Akuna, Belvedere).
- Kaggle. Popular site for data science projects/discussion and good place to familiarize yourself with numpy/pandas/scipy.

2 PROBABILITY FUNDAMENTALS

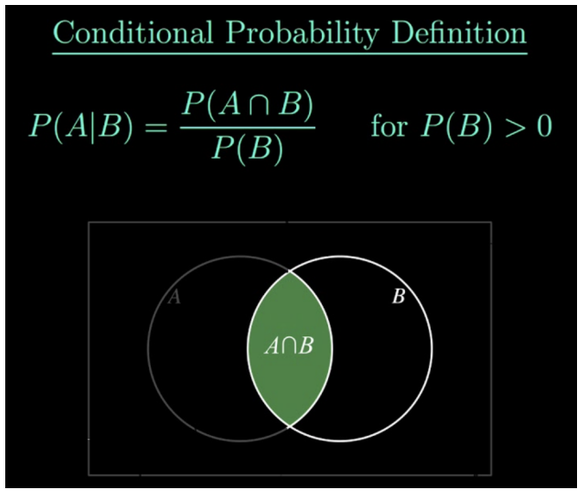
This section is an overview of all of 18.600, with most of the focus on random variables and probability distributions; I'll cover a lot of combinatorics material in the combinatorics section so this section will gloss over those aspects of 18.600 more.

2.1 Conditional Probability and Bayes' Theorem

- Quant firms care a lot about your understanding of conditional probability. In general, many real-life probabilistic events we can think of are dependent on each other (the chance someone is coughing today vs. the chance that person is sick today, etc.); for two dependent events A and B, the chance of A occurring given that B has occurred is written as the conditional probability $P(A|B)$, the probability of A given B. This conditional probability has the definitional formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

  The following diagram illustrates this:



  So we can think of the probability of A given B as the ratio of the probability of A and B occurring together vs. the probability of just B happening. In the Venn diagram above, $P(A|B)$ equals the fraction of the probability space for B that is taken up by the intersection space of A and B.

- We notice that the term $P(A \cap B)$ gives us a symmetry for conditional probabilities, i.e. we can write $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$. From this we get Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  This can be seen as a simple rewriting of the definitional formula, where we substitute $P(A \cap B)$ for the conditional probability in the other direction. Bayes' formula is useful because all three terms $(P(B|A), P(A), P(B))$ are often easily computable in real-world scenarios.

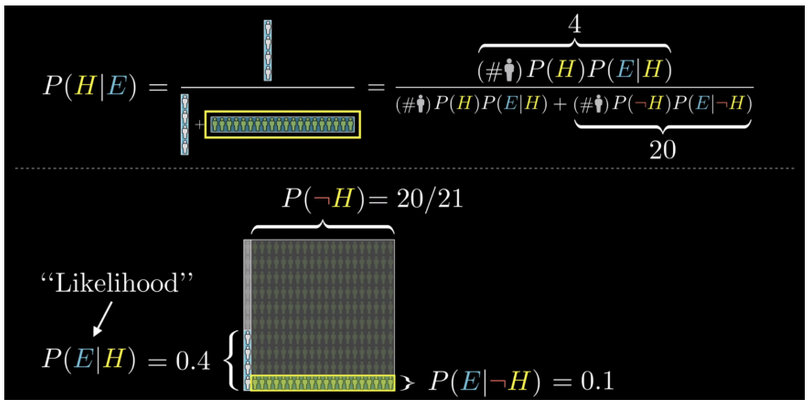
  - The term $P(B|A)$ is known as the likelihood.
  - The term $P(A)$ is known as the prior, i.e. the probability of A "prior" to new evidence (which would be B) being collected. Likewise $P(B)$ is known as the evidence.
  - The evidence term in Bayes' theorem is often calculated with the law of total probability using A and its complement $\neg A$, i.e. we can write $P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$.

- Tversky and Kahneman (famous pioneers of behavioral economics) formulated some classic brainteasers for wrapping your head around Bayes' theorem.

  - Imagine you are a member of a jury judging a hit-and-run driving case. A taxi hit a pedestrian one night and fled the scene. The entire case against the taxi company rests on the evidence of one witness, an elderly man who saw the accident from his window some distance away. He says that he saw the pedestrian struck by a blue taxi. In trying to establish her case, the lawyer for the injured pedestrian establishes the following facts. There are only two taxi companies in town, "Blue Cabs" and "Green Cabs." On the night in question, 85 percent of all taxis on the road were green and 15 percent were blue. The witness has undergone an extensive vision test under conditions similar to those on the night in question, and has demonstrated that he can successfully distinguish a blue taxi from a green taxi 80 percent of the time. What is the probability that the taxi the old man saw was actually blue?
    - Most people immediately answer that the taxi was significantly more likely to actually be blue, because of the old man's 80% accuracy rate. However, let B = taxi was blue, O = old man saw blue, G = taxi was green; Bayes' theorem gives

$$P(B|O) = \frac{P(O|B)P(B)}{P(O|B)P(B) + P(O|G)P(G)} = \frac{0.8 * 0.15}{0.8 * 0.15 + 0.2 * 0.85} \approx 0.41$$

- Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail. Is Steve more likely to be a librarian or a farmer?
  - Most people immediately answer that Steve is more likely to be a librarian than a farmer, since the personality description seems to fit much more closely with a librarian than a farmer. However, librarian is relatively a much rarer occupation, and we can estimate that there are 20x more farmers than librarians in the U.S. Even if this personality description fits, say, 40% of all librarians vs. just 10% of all farmers, Bayes' theorem gives a much greater likelihood that Steve is a farmer, as illustrated below.



- More conditional probability examples
  - Suppose 1% of people in the U.S. have Ebola. There is a test for Ebola that has a 1% false positive and 1% false negative rate, i.e. 99% of healthy people will test negative and 99% of sick people will test positive. What is the probability that a person who tested positive actually has Ebola?
    - Let H = healthy, S = sick, $+$ = tested positive. Bayes' theorem gives

$$P(S|+) = \frac{P(+|S)P(S)}{P(+|S)P(S) + P(+|H)P(H)} = \frac{0.99 * 0.01}{0.99 * 0.01 + 0.01 * 0.99} = 0.5$$

      so this test is actually pretty inaccurate. (Side note: does the effective accuracy of this test improve a lot from repeated trials, i.e. higher $P(S|+)$ if $+$ represents multiple positive test results in a row? Bayes' theorem shows that for $+ = k$ positive tests in a row, the effective accuracy becomes

$$P(S|+) = \frac{0.99^k * 0.01}{0.99^k * 0.01 + 0.01^k * 0.99}$$

      which equals 99% accuracy for even just 2 positive tests.)
  - This question is asked a lot in trading interviews, especially as part of an earlier phone screen. Suppose we have 1000 coins; 999 are fair coins and the 1000th has heads on both sides. We pick a random coin and flip it 10 times, and it lands heads all 10 times. What is the probability that we picked the unfair coin?
    - Let $10H$ = coin lands heads 10 times, UF = coin is the unfair one, F = coin is fair. Bayes' theorem gives

$$P(UF|10H) = \frac{P(10H|UF)P(UF)}{P(10H|UF)P(UF) + P(10H|F)P(F)} = \frac{1}{1 + \frac{999}{1024}} \approx 0.5$$

2.2 Expected Value and Variance

- Any random variable has a probability mass function (if discrete) or probability distribution function (if continuous); write these as $p(x)$ for the p.m.f. or $f(x)$ for the p.d.f., respectively.
- One of the most important properties of an r.v. is its expected value. Intuitively, this is the value we expect the random variable to take on any arbitrary polling of its outcome, and more formally, it is the weighted average of the values it can take, weighted by the probability of taking each value. Therefore the expected value is the same idea as the mean of a random variable. We can write the expected value as a sum or integral for the p.m.f. or p.d.f., respectively.

$$E[X] = \mu = \sum_{x \in \Omega} xp(x) \quad or \quad \int_{\Omega} xf(x)dx$$

  where Ω represents the sample space of the random variable and μ is used to denote the mean.

  - We can also think of the expected value of any function of a random variable in the same way. This would be calculated as the weighted average of the function values that the r.v. can take, weighted by the probability of taking each value. In other words, it is

$$E[g(x)] = \sum_{x \in \Omega} g(x)p(x) \quad or \quad \int_{\Omega} g(x)f(x)dx$$

    The above formulas, or even just the idea of taking a weighted average, can be very useful for quant finance interviews; you will sometimes run into questions about calculating the expected value of some more esoteric random value, and that will just come down to specifying the p.m.f. or p.d.f. and doing the weighted average or integral.

- Linearity of expectation. The expected value is a linear function so we have

$$E[aX + b] = aE[X] + b.$$

  The most important part is that the linearity works for combinations (sums) of random variables, even if the random variables are dependent:

$$E[X_1 + X_2 + ...X_n] = E[X_1] + E[X_2] + ... + E[X_n]$$

  even if $X_1, ..., X_n$ are dependent.

  - This linearity of expectation property for dependent variables is very important for quant interviews; a lot of seemingly complicated questions about expected values for some probabilistic experiment/situation can be solved very easily by setting up possibly dependent random variables for the experiment in a clever way and applying linearity of expectation on them. The Five Rings tournament question is a good example of this. Another simpler example:
  - We have a classroom of 10 boys and 10 girls, and we arrange them randomly into a single-file line of 20 students. What is the expected number of pairs of adjacent students who are different genders?
    - Let X_i be the indicator for whether the i-th and i+1-th student in the line are different gender (i.e. 1 if yes, 0 if no). The chance of any arbitrary pair in the line having different gender is $\frac{2*10*10}{20*19} = \frac{10}{19}$ so $E[X_i] = \frac{10}{19}$ for any i. We might notice that each of the X_i are pairwise dependent, since whether any one pair is different gender affects the remaining amounts of boys and girls that can be arranged elsewhere in the line. However, we can still use linearity of expectation, so the answer is $E[X_1 + ... + X_{19}] = E[X_1] + ... + E[X_{19}] = 19 * \frac{10}{19} = 10$.
- The variance of a random variable describes how much it deviates from its expected value on average. It can therefore be written as the expected value of the square of the difference between an r.v. and its mean:

$$Var(X) = E[(x - \mu)^2]$$

  - The variance also has an alternate form that comes directly from applying linearity of expectation to the above:

$$Var(X) = E[X^2] - E[X]^2$$

  - Variance of linear combination:

$$Var(aX + b) = a^2 Var(X) + b$$

$$Var(X + Y) = Var(X) + Var(Y)$$

    It's often important to know variance for the sum or average of i.i.d. random variables, i.e. random variables with the same p.m.f. or p.d.f. that are sampled independently. If $X_1, ..., X_n$ are i.i.d., each with variance σ^2, then

$$Var(sum(X_1, ..., X_n)) = n\sigma^2$$

$$Var(average(X_1, ..., X_n)) = \frac{\sigma^2}{n}$$

2.3 Random Variables

This subsection gives information on the most important classical types of random variables.

We also need to define a cumulative distribution function (cdf) for continuous r.v.s, as the probability that the r.v. takes a value less than the function input:

$$F_X(a) = P\{X < a\} = \int_{-\infty}^{a} f(x)dx$$

Below are several examples of random variables.

Discrete Random Variables

Random Variable	Experiment	PMF	Expected Value	Variance
Bernoulli	Experiment with two outcomes, 1 if yes, 0 if no; for example, a (fair or unfair) coin flip	$p_X(x) = p$ if $x = 1$, $q = 1 - p$ if $x = 0$, where parameter p = probability of success	$E[X] = p$	$Var(X) = pq$
Binomial	Experiment with n independent Bernoulli trials, count the number of successful trials.	$p_X(x) = \binom{n}{x} p^x q^{n-x}$ where parameter n = number of trials, p = probability of success of each trial.	$E[X] = np$	$Var(X) = npq$
Poisson	Experiment counting the number of occurrences of an independent event in a fixed time or space interval	$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ where parameter λ = the known mean of number of events in fixed time or space interval	$E[X] = \lambda$	$Var(X) = \lambda$
Geometric	Experiment of successive independent Bernoulli trials, count how many trials until and including first success	$p_X(x) = (1-p)^{x-1}p$ where parameter p = probability of success of each trial.	$E[X] = \frac{1}{p}$	$Var(X) = \frac{1-p}{p^2}$

Continuous Random Variables

Random Variable	Experiment	PDF	Expected Value	Variance
Uniform	Draw number in interval $[a, b]$ with equal chance for any number in interval	$f_X(x) = \frac{1}{b-a}$ if $a \le x \le b$, 0 otherwise	$E[X] = \frac{a+b}{2}$	$Var(X) = \frac{(b-a)^2}{12}$
Normal	The classic bell curve; average of asymptotically many trials of the same experiment converge to a normal r.v. under the central limit theorem.	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\{\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}}$	$E[X] = \mu$	$Var(X) = \sigma^2$
Exponential	Experiment measuring events following a Poisson distribution; measure the time from now until the first event	$f_X(x) = \lambda e^{-\lambda x}$ where parameter λ is the rate (number of events on average in one unit of time	$E[X] = \frac{1}{\lambda}$	$Var(X) = \frac{1}{\lambda^2}$

Important properties:

- A Poisson experiment comes from a binomial experiment where asymptotically n is very large and p is very small, so that $np = \lambda$.
- We can construct a Poisson point process $N(t)$ = the number of events that occur during the first t units of time. $N(t)$ is constructed by a sequence of exponential r.v.s with the same rate λ.
- Both the geometric and exponential random variables are memoryless, i.e. the probability distribution of geometric or exponential X after some trials/time has already elapsed is the same as if starting over at the first trial/at time 0. In other words, the behavior of experiments following the geometric or exponential distribution is not affected by how many trials/time has already passed.

2.4 Distributions of Functions and Joint Distributions

- If we know the pdf of a random variable X, we can compute the pdf of any strictly increasing function of X. Integrate the pdf to obtain the cdf $F_X(a)$. Let $g(x)$ be any strictly increasing function of x. Then for $Y = g(X)$, we have $F_Y(a) = F_X(g^{-1}(a))$. This gives us the cdf of Y and we can take the derivative to obtain the pdf.
- Any pair of discrete or continuous random variables can have a joint probability mass function or distribution function:

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$
$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$
$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

  From the joint distribution, we can obtain marginal distributions, which are just the probability distributions for a single one of the r.v.s (the other can take any value):

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$
$$f_X(x) = \int_y f_{X,Y}(x, y)$$
$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y)$$

  - When X and Y are independent, the joint distribution is the product of the marginal distributions, i.e.

$$p_{X,Y}(x, y) = p_X(x) p_Y(y)$$
$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

2.5 Covariance and Correlation

Covariance and correlation both describe the degree to which a pair of random variables can vary in similar and dependent ways.

- Formula for covariance of two random variables X and Y:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

  This is an "expectation of product minus product of expectations", with several useful properties:

  - If X and Y are independent, then $Cov(X, Y) = 0$. The converse is not true, however.
  - $Cov(X, X) = Var(X)$
  - Bilinearity. $Cov(aX_1 + bX_2, Y) = aCov(X_1, Y) + bCov(X_2, Y)$.
- Correlation is a scale-independent version of covariance, scaled down to between -1 and 1. Its formula is:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

  - If two random variables are independent then they are uncorrelated, but the converse of this is not true.

3 STATS FUNDAMENTALS

This section is an overview of the first third of 18.650.

3.1 LLN and CLT

- Statistics vs. probability:
  - Probability encompasses simpler problems where we can start with the initial parameters and models, then analyze the proceeding outcomes and data.
  - Statistics encompasses complex problems about randomness where our underlying parameters/distribution are unknown; we collect data and deduce the parameters through quantitative techniques.
- Main framework for statistical modeling:
  - Treat each data event as a random variable. We make assumptions for what kind of r.v. describes the data event, i.e. Bernoulli, uniform, exponential, Poisson, etc. Some additional common assumptions are independence of each data event as well as that each data event is described by the same random variable/underlying distribution; together, these assumptions are called "i.i.d" ("independent and identically distributed")
  - Formulate a link between the underlying parameter you want to estimate vs. your data event random variables. Is your desired parameter equal to the average of your random variables, or some function of the average, or something else? This function of the data is the "estimator".
    - Example. If our data events are Bernoulli, then the average $X_n = avg(X_1 + ... + X_n)$ tends to the expected value $E(X) = p$. So to estimate the unknown parameter p for a series of Bernoulli trials using an estimator $\hat{p}$, we can set $\hat{p} = X_n$.
  - Estimate your level of confidence in how your estimator predicts the underlying parameter. If your estimator is p= 0.55, are you 95% confident that your actual parameter is between 0.5 to 0.6? 0.45 to 0.65? 0.54 to 0.56?
- The law of large numbers creates this link between theoretical parameters and empirical data:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P},\ \text{a.s.}} \mu.$$

- When the mean is a function of an unknown parameter of the random variable/underlying distribution then the LLN becomes very useful, and indeed this is the case for all common r.v. types: uniform, Bernoulli, Poisson, geometric, exponential, etc.
- The central limit theorem helps us quantify our level of confidence in our estimation (through a confidence interval):

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

3.2 Confidence Intervals

- The confidence interval for some estimator in a statistical model tells us what range of values we believe the true parameter may lie in, as well as the chance that the true parameter actually lies in this range.
  - "95% confidence interval, 99%, etc." → there is a 95%/99%/etc. chance that the true parameter lies within the bounds of the CI.
  - The center of the interval often comes from the LLN and is equal to the estimator; the range of the interval comes from the CLT.
- Definition: A confidence interval of level 1-α for a parameter is an interval I where

$$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

- Starting point: the CLT tells us that if $q/2$ is the (1-$\alpha/2$) quantile of $N(0,1)$, then with probability 1-α, we have the asymptotic interval for the true parameter:

$$\theta = [\hat{\theta} - \frac{\sigma}{\sqrt{n}} q_{\alpha/2},\ \hat{\theta} + \frac{\sigma}{\sqrt{n}} q_{\alpha/2}]$$

- Clearly we need to have our confidence interval be independent from the true parameter; otherwise we don't actually know anything meaningfully new about the true parameter from the confidence interval. Unfortunately, σ depends on the true parameter, so we need to find techniques for getting rid of the variance.
- Finishing our confidence interval
  - We have a special case when the random variable is Bernoulli, i.e. $\sigma = \sqrt{p(1-p)}$. Then we have an elementary bound $p(1-p) \leq \frac{1}{4}$ so the confidence interval becomes

$$\theta = [\hat{\theta} - \frac{1}{2\sqrt{n}} q_{\alpha/2},\ \hat{\theta} + \frac{1}{2\sqrt{n}} q_{\alpha/2}]$$

  - Generally we use Slutsky's theorem which allows us to add and multiply limits in the LLN; since the variance is a function of the true parameter, we just substitute this our estimator in place of the true parameter in the variance formula, then plug that into the confidence interval.

4 QUANT RESEARCH - DATA SCIENCE

This section is an overview of the "main" chapters of Elements of Statistical Learning. The Elements of Statistical Learning (ESL) book is considered GOATed in the fields of data science, machine learning, and statistics, and is essentially the first book you'll want to consult for a comprehensive and rigorous yet concise overview of basics about regression, data modeling, and inference.

4.1 Least Squares and Nearest Neighbors

The linear model has been a mainstay of statistics for the past 30 years and remains one of our most important tools. Given a vector of inputs $X^T = (X_1, X_2, ..., X_p)$, we predict the output Y via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{p} X_i \hat{\beta}_i$$

In other words, Y is a linear combination of the input features X plus a bias term. Our goal is to fit the best set β of coefficients and bias. Y is often just a scalar (so β would be a vector), but Y can also be a vector, so that β would be a matrix. The equation above represents a hyperplane in the input-output space.

The most popular method for fitting a linear model is using least squares, i.e. picking the right hyperplane so that the sum of squares of distances of each input feature to the hyperplane is minimized across all possible hyperplanes. In other words, we are trying to minimize an objective function, the "residual sum of squares":

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

The residual sum of squares is a highly "natural" choice for the measure of error that we want to minimize for a model, and it's indeed used in many contexts besides linear regression. Due to some technical math, it actually turns out that simple linear regression, using residual sum of squares, for the linear model provides the best "unbiased" estimate among all linear models for the underlying conditional expectation of output values given input values. This conditional expectation is known as the conditional expectation function (CEF) and is the optimal theoretical predictor for our data problem from a Bayesian standpoint. It is for this reason that residual sum of squares is not only a "natural" choice of error function, but actually the mathematically optimal one in the linear context.

We can derive a simple formula for β for the linear model by taking the derivative of the RSS above; we'll do this in a few pages, and this β formula is one of the core formulas in data science and is very important to memorize by heart.

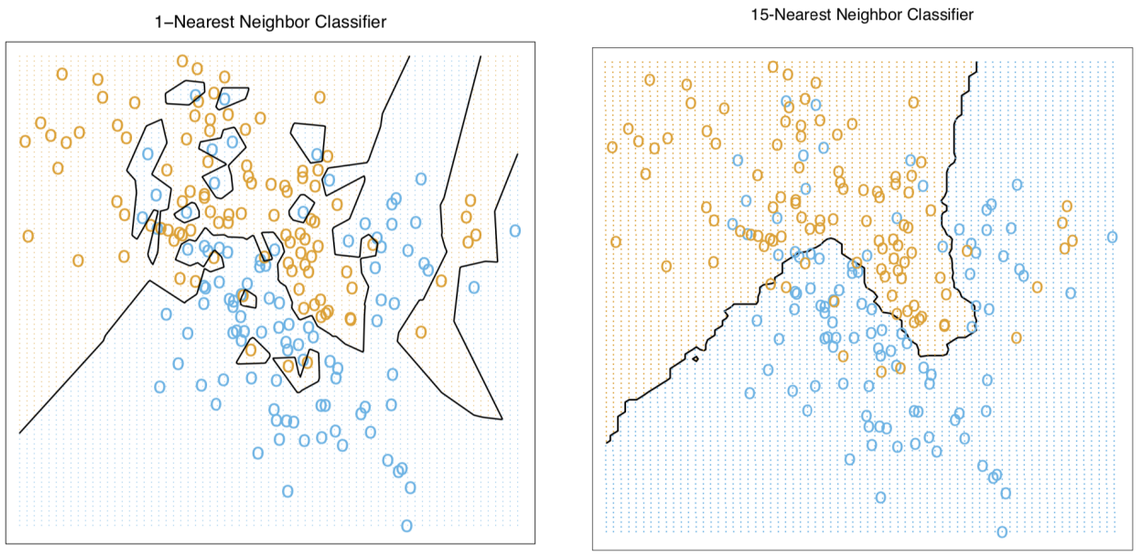
The other "simplest" model is nearest neighbors. Nearest-neighbor methods use those observations in the training set $\mathcal{T}$ closest in the input space to x to form $\hat{Y}$. Specifically, the k-nearest neighbor fit for $\hat{Y}$ is defined as follows:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

This formula essentially corresponds to taking the average of the k nearest points for x, which are symbolized by the $N_k(x)$ function. Note that nearest neighbors just involves direct calculations on the training data, and not fitting a model (except maybe the choice of k); we just have to memorize the training data and compute with it!

Decision regions and boundaries for nearest neighbors:

- For 1-nearest-neighbor, the decision boundaries/regions form a Voronoi tessellation which is easily computable, and often highly disjoint + irregular
- For highest k-nearest-neighbor, the decision regions generally get less disjoint but are still highly irregular, something that a linear model can't do
- Impt point: the "effective" number of parameters is n/k instead of just 1 (the single parameter k), because up to overlap, nearest neighbors is creating n/k different decision regions. This gives an intuitive explanation for why regions get less disjoint with higher k.



Kernel methods are augmentations of nearest neighbors that employ a varying weight, smoothly decreasing with distance between source and target, rather than a constant weight.

Comparing and contrasting the two approaches:

Linear model w/ least squares	Nearest neighbors
Low variance and high bias	High variance and low bias
Relies on assumption that linear models are the right choice for the data	Doesn't rely on assumptions about underlying data
Works well for "Scenario 1": Gaussian distributed data with uncorrelated components and different means	Works well for "Scenario 2": mixtures of Gaussian distributions, with mean of each component Gaussian in each mixture independently sampled
Efficient and accurate choice for high-dimensional data	Effective and indeed commonly used in practice when data is low-dimensional and plentiful, but suffers from the "curse of dimensionality".

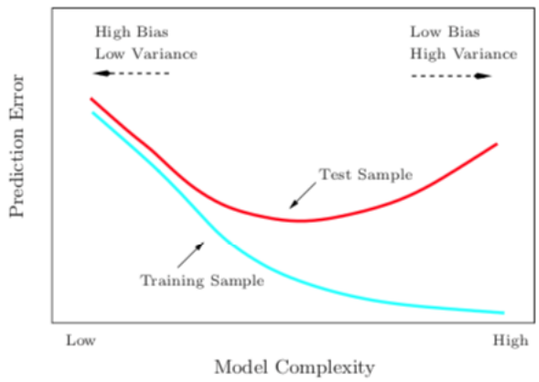
4.2 Intuition for Technical Details: Least Squares and Nearest Neighbors

There are some important technical details about the nearest neighbor and linear regression models that we will discuss here; we'll address each row of the above compare-and-contrast table, discussing their broader context and important intuitive ways of thinking about them.

First, the details about "Scenario 1" vs. "Scenario 2" are the easiest to address. When each class or target value of data follows a single Gaussian distribution, the data are more cleanly separable, i.e. a single line can effectively delineate the difference between one Gaussian and the next. This corresponds to "Scenario 1" and the use of linear models, which create these simpler delineations for prediction. On the other hand, when each class or target value of data follows a mixture of distributions that can intertwine and overlap with each other in more complex ways, the prediction must be done with a larger number of disjoint and often irregular decision regions. This corresponds to "Scenario 2" and the use of nearest neighbors, which naturally creates these disjoint regions when the linear model doesn't.

Bias-Variance Tradeoff

We expand on the bias-variance tradeoff for nearest neighbors vs. linear regression. In this context, and actually in general, we can think of bias as the degree of assumptions and constraints inherent in the choice of model; this intuition is a bit different from the formal definition of bias as the expected difference between the model's prediction vs. the actual output for an arbitrary data point. On the other hand, we can think of variance as the resulting "instability" of the model, or its degree of sensitivity to changes in the input data; fortunately this intuition pretty closely captures the formal definition of variance. These intuitive viewpoints of bias and variance allow us to think of bias as the property we directly control, via the assumptions and constraints we bake into the design and tuning of our model, and variance as the property we observe as an output result.



Any model inevitably faces a bias-variance tradeoff, in which model design and tuning that decreases bias results in increased variance, and vice versa. The relative levels of this tradeoff for nearest neighbors vs. linear regression are then closely linked to the second row of our table, which mentions the assumptions that each model makes. Nearest neighbor models are some of the lowest-bias models possible, and in fact 1-nearest neighbor is THE lowest, by exactly memorizing the training data and creating the most complex and numerous disjoint regions of classification. 1-nearest neighbor makes no assumptions about the data, no matter what the data is. One illuminating aspect of 1-nearest neighbor is that it is the only model that always perfectly classifies the training data, i.e. there is zero error; any other model will have some assumptions or generality that causes it to misclassify or mispredict at least some points in the training data. The other side of the coin is that 1-nearest neighbor displays extremely high variance; we usually make significant changes to the classification regions if we add or perturb even a single data point.

Now pivoting to linear regression, we mentioned that linear regression assumes a linear relationship is the right underlying approach for modeling the data. This assumption is not as strict as it sounds, for various reasons; for example, we can make transformations to the features such as introducing new parameters that polynomial or trigonometric functions of the original parameters, so that nonlinear relationships are captured. Even so, this lenient assumption increases the bias and decreases the variance of linear regression relative to nearest neighbors, which has virtually no assumptions (or for 1-nearest neighbors, actually no assumptions). We'll see later on that linear regression is still highly unbiased compared to most other models, which have stricter assumptions.

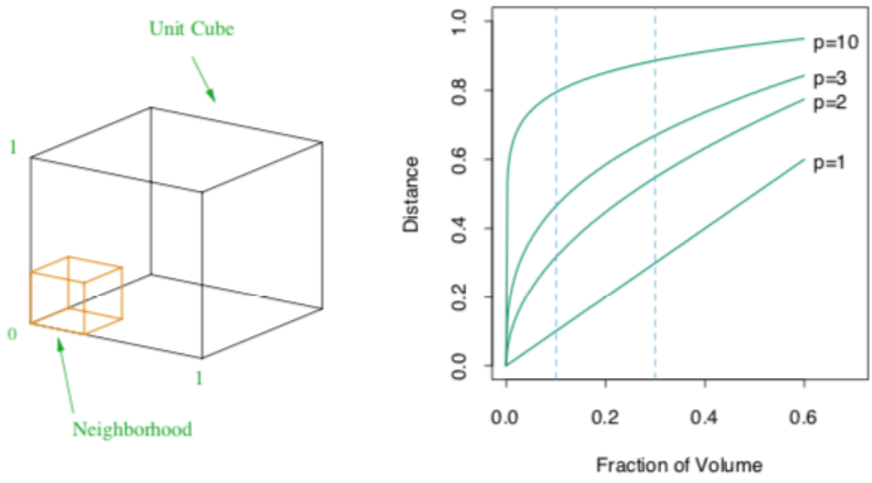
There are two other notes about bias-variance important to note:

- Bias is also directly correlated with "model complexity," which is a more common consideration in data science and statistics generally than our intuitive perspective of bias from earlier. We can think of any data science/statistics problem as taking a dataset which has some set amount of information or "complexity" baked into it, and trying to accurately capture this complexity to make good predictions. Complexity comes from both the model itself and the assumptions behind the model, which in a sense are two independent sources of complexity. Make stricter assumptions, and complexity shifts towards the assumptions and away from the model; likewise, relax your assumptions and complexity shifts back towards the model. In this way model complexity is inversely related to the level of our underlying assumptions in designing the model, and the model complexity viewpoint of bias cleanly aligns with our intuitive viewpoint about model assumptions.
- Any model has an implicit property called (effective) degrees of freedom (denoted df), which is linked to the bias-variance tradeoff. With higher degrees of freedom, bias decreases and variance increases. This makes sense with our discussion; more degrees of freedom are granted when assumptions and constraints are relaxed, but more degrees of freedom also means more room for change in the model when the data changes. We noted earlier that n/k is the effective number of parameters of k-nearest neighbors, and it's also in fact the effective degrees of freedom.

Curse of Dimensionality

Nearest neighbors and linear models are better fits for two different domains of data modeling problems, respectively low-dimensional and high-dimensional data. This dichotomy is governed by the curse of dimensionality, which is a formal way of saying that inference becomes exponentially harder as the data becomes high-dimensional for certain classes of models. We can observe the curse of dimensionality for nearest neighbors in a few geometrically intuitive ways:

- In high dimensions, the distance between two arbitrary points in some region increases on average compared to low dimensions. Because nearest neighbors classifies new points according to some nearby point a small distance away, high dimension without a proportional increase in the number of training data points results in less availability of a nearby training point for an arbitrary input point, and therefore less accurate prediction.
- In high dimensions, it becomes exponentially more likely for data points to lie on the edge of the range of input values, and boundary points are harder to predict because they more often require extrapolating from one nearby training point rather than interpolating between training points.



When data is low-dimensional and plentiful, nearest neighbors dominates and is often used in the real world for applications where available data follows this paradigm. When data dimensionality increases even a moderate amount, nearest neighbors falls off very quickly. For example, we can imagine that 2-dimensional data is very easy for nearest neighbors to classify, while 10-dimensional data has already grown to be very inefficient for nearest neighbors.

For moderate to high dimensionality, linear regression methods dominate over nearest neighbors. The assumption of linear underlying relationship provides some defense for linear models against the curse of dimensionality compared to nearest neighbor models. In general, simple linear regression is still highly impacted by the curse of dimensionality when compared to many other models; however, as we'll see soon, there are various additional assumptions and variations for linear regression we can make that strengthen linear regression in high dimensionality.

4.3 Regressions

This section will pivot away from the more general nearest neighbors vs. linear model discussion as we go into the deep technical details of linear regression.

We'll start by introducing the closed form for the beta in simple linear regression. This is the core regression equation you should know off the top of your head!

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

We can also form a "hat matrix" by rewriting the equation in terms of the set of fitted values $\hat{y}$:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

so we have a "hat matrix" $H = X(X^T X)^{-1} X$ that puts the hat on y. This hat matrix is just the beta with an extra factor X on the left and with the y removed, but it also has a more natural meaning that comes from thinking of regression as a projection of y onto X. The beta that minimizes the RSS also minimizes the distance of this projection, and the hat matrix is an operator that computes the orthogonal projection of y onto X.

BTW, the derivation of the beta closed form looks like this:

$$RSS = (y - \beta X)^T (y - \beta X)$$
$$\frac{\partial RSS}{\partial \beta} = -2X^T (y - \beta X) = 0.$$

It might happen that the columns of X are not linearly independent, so that X is not of full rank. This would occur, for example, if two of the inputs were perfectly correlated, (e.g. $x_2 = 3x_1$). Then $X^T X$ is singular and the least squares coefficients $\hat{\beta}$ are not uniquely defined. However, the fitted values $\hat{y} = X\hat{\beta}$ are still the projection of y onto the column space of X; there is just more than one way to express that projection in terms of the column vectors of X. The non-full-rank case occurs most often when one or more qualitative inputs are coded in a redundant fashion.

Sample Variance
Equation for sample variance:

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where N is the number of data points, and p is the number of inputs for each point (i.e. dimensionality). If we make some reasonable assumptions that the different outputs $\hat{y}$ are uncorrelated and share the same variance, then we find that the variance of the beta is actually equal to $(X^T X)^{-1}\sigma^2$ where σ^2 is the sample variance; this variance is distributed according to a chi-squared distribution with $N-p-1$ degrees of freedom (the same value as the denominator of our sample variance!), which gives us the t-test below.

Z-Score for Individual Beta Coefficients (T-Test)
To test the hypothesis that a particular coefficient $\beta_j = 0$, we form the standardized coefficient or "Z-score"

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where v_j is the jth diagonal element of $(X^T X)^{-1}$. Under the null hypothesis that $\beta_j = 0$, z_j is distributed as t_{N-p-1} (a t distribution with $N - p - 1$ degrees of freedom), and hence a large (absolute) value of z_j will lead to rejection of this null hypothesis.

In other words, the z-score for individual beta coefficients based on t distribution tells us how dramatically a model would change if the coefficient was removed; higher z-scores (absolute value) tell us that that coefficient is more important, and z-scores beyond a threshold (say, 2) mean that these coefficients and corresponding predictors are significant. We might choose to discard predictors whose coefficients do not reach this threshold.

The z-score for t-test also gives us confidence intervals for each individual beta coefficient:

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma},\ \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma})$$

Here, $z^{(1-\alpha)}$ is the $1 - \alpha$ percentile of the normal distribution:

$$z^{(1-0.025)} = 1.96$$
$$z^{(1-0.05)} = 1.645, etc.$$

Hence the standard practice of reporting $\hat{\beta} \pm 2 \cdot se(\hat{\beta})$ amounts to an approximate 95% confidence interval. Even if the Gaussian error assumption does not hold, this interval will be approximately correct, with its coverage approaching $1 - 2\alpha$ as the sample size $N \to \infty$.

Z-Score for Groups of Beta Coefficients (F-Statistic)
Instead of just testing for the exclusion of single coefficients like in the t-test, we may want to test groups of coefficients at once. Testing singletons vs. groups is actually two highly distinct tasks; with a lot of possible configurations for correlations between variables, it may very well be the case that, for example, there are three coefficients that do not meet the Z-score significance threshold, but a group of those three coefficients IS significant, a property that can only be detected with the F-statistic and not the t-test. The F-statistic looks like:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}.$$

Here RSS_1 is the original RSS, and RSS_0 is the RSS for the regression with all coefficients in our group set to 0. One common use of the F-statistic is to verify the significance of coefficients together after t-tests for all single coefficients have been calculated. For example, if we want to drop every non-significant coefficient, i.e. with $|zscore| < 2$, we might group all the non-significant coefficients for an F-statistic and verify whether the F-statistic is still non-significant, helping us make a final decision on whether to drop the entire group or keep some terms.

Multivariate Regression from Univariate

Algorithm 3.1 *Regression by Successive Orthogonalization.*

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
2. For $j = 1, 2, \ldots, p$

   Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, , \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \ldots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate $\hat{\beta}_p$.

The result of this algorithm is

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}. \tag{3.28}$$

In other words, when we regress on one variable and take the residuals, the set of residuals is orthogonal with respect to the original variable; then we can use the residuals as the new inputs for the successive orthonormalization.

This is actually the standard way multiple linear regression is done. The whole point of multiple linear regression is for the fitting of our model to be independent for each variable, i.e. for the beta coefficient corresponding to each single variable in the multiple regression to represent that variable's effect on the output adjusted for the effects between that variable and all other variables in the multiple regression. By orthonormalizing and regressing on the residual after each single variable step in the multiple regression, we ensure this is the case. Another way of saying this is that the multiple regression coefficient β_j represents the additional contribution of x_j only after x_j has been adjusted for the p other input variables in the multiple regression, $x_0, x_1, ..., x_{j-1}, x_{j+1}, ..., x_p$.

4.4 Dimensionality Reduction

One common goal in linear regression is dimensionality reduction, where we restrict or even vanish some of the original variables' coefficients in order to narrow down a smaller subset of important input variables with smaller dimension than the original. There are two main justifications for dimensionality reduction:

- Improving test error. One important detail of the bias-variance tradeoff is that total error is generally given by the sum variance + bias^2. We mentioned earlier that, although simple linear regression is more biased and lower-variance than nearest neighbors, it's still highly unbiased relative to many other models beyond nearest neighbors. Simple linear regression is often in the region of bias vs. variance where introducing further assumptions and restrictions will actually decrease variance more than it increases bias, resulting in improved overall error.
- We care about interpretability of the regression, and part of interpretability is pinpointing what subset of variables are the highly significant portion of the regression.

Stepwise Regressions
There are three stepwise regressions for performing subset selection (finding subset of size k of the parameters that creates the best possible model out of all such subsets). Each one takes a slightly different approach but is usually built on the regression through orthonormalization algorithm from earlier.

- Forward stepwise
  - Start with empty model, no parameters
  - At each step, find the one parameter that creates the model of best fit (vanilla univariate regression), then add that parameter to your total model and orthonormalize everything with respect to that parameter. Repeat this for further steps.
- Backward stepwise
  - Start with full model, incorporating all parameters
  - At each step, find the one parameter that contributes least to the fit, then remove it, then orthonormalize with respect to the removed parameter. Repeat this for further steps.
- Forward stagewise
  - Start with zero-initialized model
  - At each step, find the parameter most correlated w the current residual, then compute the beta coefficient for this parameter and add it to the corresponding coefficient in the total model
  - Note: takes much longer than k steps for size-k subset, which is in contrast to the first two stepwises taking just k steps, but this actually pays off with greater accuracy/effectiveness in very high dimension.

One other technique is good to know: for relatively small total dimension, i.e. 30 or less, the "leaps-and-bounds" algorithm is a highly efficient method for this.

Ridge Regression

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Note that $X^T X + \lambda I$ is always nonsingular, guaranteeing a unique solution. This nice closed form for the ridge beta (very similar to the beta for simple linear regression, just with the λI term added) was actually the original motivation for ridge regression, since $X^T X + \lambda I$ is always invertible.

One way of seeing how ridge regression increases bias is by observing its effect on the effective degrees of freedom df. It can be calculated that $df(\lambda)$ follows a monotone decreasing function, where $df(0) = p$, where p is the number of parameters, and $df(\infty) \to 0$. Increasing the regularization weight λ reduces our degrees of freedom and likewise increases bias, both results of imposing additional restriction.

More technical aspect of ridge regression: it is actually an implementation of "principal components analysis (PCA)," which aims to reduce dimensionality by transforming to new features and truncating to the first k most important new features.

PCA relies on the singular value decomposition (SVD) of a matrix X, i.e. decomposing it into $X = UDV^T$ where U and V are unitary and D is diagonal with real nonnegative entries $d_1, d_2, ...$, in decreasing order as you go down the diagonal. This is also called "eigen-decomposition".

The matrices in SVD also give a formula:

$$X^T X = V D^2 V^T$$

The SVD gives the building blocks for a transformation very similar to a "change of basis"; the columns of V are eigenvectors where each one generates a linear combination of the original vectors which is orthogonal to all the previous vectors in the sequence. These transformed vectors are the "principal components" z:

$$z_1 = Xv_1$$

and the corresponding d_i's (diagonal values of D) correspond to a "scaling" (actually sample variance) for the principal components, so that the first principal has the highest sample variance (the biggest impact on the model) and this impact decreases in order so the last principal components has the smallest sample variance and the smallest impact.

Ridge regression shrinks the last principal components.

Lasso Regression

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

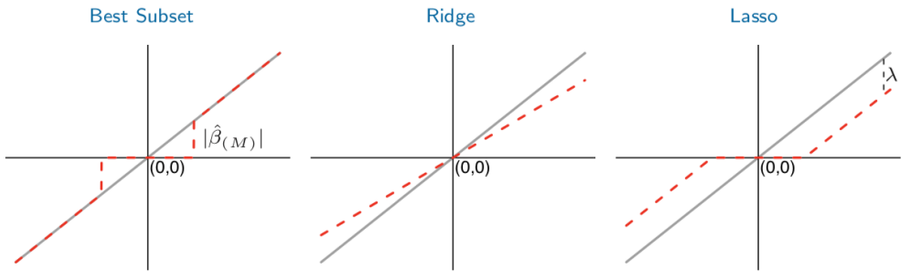
Note the similarity to the ridge regression problem; the L2 ridge penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ is replaced by the L1 lasso penalty $\lambda \sum_{j=1}^{p} |\beta_j|$. This latter constraint makes the solutions nonlinear in the y_i, and there is no closed form expression as in ridge regression.

This is a more subtle optimization problem with different results compared to ridge. While ridge has the closed form with the new beta matrix equation, the lasso has no such closed form and is actually a quadratic programming problem. An interesting point is that, if we constraint the lasso term to be $< t$ for some t, then some beta coefficients vanish as t gets smaller before hitting 0, which doesn't happen for ridge. This creates an important difference in use cases between lasso and ridge; lasso can create sparse models by removing some variables entirely, while ridge compresses variables to nonzero values, so it doesn't induce the same sparsity.

Comparing and contrasting the various dimensionality reduction approaches:

Subset selection (stepwises)	Ridge	Lasso
"Hard thresholding", completely drops/truncates features beyond the desired subset size.	Proportional shrinkage, most important features are shrunk less than least important features.	"Soft thresholding", transforms each coefficient by the same constant factor and truncates at 0.

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\operatorname{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



It's also worth mentioning that lasso and ridge regression are the $q = 1$ and $q = 2$ cases, respectively, of regularization with the L_q norm of the beta. Any non-negative values of q are possible for L_q regularization, including $q = 0$ (which corresponds to variable subset selection) and even higher powers like $q = 3$ or $q = 4$. In practice it's common for an optimal or highly effective value of q to lie between 1 and 2. For this reason, "elastic-net" is often used as a weighted average between lasso and ridge, to efficiently approximate calculation in the $q \in (1, 2)$ regime. Elastic-net uses the following regularization term for regression:

$$\lambda \sum_{j=1}^{p} \big(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\big),$$

Least Angle Regression
Similar to the forward stepwise approach, but this time instead of adding whole parameters from the feature set to the model one by one, the coefficients in the model are continuously increased, with constant tracking of which parameters (can be multiple) have the highest correlation with the residuals and continuously increasing those.

- LAR only adds to the model the amount of coefficient that it "deserves".
- LAR is useful because it is a greedy algorithm → easily computable but it also produces a very similar result to lasso, basically identical until any coeff crosses 0.
- This leads to a "lasso modification" for LAR: by dropping any variable whose coefficient crosses 0 and recomputing the joint least squares distribution after dropping, we can obtain the entire lasso path.
- After k iterations of LAR, the fit has k degrees of freedom, super elegant

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

Principal Components Regression
Continuing from the principal components derivation from ridge regression, we regress y directly on the principal components, and after obtaining the beta coefficients for the principal components, we expand them out as linear combinations of the original features to obtain a regression on the original features.

The principal components are already orthonormal so regression ends up simplifying to univariate regressions on each.

We usually drop out a bunch of the last principal components, i.e. only regress y on the first M principal components where $M < p$. This is the biggest difference between PCR and ridge; PCR directly truncates the last principal components while ridge proportionally shrinks them from first to last but retains at least some to all principal components.

Note: Principal Components Analysis
The principal components process described above doesn't have to apply to just regressions. The step of transforming the input basis into a smaller-dimension set of principal components bases can be done on the input data points before fitting any model or even considering the target values of those points; this is a form of unsupervised learning and allows the dimensionality reduction advantage of principal components transformations to be applied in many contexts besides ridge regression or PCR.

4.5 Brainteasers about Regression

- What are some of the main assumptions of linear regression?
  - The underlying relationship between the input features and the target output is actually linear (this is the main distinction that facilitates using linear regression over the other "simple" method of k-nearest neighbors).
  - Building off of the first assumption, linear regression generally relies on the target having a linear relationship with the inputs plus some amount of error. We also assume that this error term is centered around 0 so that there is no bias. We can also assume the error term follows a normal distribution but this is not strictly necessary.
  - Little to no multicollinearity, i.e. the input features are not highly correlated with each other and are independent from each other. Of course, in practice multicollinearity can happen often so at the very least we want none of the inputs to be a perfect linear combination of the other inputs.
  - Homoskedasticity, which means that the errors should have the same variance at different values of any of the input.
  - No autocollinearity, i.e. the residuals for a single variable are independent of each other. This means that the error should be uncorrelated with any input variable, i.e. it should actually be an unpredictable random error.
- Suppose I have two inputs X and Y, and I do two different regressions, each for one input w.r.t the other. In other words, I regress $\hat{Y} = \alpha_1 + \beta_1 X$ and also $\hat{X} = \alpha_2 + \beta_2 Y$. Is it true that $\beta_1 = 1/\beta_2$?
  - No; the intuition is that the first regression is trying to minimize the sum of vertical residuals while the second regression is trying to minimize the sum of horizontal residuals, and there is no reason that these sums should lead to an exact match for the betas.
- What happens to the regression if the input features are not all linearly independent?
  - The solution to linear regression is not unique
- I perform a regression on three features that are highly correlated; two of the features fit well but the third has a high standard error. How do I deal with this?
  - This is a sign of multicollinearity, which is a challenging issue in practice that has a lot of possible remedies. The simplest way to deal with it is to drop the third variable entirely. However, the third variable can still carry valuable information not present in the other variable so this might be undesirable.
  - Other remedies to multicollinearly without dropping the variable include different types of transformations, such as trying different kinds of feature transformations like polynomials, centering the variables by subtracting the mean, or using linear combinations of the variables.
  - We can also try other types of regression more equipped to deal with multicollinearity. Any of the dimensionality reduction regressions, like ridge, lasso, or PCA work well.
- Suppose I double my data points, i.e. repeat each point once; what happens to my regression?
  - The betas stay the same. To see this analytically, we can either observe the closed form $\beta = (X^T X)^{-1} X^T y$ and see that doubling the points leads to "doubling" the matrices with duplicate values so that the closed form is the same; we can also look at $\beta = Cov(X, Y)/Var(X)$ and argue both terms don't change.
- Between lasso and ridge regression, which exhibits lower bias/higher variance?
  - Intuitively we can compare the shapes of the L_q norm for lasso vs. ridge. The L_2 norm expands further outwards, i.e. takes higher values, than the L_1 norm, so that imposing a penalty on the L_2 norm is a stricter constraint than the same weight penalty on the L_1 norm. Using our model assumption perspective for bias, this implies that lasso regression should have the lower bias and higher variance, and indeed this is usually the case in practice.
  - One case where the variance difference between lasso and ridge is highly apparent is with a group of variables that are each insignificant according to their z-scores and that each have similar effects on the output. Because lasso soft thresholds by truncating some of the least significant terms, a small perturbation of the data can shift the order of significant terms and change the choice of truncated terms in lasso, which makes the lasso regression more unstable in this case. On the other hand, ridge regression performs proportional shrinkage on the insignificant terms and this shrinkage does not change much if the order of significance shifts, so ridge regression is more stable in this case.

4.6 The Econometrics Perspective

Randomized Trials

- The goal of econometrics is to deduce cause and effect relationships between different aspects of society and economics; we have access to many data science tools for doing so.
- An example of an econometrics problem that highlights some of the biggest initial roadblocks:
  - We have data samples for health insurance status, health, and other attributes for many different people in a population. We want to deduce whether greater levels of health insurance are a cause of better health outcomes.
  - Ideally we can prove causality if we have "other things equal" (ceteris paribus). However, this almost never happens and we have many other variables, known and unknown, that vary with the input and/or output variable we care about.
  - For the health insurance problem, we can measure other factors such as education level, income/wealth, age, etc. We find that all of these factors, and probably many more, also correlate with health insurance status and/or health in some way, so we don't have other things equal.
- Without other things equal, we run into selection bias with calculating the average causal effect. Mathematical explanation of selection bias:
  - Suppose we have two input possibilities, 0 = not on health insurance, 1 = on health insurance. Let Y_i be the health outcome for person i, and further define Y_{0i} as the outcome if person i is not on health insurance and Y_{1i} as the outcome if person i is on health insurance.
  - Let D be our input for health insurance, i.e. $D = 0$ corresponds to no health insurance and $D = 1$ corresponds to health insurance. Our average causal effect is

    $$Avg_n[Y_{1i} - Y_{0i}] = Avg_n[Y_i|D = 1] - Avg_n[Y_i|D = 0] = Avg_n[Y_{1i}|D = 1] - Avg_n[Y_{0i}|D = 0]$$

    From the above equations, the selection bias qualitatively arises; we see the $D = 1$ dependent outcome only for the people in the $D = 1$ group, and we compare it to the $D = 0$ dependent outcome only for the people in the $D = 0$ group. If these groups are not the same in other things equal, then the calculated comparison between Y_{1i} and Y_{0i} is faulty.
  - To extract a selection bias term, write

    $$Avg_n[Y_{1i} - Y_{0i}] = \Big(Avg_n[Y_{1i}|D = 1] - Avg_n[Y_{0i}|D = 0]\Big) + \Big(Avg_n[Y_{0i}|D = 0] - Avg_n[Y_{0i}|D = 0]\Big)$$

    $$= (average\ causal\ effect) + (selection\ bias)$$

- Randomizing the assignment of people to groups greatly mitigates selection bias. This involves randomly assigning each person in our total population set to either receive treatment or control, i.e. the $D = 1$ and $D = 0$ groups.
  - LLN in this context tells us that sample averages tend towards underlying population expectations as our sample size gets larger, i.e. every average taken for two large randomized sample groups from the same population will tend to be equal.
  - LLN implies that our selection bias term $Avg_n[Y_{0i}|D = 0] - Avg_n[Y_{0i}|D = 0]$ will also become 0.
  - "Checking for balance": after performing the randomization, it is useful to measure various sample averages and check that they are actually roughly equal.

Statistics Fundamentals in Econometrics

- Refresher on important stats fundamentals:
  - The sample mean is unbiased:

    $$E[\bar{Y}] = E[Y_i]$$

  - Population variance:

    $$Var(Y_i) = \sigma_Y^2 = E[(Y_i - E[Y_i])^2]$$

  - Sample variance of Y_i in a sample of size n:

    $$S(Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2$$

  - Sampling variance:

    $$Var(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

    We can obtain an estimator for the sampling variance by plugging in the sample variance for Y_i:

    $$\hat{Var}(\bar{Y}) = \frac{S(Y_i)^2}{n}$$

$$\hat{SE}(\bar{Y}) = \frac{S(Y_i)}{\sqrt{n}}$$

- From the data we can calculate a t-statistic as the difference between sample mean and assumed population mean, scaled by standard error:

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}$$

- The CLT states that the distribution of $t(\mu)$ approaches the standard normal distribution as our sample size grows larger. This fact allows us to test for statistical significance in two directions. We can check whether the sample mean is statistically different from the population mean by checking whether $|t|$ is above some threshold value, often 2. We can also construct confidence intervals for what values of μ are reasonable for the data:

$$\mathcal{I} = [\bar{Y} - C \times \hat{SE}(\bar{Y}),\ \bar{Y} + C \times \hat{SE}(\bar{Y})]$$

where C is a similar threshold value to before, often $C = 2$.

- Testing two sample means: we have a treatment sample group with $\bar{Y}^1$ and μ^1, and we have a control sample group with $\bar{Y}^0$ and μ^0. Our goal is to test whether $\mu^1 = \mu^0$. Our estimated standard error looks like

$$\hat{SE}(Y^1 - Y^0) = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

where n_1 and n_0 are our sample sizes. Under the null hypothesis that $\mu^1 - \mu^0 = \mu$, the t-statistic looks like

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0 - \mu}{\hat{SE}(\bar{Y}^1 - \bar{Y}^0)}$$

and our confidence intervals are constructed similarly as before.

- Important note: a large t-statistic often comes simultaneously from a significant underlying effect and from a small standard error (large sample size). Therefore, a small t-statistic doesn't necessarily mean that the underlying effect is insignificant, but rather may reflect a lack of statistical precision (high sampling variance).

Regression in Econometrics

- Without random assignment, we instead use regression to deduce causal relationships; we make the assumption that when observed variables are the same across treatment and control groups, selection bias from unobserved variables is also mostly eliminated.
  - This is the "conditional independence assumption" (CIA), or "selection on observables," i.e. the idea that there is no selection bias in the average causal effect conditional on a given set of observables when comparing outcomes conditional on the same set of observables.
- Treatment and control variables are encoded numerically
  - Raw encoding (often for continuous or discrete variables)
  - One-hot or "dummy" encoding (often for categorical or "yes/no" variables)
- A regression model links the treatment variable to the outcome, and also holds control variables fixed by including them in the model. Let X be a vector representing our treatment, and let A be a vector (or matrix) representing the controls. Our model is

$$Y_i = \alpha + \beta X_i + \gamma A + \epsilon$$

The ϵ term is the residual, or the difference between the fitted output $\hat{Y} = \alpha + \beta X_i + \gamma A$ and the actual output Y. Regression fits the α, β, γ as to make the sum of squared residuals (RSS) the least possible; the resulting estimates are called ordinary least squares (OLS).

  - Regression is commonly performed in most econometrics studies as a benchmark against more advanced techniques.
  - Technical detail: under certain technical conditions, regression provides the most statistically precise estimates possible for average causal effects from a given sample.
  - Regressing on the log of the outcome variable is useful for obtaining estimates that can be interpreted as percentage changes.
- One main roadblock in regression is omitted variables bias (OVB); since regression only eliminates selection bias across observed variables that are encoded as controls in the model, an inadequate set of controls (i.e. we failed to include important non-treatment variables) can cause selection bias to persist in our regression.

- For a given control variable, its OVB effect can be quantified by performing two regressions, one with the control ("long") and one without ("short"), and measuring the difference between the betas as follows:

$$OVB = (Treatment\ effect\ in\ short) - (Treatment\ effect\ in\ long) = \beta^s - \beta^l$$

$$= (Relationship\ between\ omitted\ and\ treatment) \times (Omitted\ effect\ in\ long) = \pi_1 \times \gamma$$

  In the above, the β^s and β^l are the regression coefficients for the treatment variable in the short and long regressions, respectively; the γ is the regression coefficient of the omitted control in the long regression; and the π_1 is the regression coefficient of the omitted control on the treatment, i.e.

$$A_{omitted} = \pi_0 + \pi_1 X + \epsilon_{omitted}$$

  - Important quote about OVB: "The importance of the OVB formula stems from the fact that if you claim an absence of omitted variables bias, then typically you're also saying that the regression you've got is the one you want. And the regression you want usually has a causal interpretation. In other words, you're prepared to lean on the CIA for a causal interpretation of the long regression estimates." (excerpt from MHE)
  - The OVB formula can't generate exact quantities for omitted variables we have no data on, but we can perform qualitative reasoning with it to deduce whether the effects of omitted variables should be positive or negative. Example in case studies.

- Robustness of regression: our confidence in a regression model grows when the OVB effect is small for any variables besides a set of a few core control variables, i.e. the treatment effect is insensitive to outside omitted variables.

Technical Details of Regression

- Technical details of regression.
  - Regression finds the best possible fit to the unknown conditional expectation function (CEF) $E[Y_i|X_i]$; this is the exact match if the CEF is linear, and a close linear approximation if the CEF is nonlinear.
  - In the bivariate case, i.e. Y and X are single variables, regression is closely related to the covariance through direct formulas for the regression coefficients:

$$\beta = \frac{Cov(Y, X)}{Var(X)}$$

$$\alpha = E[Y] - \beta E[X]$$

  - Residuals have zero mean (expectation and sample mean) and are also uncorrelated with all input variables and fitted outputs.
  - "Regression anatomy": in a multivariate regression, the coefficient for one input variable X_1 comes from the residual of its regression on the other control variable X_2. In other words, for $Y = \alpha + \beta X_1 + \gamma X_2 + \epsilon$, we have

$$\beta = \frac{Cov(Y, \widetilde{X_1})}{Var(\widetilde{X_1})}$$

    where X_1 is the residual from input on control, i.e.

$$X_1 = \pi_0 + \pi_1 X_2 + \widetilde{X_1}$$

    This makes sense because it implies that the coefficient for any input in a multivariate regression depends on its residual (the leftover) after regressing on all other variables, i.e. the coefficient for any input encapsulates information only about itself.

    * Another way of stating this. If X_2 is uncorrelated to X_1 then we expect the beta in the "short" regression (excluding X_2) to be very close to the beta from the "long" regression including X_2, and the more correlated, the most the short and long betas will differ. (MHE)

  - Standard error for a regression can be expanded out as follows:

$$SE(\hat{\beta}) = \frac{\sigma_\epsilon}{\sqrt{n}} \times \frac{1}{\sigma_X}$$

    To minimize our standard error for the estimated coefficients, we want less variance in the residuals σ_ϵ, and/or more variance in the inputs σ_2. A high residual variance means the regression is a poorer fit, but high input variance is actually good.

  - Regression with the above standard error assumes homoskedasticity (one of the core assumptions of regression), which is that the variance of residuals is uncorrelated to inputs. If the model is heteroskedastic (variance of residuals IS correlated to inputs) then we need "robust standard error" (formula not shown here).

    - Homoskedasticity can be hard to attain. Whenever the CEF is nonlinear we have heteroskedasticity because the residual variance is proportional to the square of the gap between the regression line and the CEF. Many underlying linear CEFs also do not imply homoskedasticity, such as the linear probability model (LPM) which is a regression on a zero-one input.
    - Homoskedasticity is not a very strict requirement for regression to work and in practice, heteroskedasticity doesn't make that much of a difference (the robust standard error is usually close to the conventional (homoskedastic) standard error.

- Technical details about the CEF
  - CEF decomposition. The target is equal to the CEF plus a residual (written ϵ_i) that is uncorrelated to X_i and has property $E[\epsilon_i|X_i] = 0$ ("mean independent").

$$Y_i = E[Y_i|X_i] + \epsilon_i$$

  - The CEF is the minimum mean square error predictor of Y_i given X_i. Out of all $m(X_i)$, the class of all possible functions of X, we have

$$E[Y_i|X_i] = \underset{m(X_i)}{\operatorname{argmin}} E[(Y_i - m(X_i))^2]$$

  - The ANOVA theorem (analysis of variance):

$$Var(Y_i) = Var(E[Y_i|X_i]) + E[Var(Y_i|X_i)]$$

- Technical details about the population regression solution vs. asymptotic OLS inference
  - The best regression coefficient vector, i.e. the solution to $\beta = \operatorname{argmin}_b E[(Y_i - X_i b)^2]$, is given by

$$\beta = E[X_i^T X_i]^{-1} E[X_i^T Y_i]$$

    This is also the best beta for the CEF $E[Y_i|X_i]$, not only Y_i. This insight allows us to apply "grouped data" strategies where, when we don't have access to microdata (individual data points), we can regress on aggregated data points which are averages conditional on our input features.

  - The sample analog of our population beta is

$$\hat{\beta} = (X^T X)^{-1}(X^T Y)$$

    where X and Y are our sample data points and respective targets. As we expect, $\hat{\beta}$ converges in probability to our population β from above and has an asymptotic normal distribution (with covariance matrix)

$$\hat{\beta} \xrightarrow{d} \mathcal{N}(\beta, E[X^T X]^{-1} E[X^T X \epsilon^2] E[X^T X]^{-1})$$

    Under homoskedasticity assumptions, where the residual variance is σ^2, the covariance matrix becomes $\sigma^2 E[X^T X]^{-1}$.

Extra Regression Details

- Sample weights and when to use them
  - One common situation is having samples that are nonrandom, sampled differently compared to the underlying population distribution, but our target is still the population regression function. If we know our data is constructed with sampling weights w_i equal to the inverse probability of sampling observation i, then we can use weighted least squares using w_i.
  - We can also use weighted least squares on grouped data, weighted by the underlying frequencies of the data in each group, but this is not strictly necessary (especially in macroeconomics where the scientific convention is the unweighted analysis of aggregate variables).
  - Heteroskedasticity is not a good reason to use weighted least squares over OLS with robust standard error.
- Saturated regressions are a way of constructing a regression with a guaranteed linear CEF, and are available to us when all our input variables are discrete.
  - We can create dummy variables representing all possible values of each input variable (the "main effects") as well as dummy variables for all possible products of our input variables (the "interaction terms").
  - Generates coefficients for each main effect and interaction term, and since everything is a zero-one dummy, the underlying CEF is linear so the regression will fit it perfectly.
  - Saturated models are the most restrictive strategy for modeling; creates a perfect fit but the interaction terms and their coefficients may be highly noisy/imprecise/meaningless.

5 QUANT RESEARCH - CASE STUDIES

5.1 Two Sigma - NY Housing Prices

Our goal for this case study is to take a hypothetical dataset about NYC housing and use it to predict sales prices and valuations for NYC houses in the future and for other houses whose sales prices are unknown. This is a somewhat common data science scenario and similar datasets exist out there which look something like:

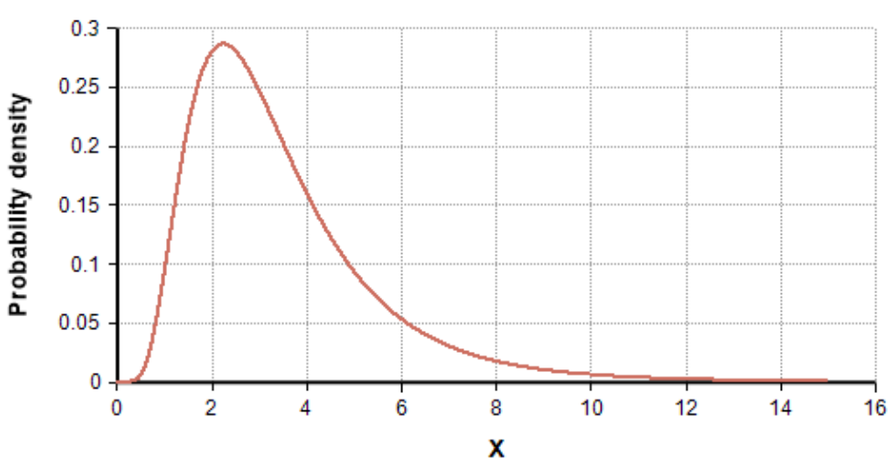
tx_price	beds	baths	sqft	year_built	lot_size	property_type	exterior_walls	roof	basement	restaurants	groceries	nightlife
295850	1	1	584	2013	0	Apartment / Condo / Townhouse	Wood Siding			107	9	30
216500	1	1	612	1965	0	Apartment / Condo / Townhouse	Brick	Composition Shingle	1.0	105	15	6
279900	1	1	615	1963	0	Apartment / Condo / Townhouse	Wood Siding			183	13	31
379900	1	1	618	2000	33541	Apartment / Condo / Townhouse	Wood Siding			198	9	38
340000	1	1	634	1992	0	Apartment / Condo / Townhouse	Brick			149	7	22
265000	1	1	641	1947	0	Apartment / Condo / Townhouse	Brick			146	10	23
240000	1	1	642	1944	0	Single-Family	Brick			159	13	36

Indeed, we assume our dataset just contains the first few variables here: i.e., we have access to most recent sales price, number of beds, number of baths, square footage, year built, and location (borough, neighborhood).

Our first and main idea is to perform a multivariate regression on the output, sales price, with respect to all our inputs, including square footage, num beds, num baths, year built, and location. Beyond this main framework, there are many considerations we need to dive into. How do we preprocess the data, and how do we deal with main regression issues that might come up?

Preprocessing
Let's look at preprocessing the data. Our first concern is encoding each variable so that we can perform linear regression properly on them. Square footage and year built are effectively continuous variables, i.e. even though they take integer values in our dataset the discretization is insignificant enough that they look continuous to the regression, so we can start of by keeping both variables as is. Number of beds and number of baths can be treated as discrete variables, which also work easily in a regression. Finally, the location (boroughs) is categorical data, so it's best treated with a one-hot encoding.

Normalization
Next, we may want to normalize our encoded data; normalization is good for mitigating numerical issues that can arise in multivariate regressions with different data types, and also generally helps with interpretation. For most of our input variables, it makes sense to just assume something like normal or uniform distribution and subtract mean, divide by variance, etc. But what about square footage? It's not exactly a normally distributed variable. We notice that it's always nonnegative and concentrated around a mean corresponding to a standard one- or two-story family home; the square footages in the left tail have less examples and are closer to the mean, since once you get to half or one-quarter of the size of a one-story family home, square footages at those sizes become very uncommon. However, the right tail has more examples and extends farther; there's plenty of large multi-story houses and mansions whose square footage can be dozens of times as large as average. The shape of this distribution suggests that square footage is lognormal:



Therefore, we can take the logarithm of square footage then normalize, using that as the input.

Correlations
Finally, we can start paying attention to correlations. Naturally, we suspect that square footage, number of beds, and number of baths are all correlated with each other, so how do we deal with this or potentially correct for this later? Our first task is to actually measure the correlation. The most well-known type of correlation, Pearson correlation, may work here but may not be the most effective. Since square footage is effectively continuous and number of bedrooms is discrete, the most appropriate type of correlation is a lesser known one called Spearman rank coefficient, which is the best for comparing continuous vs. discrete ordinal variables. Depending on the correlation value found, we may need to deal with multicollinearity later.

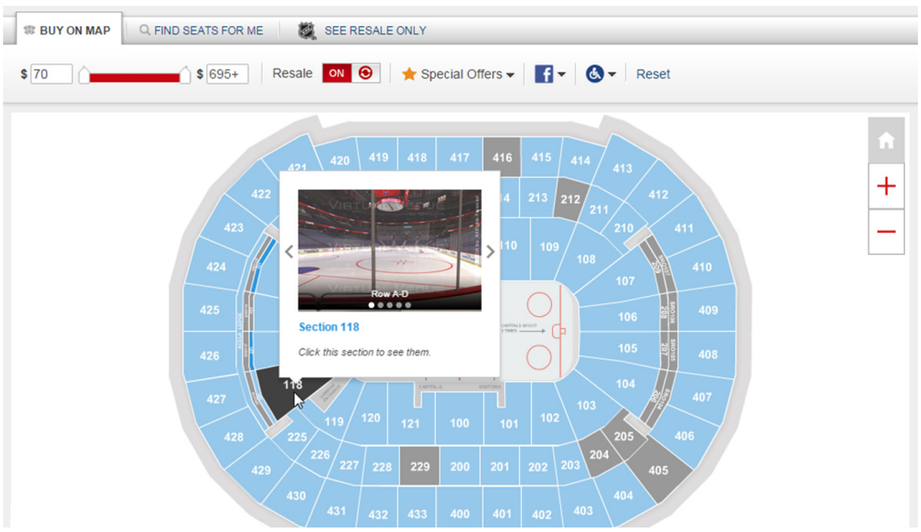
Multicollinearity
During and after we run the regression, we can explore various avenues of subset selection and dimensionality reduction, especially given the square footage vs. num beds or baths correlation we may have found earlier. We can perform ridge regression to reduce weighting on the insignificant variables, or lasso regression to cut out the less significant of the correlated variables entirely. We can also perform regression by successive orthonormalization or one of the stepwise regressions, which would prioritize the most important variables to regress on first and then, in later steps, regress on other variables' residuals with respect to the first variables, compensating for the correlation in less important variables. Finally, in the interpretation stage, we can perform t-tests to obtain z-scores on individual inputs, telling us which predictors are significant and which are not (potentially telling us to go back and try a subset selection cutting these out).

For the most part, we can successfully perform a multivariate regression for housing price prediction after dealing with these extra considerations.

5.2 QuantCo - Opera House

Our goal for this case study is to take data about ticket sales at a concert venue and use it to construct a model for how to price tickets in the future. We can assume we have access to any aspect of ticket sales we want, i.e. for each row/column/section number, we have a history of price, when it was sold, what concert it was for when that concert happened, etc. This problem can get much more involved and out-of-the-box because there are factors affecting our model that don't just come from the inputs. If we're consulting for the concert venue and trying to maximize profits (change the way we price things) instead of just predict future prices, we want to bring in other ideas such as consumption/income statistics and spending psychology to really do an effective job with building a model.

There are many good ways of answering the question, so we can just pass around a couple ideas below.



Nearest Neighbors and Regressions
If we focus on the row number/column number/section number inputs for our data, we may be inclined to do one of the classical regression/classification approaches. Even before diving into regression, the simplest approach, nearest neighbors, actually makes a lot of sense here. Rows, columns, and sections, contextualized on an actual diagram of the seats in the venue (which we can assume we have), give rise to actual 2D/3D spatial relationships. How near/far the seat is to the stage, the angle the person would see the stage from, all directly affect how good/valuable the seat is, and therefore adjacent seats should be priced very similarly while far-apart seats should be priced differently. In nearest neighbors, we can mainly control how many neighbors we want to account for for each seat, and to add complexity, we can add a kernel, which is just

putting weights on each neighbor based on their distance to the original point. Then in this model, the price of each seat is a weighted average of the prices of nearby seats. One last thing to consider is boundary points, or corner and edge seats in sections. We'll want to have a larger window (maybe including points two or three seats away) and tune our choice of kernel to properly model at the boundary.

Beyond nearest neighbors, we can perform a multivariate linear regression on sales price with respect to row, column, and section numbers. This also makes sense because, as discussed before, the distance to the front stage changes with row number and angle to the front stage changes with column, so changes in either one will affect the value of the seat. Given the distance/angle idea, we'll want to preprocess around this idea. It might be good to convert row number into ordinal data, since they are ordered by distance, or even to replace row numbers with Euclidean distance to front stage and use that as the input instead; the same thing can be done for columns to some extent. We can even pair up points based on left-right symmetry in the venue layout. One main concern is the correlation; it's likely that row, column, and section are all highly correlated with each other. If we want to incorporate multicollinearity approaches, we can use lasso or subset selection techniques to find which variable to drop, and if we don't want to entirely drop any variables, we can do ridge or regression by successive orthonormalization (regressing on residuals).

With only nearest neighbors and regressions on row/column/section, we can either run into a lot of outliers and strange behavior in the data or we might not capture all the complexity of the data. This is where aspects of consumption/income microeconomics and psychology might come into play.

Advanced Ideas
Ideas in this section are taken from
`https://towardsdatascience.com/statistics-for-dynamic-pricing-of-theatre-87df073a0848`

We haven't used sales times in our modeling yet, but from real-life knowledge the amount of time left before the concert happens has a major effect on what the price of buying a ticket would be at that point in time. The rate of people buying tickets (not looking at price yet) continually increases as the date gets closer to the time of the concert; we can make a reasonable assumption that this increase is close to exponential, and therefore we can perform an exponential fit on number of tickets sold vs. (concert date minus sell date). Then, if we interpret this as demand increasing exponentially with time and also supply decreasing exponentially with time (total number of seats in venue minus seats already bought), then we infer that price can also correspond to time in an exponential fashion.

One straightforward way to incorporate this information is to take the log of (concert date minus date sold) and add that as an input in our multivariate regression from earlier. Since our time analysis is pretty far removed from the spatial analysis of row/column/section earlier, we don't expect much correlation between the log-time variable and the row/column/section variables, and we've now captured an important, previously unseen aspect of the complexity of the data. (Obviously, we should still check for multicollinearity with the techniques we've been discussing, just in case.)

We can loop back around to our point about supply decreasing dramatically as the time gets closer to the concert date. Here, some spending psychology effects may come into play and further increase upward pressure on prices. For example, if there are very few seats left, loss aversion/FOMO comes into play as people may miss out on the concert entirely if they don't purchase on the last few seats, therefore driving up their willingness to pay.

One way we can incorporate this into our model is to make a new variable for scarcity, i.e. aggregating the tickets sold before the current ticket and subtracting that count from the total available, then adding scarcity as an input into our multivariate regression. It's more likely that there will be some correlation between scarcity and log-time, so we can incorporate multicollinearity approaches and pay attention to these two variables in that context. The scarcity idea also helps with interpretation. One thing that we might notice is that the very last row in the back has prices higher than what linear regression on row/column/section should give, and that can be explained by scarcity, as the undesirable back row seats are usually the very last to go and may be filled by FOMOed people with much higher willingness to pay.

As a final note, if we're maximizing profits, we may want to incorporate microeconomic data about consumption. For the general population, the distribution of disposable income is lognormal (like square footage from the housing case), and willingness to pay for concert tickets likely follows the same lognormal pattern. Since people will buy the tickets if they're at or below, but not above, the price they're willing to pay, we can approach maximizing profits by modeling the distribution for willingness to pay, then adjusting prices across row/column/section/log-time/scarcity so that we capture just enough willing-to-pay people as can fit in our venue. This goes beyond our main multivariate regression, but it's a start for how we would actually modify prices to increase profits.

5.3 Two Sigma - CitiBikes [Advanced!]

This case study is more open-ended than our earlier housing and opera cases. Our goal here is to build a predictive model for CitiBike usage across Manhattan; for any CitiBike station that we specify, given any recent data we think is relevant to CitiBike usage patterns, we want to predict the amount of CitiBikes docked in the station, as well as the changes in this amount, over the near future. We no longer have an explicit set of available input variables that we have to use, but rather we have to brainstorm what input variables are relevant from all the possible data about Manhattan before we even draw out the technical details of regression on these variables.

In an interview context, you would probably iterate back and forth between brainstorming new things to include in your model, vs. performing data preprocessing and model tuning/evaluation for these new variables. For this writeup, we'll do two iterations of introducing new variables.

In our first round of brainstorming, we come up with a handful of the more obvious variables that are very likely to have strong effects on CitiBike usage patterns in Manhattan. These first-pass variables include the time of day, location/neighborhood, month/season, temperature, and weather pattern (i.e. sunny, cloudy, rainy). Each of these variables has a relatively complex treatment in our preprocessing.

First, the time of day and the month/season are both cyclical variables; although they are both ordered, this order loops back on itself so that we have no definition of least to greatest. A naive approach would be to make these into ordered discrete or continuous variables anyway, i.e. the time of day would be used as the numerical hour and minute (0 to 23, 0 to 59) while the month is also used numerically (1 to 12 for January to December). This can produce workable results, but the blatant problem is the discontinuity that the ordering introduces, discarding the cyclicality of the data. One common approach to processing the cyclical data is to encode it as one-hot, so that the hour and month become 24 and 12 distinct indicator variables, respectively, in a one-hot vector. Since this introduces many more dimensions of data and also removes the ordering, we may want to bucket the variables into smaller groups before one-hot encoding. For example, the time of day can be bucketed into morning, afternoon, evening, and midnight, while month can be bucketed into the four seasons; these smaller handfuls of categorical variables can then be one-hot encoded with smaller dimensionality than before. The bucketing also recovers some of the ordering of these cyclical variables, as sequential slices of the variables are collected into each bucket. Another more advanced approach is to directly capture the cycles with a cyclical function; this entails a basis transformation with a cyclical function, which we can achieve in a few ways, perhaps with a trigonometric function or a cyclic spline. (Splines are essentially piecewise polynomial transformations, and their technical details are pretty outside the scope of this bible; feel free to Google and learn more about splines on your own time.)

Next, location/neighborhood is potentially the most complex variable to treat, since neighborhoods are irregularly distributed around the geography of Manhattan and also have irregular, arbitrary patterns of activity with complex social factors way outside the scope of this regression problem. One basic approach is to make a reasonable assumption that CitiBike usage within the same neighborhood, or even between adjacent neighborhoods, is more uniform, so that we can directly encode the neighborhood in a one-hot fashion. Another approach, unique to this context, arises when we notice that location/neighborhood can be encoded as a coordinate pair of latitude and longitude, both of which are continuous variables. Then the irregularity of neighborhood activity patterns reduces to arbitrary nonlinear dependencies of neighborhood activity on latitude and longitude, which requires a basis transformation. If we reintroduce the assumption of uniformity of activity within neighborhoods, we can think of the problem as regressing CitiBike usage on neighborhood and simultaneously regressing neighborhood on some nonlinear transformation of latitude and longitude. We can eyeball an argument with the linear nature of the regression equation that these two layers of regressions have an equivalent effect to directly regressing CitiBike usage on the exact same nonlinear transformation of latitude and longitude. With the relatively small number of neighborhoods, we can argue that a medium-degree polynomial transformation of latitude and longitude can adequately capture the nonlinearity. Since even a medium-degree transformation can introduce a large number of new variables, we can further look at low-degree polynomial splines, such as quadratic or cubic splines, to attempt to reduce the dimensionality of this transformation.

Finally, the temperature and weather variables are the simplest to encode in isolation. Temperature is a continuous variable that can be encoded directly (perhaps after some normalization), and weather is a categorical variable that can be one-hot encoded. The trick part of these variables is the correlation with each other and with our other variables. First, temperature and weather condition can be highly correlated with each other. The solution here is to first include both temperature and weather in the regression, then calculate the correlation between the two; since we are dealing with a categorical vs. continuous variable, some appropriate correlation measures can come from the "point biserial correlation" or from logistic regression. If the correlation is high enough to be a concern, we can then turn to the various dimensionality reduction or subset selection approaches to perhaps truncate one of temperature or weather condition, or proportionally shrink both variables in the case of ridge regression, or even perform some transformation that combines the two variables into one. Another concern is the correlation between temperature/weather and the time/season variables from earlier, as the average temperature as well as trends in weather conditions explicitly follow along with the time of day or the season of the year. For variables that follow a cyclical pattern with a cyclical variable also in the inputs, a common way to patch up this multicollinearity is to normalize the variable by subtracting the means from the previous cycle. For example, temperature can be normalized into a "temperature difference" with respect to the time of day by subtracting the temperature at the exact same time from the previous day. These normalized variables can then be used as the inputs in the regression, their correlations evaluated afterwards, and dimensionality reduction/subset selection approaches used if necessary, as usual.

With the first batch of variables preprocessed and tuned adequately, we can now turn to brainstorming another batch of variables. Examples of less obvious variables we can think of are day of the week, holidays, traffic intensity, subway usage, inches of rain, air quality index, local housing prices, and local business activity. These variables also have complex considerations, but these considerations are similar to the cyclical, correlational, or geographic concerns that we discussed in detail with the first batch of variables, so we don't need to discuss the preprocessing here. Instead, it's worthwhile to note the tuning aspect of our regression after we've incorporated all ten variables. With a relatively high number of variables, it will be necessary to test for inclusion/exclusion as well as try various dimensionality reduction approaches. First, we will want to calculate z-scores for each individual variable, and note which variables don't pass the significance threshold. Depending on the correlations between these insignificant variables, we will then want to calculate F-statistics for groups of correlated insignificant variables to verify whether the entire group can be excluded. Then, for dimensionality reduction tasks, we may opt for a regularized regression such as lasso or ridge. Because of the high number of intercorrelated variables with various variances, we can argue that different attempts of elastic-net, which incorporates both the proportionality shrinkage aspect of ridge and the soft thresholding/truncation of lasso, may result in the most effective dimensionality reduction. If we want to extract a big-picture view of the regression, we can opt for the subset selection approaches such as forward or backward stepwise regression.

6 QUANT TRADING - MARKET MAKING

6.1 What is Market Making? by Evan and Guang

When you (or an institution) go to place an order, ex. "Buy TSLA Calls," you need to buy those calls from somebody. Same for an order selling AAPL stock–you require another party to sell to who will give you money in exchange for that stock. What happens, though, if there is nobody available to buy a stock from (there are 100,000 primary security equities... do you think you can always find someone to trade all of them)? What if you can find somebody, but they don't want to sell you the financial product at the market price because they are the only seller and you want to buy. This is where *Market Makers* come in. In its simplest form, the job of a Market Maker is to always be available to buy or sell a particular financial instrument at the prices they quote. Many exchanges partner with Market Makers to ensure that securities traded on the exchange are "liquid" (not subject to sudden price fluctuations due to trade volume), and in return Market Makers are given special access to information about order flow so they can quote their prices in a more favorable way.

In most cases, a Market Maker will take on a particular financial instrument (a stock, option, bond, warrant, ETF, etc..) by always offering a bid-ask spread. The bid-ask spread is usually expressed as $x@y$, where x is the bid and y is the ask (also called the offer). The *bid* price is the price the Market Maker is willing to buy the security at, and the *ask* price is the price the Market Maker is willing to sell the security at. Sometimes, the bid-ask spread has a restriction on its *width*, meaning ask $-$ bid $< k$ for some given value k (often very small). In almost all cases, Market Makers are *required* to execute a transaction at the bid or ask prices they quote.

The execution of a transaction (say the Market Maker buys a stock from an individual who wants to sell), carries risk for the Market Maker. The risk in the aforementioned example is that the price of the security they just bought will go down. In theory, the bid-ask spread should compensate a market maker for the risk in taking on either side of a trade. After the trade, the Market Maker can use hedges such as options or other financial instruments to mitigate that risk. This principle–mitigating risk through hedged betting–can show up in interviews for Market Making companies.

6.2 Theory by Ravi

In practice, there are three main determinants of your market (bid@ask).

Theoretical Value: This is what you think it's worth. If it's a market on the outcome of a die roll then you can make a tighter market since it is a known quantity (3.5). If it is something obscure, like the number of ping pong balls that can fit in the Empire State Building you will need to make your market wider. The interviewer wants to see that you are adjusting your market for risk due to uncertainty.

Last Price Traded: This is the going market price. Sometimes the trading price deviates from your theoretical value. In this case the trader needs to balance his faith in the market with his faith in his models. Interviewers don't usually give you the last traded price, but if you are playing an iterative game that involves multiple trades with an interviewer, knowing the last traded price will help you adjust your market over time.

Current Position: This is your net long/short exposure. Ideally, market-makers like to be flat, meaning they have no exposure to movements in the asset price. If you have accumulated a serious position you would want to make your markets asymmetric to make either buying or selling more desirable. For example, say you are extremely long an asset that is worth \$0.50. A reasonable market may then be \$0.43@\$0.53. This means you are willing to sell for less theoretical profit than you are willing to buy. You are giving up some "edge" to reduce your exposure. Generally, if you are flat your market should be symmetric (i.e. bid/ask are equidistant from theoretical value).

A trader's market is a balance of these three things. It is not uncommon for interviewers to ask for your confidence interval.

Confidence Interval: A confidence interval is an interval where the true value will fall within your interval a certain (given) percentage of the time. For example, if I am picking real numbers from a normal distribution with mean 100 and standard deviation 10, a 95% confidence interval would be 80@120, as 95% of the data will fall between 2 standard deviations of the mean. These confidence interval questions can be less straightforward; for example, SIG has asked people to generate a 90% confidence interval on the number of windows in their building, which is not a well-known quantity that can be derived by some formula. Thus, it is important to remember that confidence intervals should be wider if there is more uncertainty, just like a market.

Finally, it is important that your markets are realistic. Answers like 0@1billion might encompass the true value, but they will never get trades which defeats the purpose of a Market Maker. In the real world, the highest bid and the lowest ask determine the market. The tightest and fastest markets usually get the majority of trades and no one will deal with absurdly wide traders.

If we revisit the windows example, i.e. how many windows are there in SIG's building, we can play out a scenario that illustrates the importance of market width and informational asymmetry. Let's say that you are asked to make a market which will then be traded on by your interviewer on the number of windows in the building.

Your bid is the price at which you are willing to "buy" the number of windows. Generally, your bid will be below your theoretical estimated value of the number of windows, but there are some cases where it may be above (we will talk about those later). Your ask is the price at which you are willing to "sell" the number of windows. The notion of "buying" and "selling" the number of windows might be a little bit confusing, as these are non-financial instruments, but the process is very intuitive. Suppose you "buy" the number of windows at 300 windows. If there were 350 total windows, then you make 50 units of whatever you and your counterparty were trading (dollars, meal swipes, hours of homework help, etc.), so you gain some desired commodity. If there were only 275 windows, then you would lose 25 units of whatever you and your counterparty were trading, so you lose some desired commodity.

Back to the windows example, suppose there are 400 windows in total (you don't know this), and your first market is 350@370. Your interviewer (if they know the actual number), will buy from you at 370. On that trade, they will make 30 units and you will lose 30 units, as $400 - 370 = 30$. If they didn't know the true value and happened to sell at 350, you would make 50 units and they would lose 50 units.

Taking the simple case of just you and the interviewer, even if you know with 100% certainty the number of windows, its in your interest to make the widest market possible at which you think your counterparty will actually trade. The wider your market, the less probable your counterparty will trade. Therefore if you think your counterparty has a very different view from you, you might give a wider market, or skew the market one way. For example, if you think the interviewer will estimate that there are at least 600 windows, you might open your market as 550@600, because you know the interviewer is more likely to buy at 600 windows than sell at 550 windows. Even though your bid of 550 windows is higher than your initial theoretical value, shifting the market upwards will give you more compensation when the interviewer chooses to buy from you.

6.3 Cases by Ravi

Many quant firms ask interview questions about market making for the purpose of simulating a real-life trading marketplace. It's really important to keep track of your current position and adjust your markets based on trades. Another important aspect is having good "trader memory," or being able to keep track of the history of your positions and the PNL you've made on these positions up to your current one.

We illustrate many of the principles in market making theory with a few "case studies" similar to market making scenarios you may be asked in interviews.

Case 1: Sports Betting

Make a 10-wide market on the number of regular season games the Boston Red Sox will win in 2021. Then make a 50% and a 90% confidence interval. The interviewer will then trade (buy or sell) and you will then make a new market, and so on.

The first step towards making a market is to generate a theoretical value. In generating this theoretical value, it is important to lay out any assumptions you are making to the interviewer. Let us assume that there will be a full 162-game season, as the commissioner indicated. The next assumption comes with a bit of baseball knowledge, so it is not completely necessary but will help with generating a good theoretical value. The Red Sox had multiple injuries in 2020 especially to their pitching staff, and they improved their team in the offseason, so it is safe to assume that they will perform better this year.

In 2020, they only won 24 games, but the season was only 60 games long, so that is about a 0.400 winning percentage, whereas in 2019, they won 84 out of 162 games for a 0.519 winning percentage. In 2018, when they won the World Series, they won 108 out of 162 games, good for a 0.667 winning percentage. Using the data from the past three seasons (we could look at more but roster turnover would imply that more recent season predict future success better), we can reasonably conclude that the Red Sox winning percentage this year will likely be closer to 0.519 than either 0.400 or 0.667. Using this, we can construct a theoretical value. We shade our winning percentage estimate a bit down from 0.519, given that the 2020 season was more recent and the team has lost a lot of its members from the 2018 team. Suppose we say that the Red Sox will win 50 percent of their games this year. That yields a theoretical value of 81 games, and now we can construct a market. Given the 10-wide requirement, our market will be 76@86 games.

As for confidence intervals, we can assume that the number of wins is a Binomial random variable with $N = 162$ and $p = 0.5$. We can assume N is sufficiently large and that this distribution will be approximately normal. Then we have that the expected number of wins is $162 * 0.5 = 81$ and the standard deviation of the number of wins is $\sqrt{162 * 0.5 * 0.5}$, which is approximately 6.36.

Since the z-score for a 50% confidence interval is about 0.67 and the z-score for a 90% confidence interval is about 1.64, we can generate our approximate confidence intervals:

50% confidence interval: $81 \pm (0.67 * 6.36) \implies [76.75, 85.25]$
90% confidence interval: $81 \pm (1.64 * 6.36) \implies [70.5, 91.5]$

Returning back to your original market of 76@86 games, we now play out the trading portion of the game. Your interviewer will trade on each market, after which you will have to give a new market. This will continue over a series of trades until the interviewer chooses to stop the game. Suppose the conversation between you (Y) and the interviewer (I) takes place as follows:

Y: 76@86
I: Buy!
Y: 81@91
I: Buy!
Y: 85@95
I: Sell!
Y: 83@93
I: Buy!
Y: 84@94
I: Sell! Stop the game. What is your PNL and current position? If the Red Sox win 90 games, how much do you make/lose? What is the breakeven number of wins for you?

Before answering the interviewer's questions, we will go through the process of generating each market. Your opening market is 76@86, which we calculated before. Now the interviewer buys. You should now move your market up for a few reasons. First, the interviewer likely thinks the value is higher than 86, and you want to incorporate this additional information into your new market. Second, if the interviewer buys for 86, then you know they will very likely buy for 86 again if you show 76@86 as your second market. So, even if you are happy selling at 86 wins, you can raise the market to sell at a higher price. Finally, the interviewer bought from you, so you are short one contract, and want to ideally flatten your position. Obviously, if you move your market down, the interviewer will buy again, and if you move your market up, the interviewer becomes more likely to sell to you. Now that we have established that you should move your market up, we begin to generate the new market. Unfortunately, there is not really a set formula to this, but given our relative lack of information about the interviewer's valuation, we move our market more at the start. We moved our market up by half our market width and up to the edge of our 90 percent confidence interval, so that we are relatively sure that we are okay if the interviewer buys again, and are very happy if the interviewer sells. Thus, our second market is 81@91.

Once the interviewer buys from us at 91, we have to move our market up a bit more (maybe a little less this time since we still have some faith in our original prediction). Our third market is 85@95. The interviewer sells to us at 85. This is a good spot, because we just sold at 91 and bought at 85. Now, we can reasonably bound the price between 86 and 90 (when the interviewer buys on our 81@91 market, they probably think the fair value is on the greater side of the midmarket, 86, and when the interviewer sells on our 85@95 market, they probably think the fair value is on the lower side of the midmarket, 90). The interviewer just sold to us on our 85@95 market, so let's move it down to 83@93 (notice the midmarket of this is 88 which is right in the middle of where we bounded the interviewer's fair value). Once the interviewer buys from us at 93, we should move up the market, but keep note that they sold 85@95, so let's try 84@94. The interviewer sells to us at 84 as the last trade in the game. Note that these last two trades generated riskless PNL, just from having an idea of the interviewer's theoretical valuation. We sold to the interviewer at 93 and then bought from them at 84, generating a riskless PNL of 9 on just 2 trades.

Let us move onto the interviewer's supplemental questions. Our current position and PNL are easy to calculate if you pair up opposite sides of trades. Here is the complete list of trades: sold at 86, sold at 91, bought at 85, sold at 93, bought at 84. If we pair up the first and third trade, we see a PNL of 1 unit and a flat position. If we pair up the fourth and the fifth trade, we see a PNL of 9 units and a flat position. This just leaves the with the second trade, so our current position is short 1 contract of 91 wins and a PNL of 10. You could also give a different PNL and contract that you were short if you paired up the trades differently, as long as the PNL and contract number of wins add up to 101. If the Red Sox win 90 games, our short position makes 1 PNL, so our total PNL is 11. The breakeven number of wins is 101 wins, meaning that as long as the Red Sox win less than 101 games, which we seem fairly sure about, we will generate positive PNL.

This game illustrates many aspects of market making. It focuses on the importance of theoretical value generation, confidence intervals, informational asymmetry, and market adaptation in response to trades. If you can go through a similar example that stresses these points, you will be all set for any pure market making scenario during an interview.

Case 2: Country Population

Make a market on the population of Tanzania. The interviewer will trade on the market, after which you will continue to make markets and the interviewer will continue to trade. As an additional rule, the ask on each market can be at most 1.5 times the bid.

The actual population is 56 million. Unless you're an African country aficionado, you probably don't know that exactly, so let's assume we are off with our first theoretical value. We can play the game out from there, as follows (note that the markets are assumed to be in millions of people):

Y: 20@30
I: Buy!
Y: 60@90
I: Sell!
Y: 40@60
I: Buy!
Y: 50@75
I: Sell! Stop the game. What is your PNL, assuming one dollar per million people, and current position?

We know that the US population is 325 million, and given that Tanzania is much smaller, let us generate a theoretical value of 25 million. As mentioned above, the true population is more than double this number, but we will play this game out from here to show how to do these exercises when starting with little information. Given our estimation of 25 million people in Tanzania, our opening market is 20@30.

When the interviewer buys, we have to move our market up. Since we don't have any other information besides this 1 trade, we should err on the side of moving our market up more, given that the potential downside of moving our market up very little outweighs the potential downside of moving our market up too much. Given the US population is 325 million, we can reasonably bound the Tanzanian population below 100 million. Now, we can triple our original theoretical value to get a new theoretical value of 75 million, and build our market around that, which gives 60@90.

When the interviewer subsequently sells, we are in a much better spot, as we have some reasonable bounds for our population estimate. It is likely between 30 million and 60 million but could be slightly outside that range. So our new market could be anything from 30@45 to 40@60. Given that we made the first market with less information, moving the market a bit less this time and keeping it closer to 60 million than 30 million makes sense. Additionally, the first trade suggests that the true value is greater than 25 million, while the second trade suggests that the true value is less than 75 million. The middle of those values is 50 million, and building a market around that yields 40@60. Thus, 40@60 seems like a reasonable market given the previous trades.

When the interviewer buys, we are in a really great spot. In our second market, the interviewer sold to us at 60 rather than buying at 90, and here the interviewer elects to buy from us at 60 rather than sell at 40. This suggests that the true population value is quite close to 60 million. Taking a similar approach as before, the second trade suggests the true value is below 75 million, while the third trade suggests the true value is above 50 million. Given that the interviewer bought from us when we gave 40@60 as our market, we need to

move our market up. 50@75 seems very reasonable, given the explanation above, but let's just validate the size of our market moves. From the first to second market, we moved the midmarket by 50 (25 to 75). From the second to third market, we moved the midmarket by 25 (75 to 50). Thus 50@75 would then move the midmarket from 50 to 62.5, which is 12.5. This validates our new market, as when we have trades on both sides that bound our estimate, we generally want to move our market by less on subsequent trades (since our markets become more and more accurate over time). Thus, our fourth market is 50@75.

The interviewer sells and asks us our PNL and position. The position is quite simple, as the interviewer bought twice and sold twice, so we have a flat position. Our PNL can easily be calculated by pairing up buys and sells. Pairing up the first two trades, we sold to the interviewer at 30 and bought at 60, giving us a 30 dollar loss. Pairing up the last two trades, we sold to the interviewer at 60 and bought at 50, giving us a 10 dollar profit. Overall, our PNL is -$20.

Note that we lost money playing this game. That's more than okay and will often happen in these games. The interviewers do not expect you to know the population of Sub-Saharan African countries. This game is more of a test of your reaction to trades and moving your market so as to maximize your gains and minimize your losses.

Case 3: Trade or Tighten

Suppose you and your interviewer are going to play a game that takes place as follows. Before the game, each one of you is given a theoretical value for stock XYZ. Each theoretical value is drawn independently from a uniform distribution from 90 to 110. You will open with a market, after which the interviewer can choose to tighten the market (by increasing the bid and/or decreasing the ask) or to trade on the current market. If the interviewer tightens the market, then you have the choice to tighten their new market or trade on their market. The process of tightening continues until someone elects to trade. The theoretical value you receive for XYZ is 100. The interviewer also receives their theoretical value, which you do not know. Your goal is to now make the best possible trade according to your theoretical value.

A well-played game between you (Y) and the interviewer (I) might go as follows:

Y: 90@110
I: 95@107
Y: 97@106
I: 100 bid
Y: 101 bid
I: 102 bid
Y: @105
I: 103 bid
Y: @104
I: Buy!

Notice that we open up with 90@110, since we know that distribution that the theoretical values are drawn from ranges from 90 to 110. The interviewer then responds with 95@107. While we should not trade on this market (both selling for 95 and buying at 107 would be losing trades for us), this new market does give us information about the interviewer's theoretical value. If the interviewer's market was symmetric around the theoretical value, then this suggests a theoretical value of 101. Now we can't assume that the theoretical value is directly at the midmarket, but it is more likely that the interviewer's theoretical value is greater than 100, and thus greater than our theoretical value.

Using this information, we can respond with 97@106, giving the impression that our theoretical value is also higher than 100. We are also happy if the interviewer trades on this market. The interviewer then responds with 100 bid, which cements our confidence that interviewer's theoretical value is greater than 100.

Now, the inside market is 100@106 (the interviewer is on the bid and we are on the ask). We have an option here. We can either say @105, or be 101 bid. The former is a less risky strategy but might reveal too much of our pricing, so let's elect the latter option. If we respond with 101 bid, this option will have a higher reward to compensate us for the risk. This is a good risk to take because the interviewer is unlikely to sell to us at 101, given that they were 100 bid and the fact that both parties have been pushing the price upwards with the past few markets. Thus, our 101 bid gives us the option to sell at a much higher price in the future than if we were @105.

The interviewer is then 102 bid after we were 101 bid. Now, this is a pretty clear spot between showing @105 or 103 bid. We will elect the former since the risk of getting sold to at 103 is way too high to get a marginally better price. After we show our offer of 105, the interviewer responds with 103 bid.

Now, the inside market is 103@105 (the interviewer again is on the bid and we are on the ask. Obviously, we can't buy from ourselves, so we then show @104 as our final tightening of the market. The interviewer buys, and we make a PNL of 4 marked to our theoretical value that we were given.

The interviewer's theoretical value was 105 in this game (which we obviously didn't know, but as the game went on, we got a better idea of it). As a bonus exercise, see if you could have figured out that theoretical value given the trades that the interviewer made.

7 QUESTION BANK

7.1 Preliminaries

Beyond the company-specific questions discussed later in this section, there are plenty of "classic" types of quant questions that can pop up in any interview, especially earlier phone rounds. There are dozens of not hundreds of these questions and rather than single out a few to put into this section, we decided that they're all important enough to brush up on every single one as you head into recruiting season. You'll find some important questions from two books we mentioned in the intro to this bible:

- *Heard on the Street*, Chapter 2. This contains 70 math brainteasers that you'll probably run into in any finance recruiting process, and certainly in the quant process.
- *A Practical Guide to Quantitative Finance Interviews*, Chapters 2 and 4. This is the quant "green book" you might hear people mention at MIT.

Since these are more fundamental questions compared to the pretty complex stuff in the rest of the section, we recommend looking at these books first, but after you've built up the math (and CS) coursework and reviewed probability, stats, and data science with the rest of the quant bible.

7.2 JANE STREET by Evan and Brian

(Round 1)

- There are 10 people in a room; how many ways to choose three of them?
  - $\binom{10}{3}$
- We are going to play a dice game with a d6.
  - You pay to get the amount in one roll; what is a fair price?
    - 3.5; $sum_i = 1^6 i * 1/6$
  - You will roll the dice, then choose whether to re-roll. If you re-roll, you get whatever the value of the re-roll is. What is a fair price?
    - We should only re-roll if we get less than 3.5 in our first roll (since on the next roll we also expect to get 3.5). Thus we get $\frac{3}{6} * (3.5) + \frac{1}{6} * (4 + 5 + 6) = 4.25$ for our expectation.
  - You will roll the dice, then choose whether to re-roll. If you re-roll, you get whatever the value of the re-roll is. The re-roll costs \$1. What is a fair price?
    - Now if we re-roll we only expect to get 2.5 (after paying 1), so we only re-roll if we get less than 2.5. $\frac{2}{6} * (3.5 - 1) + \frac{1}{6}(3 + 4 + 5 + 6) = 23/6$
  - Same as above, but now you have infinite re-rolls *each* of which cost \$1. What is a fair price/optimal strategy?
    - Payoff = (Payoff from roll)*(Prob of stopping) + (Payoff from re-rolling - 1)*(Prob of Rerolling); recognize that Payoff from re-rolling = Payoff (same game!), and let X=payoff, p=prob of re-rolling. $X = 3.5 * (1-p) + (X-1) * p \rightarrow X(1-p) = 3.5 - 4.5p \rightarrow X = (3.5 - 4.5p)/(1-p)$. Recognize that this is asymptotic to $p = 1$, so the question is just which side of the asymptote lines lie. Test points determine that $p = 0$ gives a payoff of 3.5 is optimal

(Round 2)

- Rank the following in order of expectation: the product of two d6's, the square of one d6, and the square of the median of 5 d6's
  - Product of two independent d6's can be computed via independence as $3.5^2 = 12.25$. Square of one d6 can be computed through definition of expectation as 15.16. Only question then is where the square of the median is. This can be computed via order statistics, but a nicer way to think about it is in comparison to the above two. We know the distribution here is skewed towards the middle (we would need 3 6's or 3 1's to get either extreme). Note that we gain much more through the squared 6 than we lose in the squared 1. This ranks it below the squared roll of one dice (which rates both of these equally with higher probability). When comparing it to the product of two die, my approach was shaky in comparing probabilities for (3, 4) to (1, 6), so I would recommend finding a better one. The correct answer is ([Least], Product of two, Square of median, Square of one, [Most])
- We are going to play a dice game with a d100; you will roll once for a fixed price (x), and then you can choose whether to re-roll. Each time you choose to re-roll, it costs \$1. When you finally decide not to re-roll, you get the value on the d100. What is the optimal strategy, and what is the fair fixed costs (\$x)
  - X = (50.5)*p

(Round 3)

- You and I will play a game; we each have a d6 (hidden from the other person). We each roll, and then a third party sees both of our rolls and puts the sum in a jar hidden from us. We then take turns bidding on the jar. At each step, you can either raise the bid by a positive integer amount (+1, +2, ...) or pass. If a player passes, the person who made the last bid pays their bid to the third party and receives whatever was in the jar. What is your strategy?
  - This is not a solvable problem; they want to see how you think, and there are a variety of ways to approach it. In my case, I chose to take the ultra-conservative approach and guarantee I would never lose money through my betting strategy. In the end you are required to come up with a specific strategy regarding what you will do for each potential dice roll you get, and doing so required me to think through each step of the bidding process to evaluate what my opponent could have for their die. One useful assumption to make is that your opponent has the same strategy as you do (you are, after all, claiming your strategy is optimal).

(Onsite; Virtual)

- Make me a market on the number of magazines in the AirBnB I am staying in (more just to check how well you know Market Making terminology; Fermi is not big)
- Market Making on a variety of games involving die (product of d6, d12, max of d3, d4, d8, etc...). They are guaranteed to take one side, but they will choose which.

- Ski-Ball! You have a virtual ski-ball machine where you are given (for each game) the probability of successfully getting a ball in a hole if you aim for that hole. You are then put up against a series of opponents who play with various strategies. You are given 5 practice rounds (to evaluate your opponent's strategy) and then play 1 round against the opponent. You put down a certain number of chips to play the game, and if you win the game against the opponent (get more points than they do in 10 throws), you win whatever you put down and receive your down back (so 2x). If you lose, you get nothing. Strategies opponent's played with that I observed: 1) always go for the 10-point hole (that hole had a 100% chance of success if you aim for it); 2) Always go for the 100 point hole unless they get it in which case they go for the 10-point hole, although sometimes if they miss the 100 point hole too much they switch to just going for the 10-point hole anyway
- There are 10 coins, 9 of unknown weight and 1 of known weight (50/50). You are given the PMF over the weights of the coins. You pay 50 chips to play the game, and then are given 100 flips of the coins. Every head you get earns you one chip. First question: do you want to play the game (answer yes by observing PMF). Actually playing the game, you are given the ability to record the observed H/T ratio for each of the coins as well as the number of times you have flipped it. You need to find an exploration vs exploitation strategy.
- You play a game where you throw your chips against the wall, playing against your interviewer. At each step you have two options: throw a chip, or pass. If you pass, it moves to your opponent's turn. If two people pass in a row, the person who threw the chip that has landed closest to the wall (over all throws) gets 20 chips, the other person gets nothing. You do not get back any chips you threw at the wall. Since this was a virtual interview, instead of a wall, we had two entirely unknown random number generators which output a distance to the wall.

Extras

- Three correlations problem. This is an important brainteaser example in linear algebra that I haven't seen in any of the "classic" quant interview books, and although I haven't heard it asked in any interviews in the past year or two, I'm throwing it in the Jane Street section because apparently it used to be asked at Jane Street in the past. The question: suppose we have three random variables X, Y, Z, and we know two of the three pairwise correlations. We know X, Y have a correlation of 0.9, and Y, Z have a correlation of 0.8, but we don't know the correlation of X, Z exactly. What are the best bounds that we can find for the correlation of X, Z?
  - This problem relies on a property of all correlation matrices. To illustrate the correlation matrix for this problem, let c be the unknown correlation between X, Z, and create a 3x3 matrix (rows/columns indexed in the order of X, Y, Z), so that the i, jth entry is the correlation of the corresponding random variables:

    $$\begin{bmatrix} 1 & 0.9 & c \\ 0.9 & 1 & 0.8 \\ c & 0.8 & 1 \end{bmatrix}$$

    One important fact about all such correlation matrices is that they are positive semidefinite, which means all their eigenvalues are nonnegative; since the determinant equals the product of the eigenvalues, the determinant must be nonnegative. We can obtain bounds on c from this property: $det = -c^2 + 1.44c - 0.45 \geq 0$ gives (rounded to two decimal places) $0.46 \leq c \leq 0.98$.
  - Discussion of the ideas behind this problem from a geometric perspective can be found at: `http://www.johndcook.com/blog/2010/06/17/covariance-and-law-of-cosines/`
- Ants on a circle problem. This question may or may not have actually been asked at Jane Street before, but it's very Jane Street-esque. We arrange n ants equally spaced around a circle. All ants walk at a constant speed that allows them to complete 1 revolution around the circle, or the equivalent of 1 revolution, in 1 minute. We run an experiment where we randomly set each ant to face one of the two possible directions (clockwise or counterclockwise), and then have the ants walk for 1 minute. Whenever two ants bump into each other, they both turn around and start walking in the opposite direction. At the end of the minute the ants stop and we check their positions. What is the probability that all n ants are in the same positions as they were originally?
  - First key insight: the set of the n ant positions at the end is the same as the set of the original positions. We can see this by considering a simpler variant where the ants are indistinguishable; then when two ants meet and turn around, it looks the exact same as if the two ants pass right through each other. Then we can consider the simplified experiment as all n ants walking in a full revolution uninterrupted and ending up with the same set of positions.
  - Second key insight: we have an invariant at any point in the experiment which is the total number of clockwise-moving ants. Whenever two ants collide, we transition from one clockwise and one counterclockwise ant to one counterclockwise and one clockwise ant, respectively, so the total number of clockwise-walking ants has not changed. We can reformulate this invariant as the total clockwise speed of the ants, which also is determined at the start of the experiment and stays the same throughout; this constant total clockwise speed leads to a total clockwise distance traveled across the n ants by the end of the experiment, and this is the final insight we need.
  - These two key insights are enough to solve the problem. For the n ant positions to be the same at the end, we need every ant to have traveled the same number of revolutions, whether it's 1 revolution clockwise, 1 revolution counterclockwise, or 0 revolutions. This means the total

clockwise distance traveled must be either 0, n, or $-n$ revolutions traveled, which corresponds to a total clockwise speed of 0, n, or $-n$ respectively. Then we have all n ants in their same positions at the end if and only if the start of the experiment either set all n ants clockwise, all n ants counterclockwise, or equal amounts of ants clockwise and counterclockwise. There are 2^n possible configurations of directions for the n ants, and 1 configuration each for the case of all n ants clockwise or counterclockwise. For the number of configurations with equal numbers of ants clockwise and counterclockwise, we have 0 configurations if n is odd and $\binom{n}{n/2}$ if n is even. Then our final probability is $\frac{1}{2^{n-1}}$ if n is odd, and $\frac{2+\binom{n}{n/2}}{2^n}$ if n is even.

- Ball drawing problem. This is also a random question that is Jane Street-esque. We have two bags of colored red and blue balls; bag A has 8 blue and 4 red balls, while bag B has 4 blue and 8 red balls. We pick one of the bags at random, and we draw 3 balls without replacement; we see that we drew 2 blue balls and 1 red ball. Conditioned on the result of these draws, what is the probability that we originally picked bag B?
  - This conditional probability is tractable enough to calculate by brute force, but the super-fact Bayesian statistics solution is this: in conditional probability, any subset of the provided condition (or the evidence, in more rigorous terms) that provides zero information can be deleted without changing the result. A drawing of 1 red and 1 blue ball is symmetric with respect to the choice of bag A or bag B, and therefore provides zero information. Then we can delete this subset of the drawing from the condition, and the question reduces to asking which bag we originally picked given that we drew 1 blue ball; the answer is clearly 2/3.

7.3 VIRTU FINANCIAL by Evan

- There is a 40% chance is rains on Saturday and a 70% chance it rains on Sunday; what is the probability it does not rain this weekend (what assumptions do you make?); give a lower and an upper bound on the probability it does not rain this weekend.
  - Assuming independence between the two events, the probability it does not rain is the probability it does not rain on saturday * probability it does not rain on sunday $= 0.6 * 0.3 = 0.18$
  - With no assumption, we let A = it rains on Saturday and B = it rains on Sunday. We desire bounds for $P((AUB)^c) = 1 - P(AUB) = 1 - (P(A) + P(B) - P(AB))$. Analysis of this last expression shows the $0.7 <= P(AUB) <= 1 \rightarrow 0 <= P((AUB)^c) <= 0.3$
- We distribute 20 points randomly along a circle and then choose 4 of these points at random, labeling them A, B, C, and D (e.g. choose A at random, from remaining choose B, ...). We draw the chords AB and CD; what is the probability they intersect within the circle?
  - The first thing to realize here is that the 20 points do not matter. Choosing 4 points at random from 20 points at random is the same as choosing 4 points at random (note that replacement does not matter since the probability of any individual point being chosen is 0). With this in mind, let's consider the arrangements for $ABCD$ if we read along the circle clockwise. We can write them in the order we see them (e.g. $BACD$ would occur if we first see B, then A, then C, then D walking along the circle clockwise). The key is to realize that if C (or D) separate A and B, the chords will intersect, but if both C and D do (e.g. $ACDB$) then they will not. Once you convince yourself of this, convince yourself that you are equally likely to see any of the 24 possible permutations of $ABCD$. The easiest way to do this is to see that the joint density for $ABCD$ is uniform over 4-d theta space (where theta is the angle from 12 o clock), and symmetry gives that the partitions of this space corresponding to each permutation are the same size. With these two arguments made clear, this turns into a counting problem. There are two arrangements for the 4 points for indices ($ABCD$ and $CADB$) and 2 ways to arrange the (A, B) and (C, D) in each. This leads to $2 * (2 * 2) = 8$ combinations out of $24 \rightarrow \frac{1}{3}$ for our answer.
- We have two planes, a 4-engine one and a 2-engine one. Each engine fails independently from every other one with probability p; the 4-engine plane goes down if 3 engines fail, the 2-engine plane goes down if both fail. Is either plan safer to be on outright? What is the value of p that makes them both equally safe to be on?
  - These are both binomial random variables, one with p and $n = 2$, and one with p and $n = 4$. This gives the probability the first plane fails is $\binom{4}{3} * p^3 * (1 - p) + \binom{4}{4} * p^4$ and the second is $\binom{2}{2} * p^2$. Test points of $p = \frac{1}{2}$ and $p = \frac{1}{4}$ show that there is no p for which one plane is uniformly safer to be on.
  - Setting the above probabilities equal gives $p = \frac{1}{3}$.
- We are going to play a game where you flip a coin, and you receive an amount equal to the number of flips it takes to get a head; what is a fair price to pay for this game? What if instead you get paid $2\hat{}$(# of flips to get a head)? Taking that last situation, how much would you *actually* pay?
  - In the first case, let X be the amount we receive; then we have $X = \frac{1}{2}(1) + \frac{1}{2}(X + 1)$ by conditional expectation, which we solve to give $X = 2$. Thus $X = 2$ is a fair price.
  - In the second case, we use the definition of expectation to get the our expected profit is $\sum_{i=1}^{\infty} 2^i * (\frac{1}{2})^i = \sum_{i=1}^{\infty} 1 = \infty$. Of course, we would not pay infinite money to play this game. There are many approaches to say how much you would pay. I chose to say that $p = 1/32 = 0$, and thus assume that anything with $\leq 1/32$ probability was too small to be considered. Definition of expectation allows you to arrive at an answer.
- You have a camera which can hold 1 picture in memory at all times (so if you take a new picture and there was one in memory, the one in memory is lost forever). You have a row of houses to photograph which has unknown but finite length. You are given a Uniform(0, 1) generator. Devise a strategy to photograph the houses s.t. After the last house, any house is equally likely to be in the camera's memory
  - Intuition tells us that we should consider an approach where for each house (i), we have a probability p_i of taking a picture of the house. If intuition doesn't tell you this, the interviewer will likely guide you there. The question becomes how to assign p_i when we don't even know how many houses there are. We should certainly take a picture of the first house, since if there is only one house and $p_1 = 0$, we do not have a uniform distribution over the houses. Now that we have taken a picture of the first house with $p_1 = 1$ probability, we have a p_2 probability of taking a picture of the second house and a $1 - p_2$ probability of not taking a picture of the second house. If there were only two houses, we would need both these values to be $\frac{1}{2}$, giving $p_2 = \frac{1}{2}$. Similarly reasoning tells us that the n'th house must have a $1/n$ probability of being in our camera if it is the last house, so $p_n = 1/n$.
- You have a set of numbers of size n: $\{0, 1, 2, ..., n - 1\}$. At each step, you select a number randomly from the available numbers and then reduce your set to all numbers less than that number (e.g. $n = 4$: $\{0, 1, 2, 3\}$; select 2, now you are left with $\{0, 1\}$). You then continuously repeat this until you select the number 0. What is the expected number of selections required for this to happen (e.g. $n = 4$: $\{0, 1, 2, 3\}$; select 2; $\{0, 1\}$; select 1; $\{0\}$; select $0 \rightarrow 3$ selections, the last one included).

- Let X_n be the number of selections required with the numbers $\{0, ..., n-1\}$. Let's start with $n = 1$. The answer is clearly 1. Now consider $n = 2$. There is a $\frac{1}{2}$ chance we choose 0, and a $\frac{1}{2}$ chance we choose 1. If we choose 1, we are back to the zero case, so $X_1 = \frac{1}{2} * (1) + \frac{1}{2} * (X_0 + 1) = \frac{1}{2} + \frac{1}{2} * (2) = \frac{1}{2} + 1$. Now X_3. We have a $\frac{1}{3}$ chance of choosing zero, a $\frac{1}{3}$ chance of 1, and a $\frac{1}{3}$ chance of 2, so we get $X_3 = \frac{1}{3}(1) + \frac{1}{3}(X_0 + 1) + \frac{1}{3}(X_1 + 1) = \frac{1}{3} + \frac{1}{2} + 1$. Clearly there is a pattern here. Let X_n be the number of selections required for n numbers. We desire X_{n+1}. There is a $1/(n+1)$ chance that we choose 0, and if we don't choose zero, then this problem is equivalent to the same problem on n numbers, so we get $X_{n+1} = 1/(n+1) + X_n$. Inducting with $X_0 = 1$, we get $X_n = \sum_{i=1}^{n+1} 1/i$

- How do you test the significance of a regression slope in simple linear regression?
  - T-Test (REFER TO QR DATA SCIENCE SECTION)
- Suppose you have linear regression with several covariates; how do you prevent overfitting?
  - L_p regularization (lasso, ridge, elastic)
- We want to create an algorithm which will keep a running 30-minute minimum for price data from SPX; your code will be queried for the minimum at random times, and you receive a completely random amount of price data at random times. How do you design an efficient algorithm to do this?
  - We can keep a sorted list of all recent price points within the 30-minute limit; alongside this list, we have a dictionary where each price is mapped to the time difference between its receipt and the most recent query (either calling the minimum price or adding new price data). We have a variable keeping track of the time of the most recent query, and whenever a new query is made, the time difference with the previous most recent query is calculated and added to all times in the dictionary. If any receipt time difference for any price in the list now exceeds 30 minutes, the price is removed from the list and dictionary. When we query for the minimum, we perform this update and then return the minimum in the sorted list. When we "query" by adding new data, we perform this update and then insert the new data into the list, maintaining sorting, as well as add the new data into the dictionary each initialized with receipt time difference 0.

7.4 OPTIVER by Ravi

- Mental math – square root questions
- 9 AM start time for work; leave home and arrive in 20, 40, 35, 15, 25 minutes, want to be early 95% of the time.
  - Assuming normal distribution, around when is the latest you should leave?
    - 8:15 AM
  - Would you rather reduce mean or standard deviation by a minute if you want to leave later but still be at work on time the same proportion of time?
    - Standard deviation
- Shuffle function can shuffle two letters of ABCD.
  - How many ways for 1, 2, 3, 4 letters to be out of order?
    - 0, 6, 8, 9
  - Expected value of shuffles to return back to order given that the list is out of order?
    - 2
- Multiple parts
  - Most number of singles possible with 100 AB, 0.300 BA, 0.500 SLG?
    - 23
  - How many arrangements possible if 100 AB, 0.300 BA, 0.500 SLG, no home runs?
    - 11
  - General formula for x AB?
    - x/10 + 1, as long as x is divisible by 10

7.5 AKUNA CAPITAL

- We take n i.i.d. samples from a Uniform(0,1) distribution. What is the probability that our n samples add up to at most 1?
  - This is a geometric probability question. Since each of the n terms is uniformly distributed between 0 and 1, the space of possible outcomes for the sequence of n samples is the n-dimensional hypercube, i.e. between $(0, 0, 0, ..., 0)$ and $(1, 1, 1, ..., 1)$. The subset of the possible outcomes where the n samples sum to 1 or less is the region of the hypercube below the hyperplane $x_1 + x_2 + ... + x_n = 1$, where x_1 through x_n are the coordinates. In two dimensions this is an isosceles right triangle, and in three dimensions this is the right tetrahedron with the three right-angled edges; in higher dimensions the region is the generalization of these shapes, i.e. the polytope formed by connecting the origin with all points along the coordinate axes that are 1 away from the origin. The area of the two-dimensional right triangle is $1/2$, and the volume of the three-dimensional tetrahedron is $1/6$. It is reasonable to generalize these volume formulas as $1/n!$, where n is the number of dimensions. Then the geometric probability is the volume of our polytope divided by the volume of the unit hypercube, which is 1; then the geometric probability ends up at $1/n!$.
- Using coin flips for a fair coin, we want to simulate different distributions of $1/n$ probabilities.
  - Simulate three equal probabilities of $1/3$?
    - We need two coin flips at least, which creates 4 equally likely outcomes; assign three of the four outcomes, say HH, HT, TH, to the three outputs, and the fourth outcome, say TT, is thrown out, i.e. if we get the fourth outcome we redo the coin flips.
  - In general, how many coin flips needed for n equal probabilities $1/n$?
    - We need k coin flips to produce 2^k outcomes, and we need $2^k < n$, i.e. we need at least $\log_2 n$ coin flips.
  - Suppose we only need one $1/3$ probability. How do we generate this probability with the smallest expected number of coin flips, and what is this smallest expected number of coin flips?
    - We can begin with the earlier strategy of flipping the coin twice and throwing out one of the four outcomes. We now only need one of the three valid outcomes to map to a single $1/3$ probability. Then for some of the remaining outcome space, we can combine these outcomes into just the first coin flip. To illuminate this idea, suppose we always throw out TT, and we take TH as the single valid output for the $1/3$ probability. Then the remaining outcomes are HH and HT, which we can account for already when the first coin flip is H. Our strategy is to flip the coin once, and output that we missed the $1/3$ probability if we obtain H; if the first flip is T, then we flip again, outputting the $1/3$ probability if we obtain TH and starting over our two flips if we obtain TT.
    - This strategy indeed has the smallest expected number of coin flips. We can calculate this expectation as $\mathrm{E} = 1\frac{}{2*1+\frac{1}{4}*2+\frac{1}{4}(2+E).}$

7.6 CITADEL

- 2 ants at opposite ends of an octahedron. Ant A moves to random neighboring vertex, each with probability $\frac{1}{4}$. Ant B then does the same. They keep going until an ant moves onto a vertex with the other ant on it. What is the probability that if A goes first, the game ends with A moving onto B's vertex?
  - State probabilities. Any point in the game can be described by how far apart A and B are (1 or 2 vertices) and whether A or B is moving next, so we have four probabilities $P[A1]$, $P[B1]$, $P[A2]$, $P[B2]$ each representing the probability that A wins (moves onto B) at the state described. Can make system of equations for the move transitions and also noticing that $P[A1] = 1 - P[B1]$ and $P[A2] = 1 - P[B2]$. Answer after doing out and solving these equations is $\frac{2}{5}$.
- $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 4)$. Given the PDF function of a random variable, you can look this up since it was given during the interview. Calculate $E[|Y - X|]$, i.e. the expectation of the absolute value of the difference in Y and X.
  - This is the options straddle trick; $10/\sqrt{10pi}$
  - PDF is $Y - X \sim \mathcal{N}(0, 5)$ and then the pdf of the absolute value $|Y - X|$ is just the right hand side of the normal pdf, doubled in scale to have 1 under the integral, i.e. we actually know the pdf of $|Y - X|$ and can therefore calculate its expectation by hand; remember it would be $\int_0^\infty x * f(x)dx$ where f(x) is the $|Y - X|$ pdf. This comes out to be $10/\sqrt{10pi}$.
- What is the probability that if I take 6 sets of 3 cards from a deck of 52 standard cards, exactly one of the sets will contain exactly one ace?
  - 4*(48C17)/(52C18), this is approx. 0.4
- We have 100 airplane passengers labeled 1, ..., 100, the airplane has 100 seats labeled 1, ... 100. Each passenger boards in numerical order and they're each supposed to take the seat with the same label. However passenger 1 is drunk and boards first and takes a random seat. For each subsequent passenger, their will board their labeled seat if it is free and board a random remaining seat if their seat is filled. What is the probability passenger 100 gets to sit in their own seat?
  - Notice that there is a symmetry with any sequence of passenger seatings; if passenger n has their own seat full, they have equal probability of picking seat 1 (so that the last passenger is guaranteed to sit in their own seat) vs picking seat 100 (so the last passenger cannot sit in their own seat); otherwise they pick another random seat and we continue to a future passenger who needs to pick a random seat. Because of the symmetry between random choice of seat 1 or 100 at any point the answer is $\frac{1}{2}$.
- (Systematic) How to find k maximum elements in an array of size n? Most efficient algorithm?
  - Most efficient in sorting regime is using an $O(n \log n)$ sort such as merge or quick, then querying the k elements at the start of the array.
  - Can have an answer array of size k storing the k maximum; pass through array once and then compare to sorted k array, add/replace and resort k array if necessary. $O(nk)$ for the pass, $O(k \log k)$ for keeping the array sorted; total $O(nk + k \log k)$
  - Conclusion: The single pass is better when k is small, but runtime gets high for large k so the full mergesort or quicksort is better for large k.
  - Extra solution: can build max heap in $O(n)$ and then query for max k times, each query takes $O(\log n)$ so total $O(n + k \log n)$ time
- (Systematic) How to find number of pairs of elements in array of size n that sum to some number m?
  - Can make dictionary enumerating each unique value and # of times it appears in array, in linear time. Then pass through dictionary in linear time, for each key k, query the key m-k and add the product of the values to a running sum; return answer in linear time
- We have a half hour of stock price data; how do we find the maximum profit we could have made from a single transaction? No shorting, so the transaction must be a buy followed by a sell.
  - This is a Leetcode problem that can be solved with a single pass through the data, i.e. in linear time. We do the backwards pass approach, although the forwards direction should be essentially similar. As we iterate backwards through the stock price time series, we keep track of the minimum price seen so far as well as the greatest profit possible so far (initialized to the final price and 0, respectively, when we start the pass). At every price point we see, if the current price point - minimum price point > greatest possible profit, we update the greatest possible profit; also, if the current price point is below the minimum so far we update the minimum. When we reach the start of the time series we should have our answer.
- Suppose we do simple linear regression with one input feature, i.e. the input X and output Y are vectors. We obtain a beta when we regress Y on X, and on the flip side we obtain a second beta when we regress X on Y. Can you bound the product of the two betas?
  - We use the closed form of beta here. Our two betas are $\beta = (X^T X)^{-1} X^T Y$ and $\beta' = (Y^T Y)^{-1} Y^T X$, and the product becomes $(X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T X$. Now we use some linear algebra tricks to simplify the product. $(X^T X) = |X|$ for any vector X, and also $X^T Y =< X, Y >$, the dot

product, for any two vectors X and Y. Then $\beta\beta' = \frac{<X,Y>^2}{|X||Y|}$. Notice that this is the formula for the cosine between X and Y, and cosines always only take values between -1 and 1, so we can bound the product of the betas between -1 and 1.

  - Note: in general, the inequality $< X, Y >^2 \leq |X||Y|$ is the Cauchy-Schwarz inequality, and applies to more generalized forms of the dot product known as inner products. You won't really need to know this for quant interviews, though.

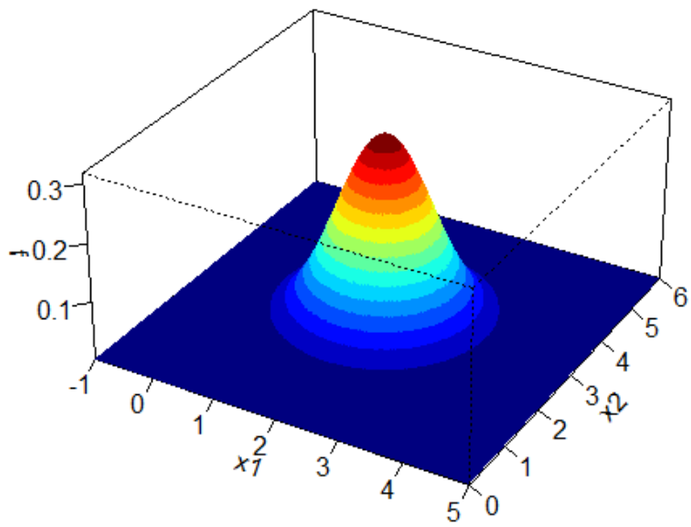
- Suppose we have a sequence of i.i.d. numbers with mean μ and variance σ^2. We replace this sequence with the sequence of subsequent differences, i.e. if the original sequence was $a_1, a_2, ..., a_n$, we now have $a_2 - a_1, a_3 - a_2, ..., a_n - a_{n-1}$. What are the (expected) mean and variance of the new sequence and how do these relate to the original μ and σ^2?

  - The sequence of differences displays a "telescoping sum", i.e. when we add up all the terms, we get a lot of cancellations and end up with $a_n - a_1$. Because our original a's were i.i.d., we expect $a_n - a_1 = 0$ so we expect the new mean to be 0.

  - Variance follows linearity, i.e. $Var(a+b) = Var(a) + Var(b)$. This implies that $Var(a_i - a_{i-1}) = 2Var(a_i)$ for all the terms in our new sequence. We can use this insight to intuit that the variance should be expected to double compared to the original, i.e we now have variance $2\sigma^2$.

7.7 HUDSON RIVER TRADING

- We have two i.i.d. standard normal variables X, Y. What is the probability that $Y > 3X$?
  - Solution 1: We want $P(Y - 3X > 0)$. Actually we can easily compute the distribution of the r.v. $Y - 3X$ as $\mathcal{N}(0, 10)$, since $3X \sim \mathcal{N}(0, 9)$, $Y \sim \mathcal{N}(0, 1)$ and sum of two normals is normal with sum of two means and sum of two variances. So clearly $P(Y - 3X > 0) = \frac{1}{2}$ for $Y - 3X$ $\mathcal{N}(0, 10)$
  - Solution 2: Looking at the joint pdf of $3X, Y$ in 3D space, it looks like a symmetrical upside-down bell shape (below). Note that this joint pdf has radial symmetry. Then $Y > 3X$ corresponds to the plane $y = 3x$ which cuts through the middle; because of radial symmetry it cuts the joint pdf exactly in half so the answer is $\frac{1}{2}$.



- We have two i.i.d. standard normal variables X, Y. What is the probability that $Y > 3X$ under the condition that X is positive?
  - Only solvable under the joint pdf idea. We still have the radially symmetric bell shape; the condition that X is positive corresponds to the half of the joint pdf on the positive side of the x-axis plane, and $P(Y > 3X|X > 0)$ is the region of the pdf on the positive side of both planes $y = 3x$ and $x = 0$. Because of the radial symmetry we only need to take the ratio of the angles these two regions are subtended by. The $x = 0$ space is subtended by angle 180 degrees (or pi) and the $y = 3x$ and $x = 0$ space is subtended by $\arctan(\frac{1}{3})$ so the answer is $\arctan(\frac{1}{3})/\pi$.
- We stand on a street and keep track of the heights of people that walk by. We note the height of the very first person who walks by; then for each subsequent person, we compare their height to the first person. What is the expected number of people that pass by after the first person before we see someone taller than the first person?
  - The expected number depends on how tall the first person was; if they were very short then next person is very likely to be taller, while if they were one of the few tallest people in the world then we will almost never find a taller person (almost infinitely expected value.) Notice that if we assume the first person is in a certain percentile of height then we have an easy solution for the expected number; if first person is p percentile (taller than proportion p of total population) then $E_p = (1 - p) + p(1 + E_p) \to E_p = 1/(1 - p)$. The percentile p is actually a uniform r.v. $[0, 1]$, so the answer expected value comes from integrating the conditional $E = 1/(1-p)$ for Uniform(0,1). So the answer is $\int_0^1 1/(1 - p)dp = \infty$
  - Note that we only used the percentile and not the actual pdf of heights; percentile is uniform [0,1] no matter what the actual height pdf is so we don't rely on any assumptions about height pdf. It is common to assume height is Gaussian but that actually doesn't matter at all.
  - Intuition: Nonzero chance of getting a first person on the extreme right tail of height (one of if not the tallest person in the world), and then infinite/almost infinite # of people need to walk by before finding someone taller. The right tail causes the expected value to diverge to infinity.
- Starting from a running sum of 0, we draw a number from the uniform distribution from 0 to 1 and add it to the running sum; what is the expected number of draws before our running sum exceeds 1?
  - Let $f(x)$ = the expected number of draws to get to x starting from 0. The answer we want is $f(1)$, and after a single draw from the uniform distribution, say drawing a number t, we have added 1 to the expected value and are now at $f(1 - t)$ expected draws left. Since t is from the uniform distribution between 0 and 1 we can quantify a single draw starting from 0 as

$$f(t) = 1 + \int_0^1 f(1 - t)dt = 1 + \int_0^1 f(t)dt.$$

  - This step requires cleverness but from looking at the equation, it seems like f is the integral of itself (i.e. $f = e^x$) so we try $f = e^x$ and notice that it actually works: $f(1) = 1 + f(1) - f(0)$ [i.e. $f(0) = 1$, true for $f = e^x$] so $f(1) = e$. Our expected number of draws is e.

7.8 TWO SIGMA

Note: Two Sigma very rarely asks brainteasers so still mostly prepare for data science, but I was asked these brainteasers in one round.

- You have several bottles of wine, where only one bottle is poisoned; need to find a strategy to test the bottles of wine on several rats to determine which is poisoned.
  - Have 1000 bottles of wine and one poisoned, 10 rats to test it on, any rat dies after one hour if they drink any amount of poison. What is the strategy?
    - Do binary encodings; label each bottle 1, 2, ..., 1000, convert each label into 10-digit binary number, feed to rat i if i-th digit is 1 and don't feed to rat i if i-th digit is 0. The pattern of rats that die will exactly correspond to one of the binary encodings which tells us which bottle is poisoned.
  - Still 10 rats and 1 hour, maximum bottles of wine we can still narrow down to one poisoned bottle?
    - 2^{10}, with the binary encodings
  - n rats
    - 2^n
  - What if we have two hours?
    - 3^n. We can do a base 3 encoding in a very similar way as the 1-hour case. Here, label each bottle $1, 2, ..., 3^n$, and convert each label into an n-digit base-3 number; if the i-th digit is 1, feed to rat i in first hour, if i-th digit is 2, feed to rat i in second hour, don't feed if 0. This ensures a unique encoding for each bottle in terms of the pattern in which the rats would die for each one.
  - What if we have k hours?
    - $(k+1)^n$. Generalization of the 2-hour case to base-$k+1$ encodings.
- We construct a sequence of die rolls as follows. For each roll, if it is 1, terminate the sequence; if it is even, add it to the sequence and keep going, and otherwise (if 3 or 5) throw out the whole sequence and start over with the sequence and die rolls. What is the expected length of such a sequence?
  - Notice that our sequence can never contain 3 and 5, i.e. we are excluding them as possible rolls in the conditional space and we are only including 1, 2, 4, 6 as possible rolls in our sequence. Then we are looking at the expected number of rolls of 1, 2, 4, 6 before we get 1. We expect to roll 1 from these 4 possibilities after 4 rolls so the expected length is 4.

7.9 FIVE RINGS

- We set up a single-elimination tournament; each round randomly pairs each team left in that round, and all teams are equally skilled so equal chance of either team in a game passing to next round. If odd number of teams in a round, one team randomly selected for a bye (automatically pass to next round). We have two teams A, B in the tournament, what is the probability they play in some game?
  - What if the tournament starts with 8 teams?
    - Can calculate with casework on what round A and B can play in and whether they both make it to that round. Answer is 2/8.
  - 9 teams?
    - More casework. 2/9
  - n teams?
    - Notice that A and B can either play in 1 game in the tournament or not play at all, so the probability is the same as the expected number of games they both play in. By linearity of expectation this is equal to the expectation that A and B play in any given game. There are $n-1$ games in a n game tournament and each game has $1/\binom{n}{2}$ chance that A and B play in that specific game, i.e. the expectation for a specific game is $1/\binom{n}{2}$. So the expectation i.e. probability is $(n-1)/\binom{n}{2} = 2/n$.
    - Important note. The core insight above that the probability of A and B playing together equals the expected number of games they play together across the tournament is unique to this context. The variable that describes whether or not A and B play a game together is an indicator variable which is 1 with probability p (the answer we're looking for) and 0 with probability $1-p$. The expected value of this indicator is therefore p. This is a very elegant trick we can generally perform with indicator variables. Keep in mind that for a lot of the other "elementary" random variables (i.e. Bernoulli, Poisson, exponential, etc.) also have expected values (means) that are simple functions of their main parameter; it's just the case that the indicator/Bernoulli variable's main parameter is a probability.
- We set up a game where we have n pairs of socks in a drawer, each pair labeled 1 through n respectively. In our game, we randomly draw 2 socks from the drawer without replacement to form a pair; the pair of drawn socks is "satisfactory" if their labels are at most 1 apart, i.e. drawing a $(2,2)$ or $(1,2)$ is satisfactory but a $(1,3)$ is not. We lose the game if we draw any unsatisfactory pair; we win if we draw all socks from the drawer, i.e. draw a pair n times, and get a satisfactory pair each time.
  - First examine the small case of $n = 3$. What is the probability that we win the game?
    - From the definition of satisfactory, we win as long as we never draw a $(1,3)$ pair. The different configurations of drawing 3 pairs is tractable to hand-calculate in a few minutes; the answer turns out to be 2/3.
  - General case of n. How do we find the probability of winning the game?
    - First key insight. Imagine drawing n random pairs of socks first, and then checking if we won or lost after the fact, i.e. if all n pairs are satisfactory then we won and if at least one pair is unsatisfactory then we lose. This is identical to the original setup in terms of win probability. The difference here is any ordering of the n random pairs is the same when we check after the fact. In other words we can consider our drawn socks in any order, and since we need to draw all socks to win (or to finish the game at all in our modified setup), we can condition on the outcome of some specific pair... maybe some edge case?
    - Second key insight. When we draw a satisfactory pair whose labels are 1 apart, we need to draw that same pair again or else we can't win. Intuitively, when we draw a 1-apart satisfactory pair, say $(k, k+1)$ for some k, we have one remaining k sock and one remaining $k+1$ sock each of which need to be paired somehow. If k pairs with $k-1$ then later the other $k-1$ must be paired, and so on; this inductively reduces to a single 1 sock having no possible satisfactory pair so we lose. The argument is symmetric for the $k+1$ sock. Then another $(k, k+1)$ pair must be drawn.
    - The two key insights lead to a recursion solution. Let P_n be the win probability for the game with n socks. Any winning game must draw pairs with the label-1 socks. One possibility is that a single $(1,1)$ pair is drawn; this occurs with probability $\frac{1}{2n-1}$, and the rest of the draws are governed by the game with $n-1$ socks, i.e. the win probability for this case is P_{n-1}. The other possibility is that two $(1,2)$ pairs are drawn; this occurs with probability $\frac{2}{(2n-1)(2n-3)}$, and the rest of the draws are governed by P_{n-2}. The entire recursion looks like $P_n = \frac{1}{2n-1}P_{n-1} + \frac{2}{(2n-1)(2n-3)}P_{n-2}$. (This can be solved for a closed form for P, but it's just tedious algebra, and Five Rings expected just the recursion idea.)
- You and the interviewer bet on opposite sides on the $n = 3$ sock drawer game from above. The interviewer allows you to bet with 1 : 1 odds, i.e. if the sock drawer game wins then you double your money (and if the sock drawer game loses then you lose what you betted). If you start with 100 dollars, how much money do you bet on this game?
  - First calculate the expected return of the game.

    - Earlier we calculated that the win probability for the $n = 3$ sock drawer game is $2/3$. When we bet 1 dollar, our outcomes are 0 dollars with probability $1/3$, and 2 dollars with probability $2/3$, i.e. the expected outcome is $4/3$, and the expected return of the game is $1/3$.

- Suppose we only do one round of betting. How much money do you bet?
    - No exactly right or wrong answer; the interviewer is looking for how you manage risk relative to your total wealth. For this kind of question the number you say isn't any form of math problem solving, but probability illuminates your risk adversity.

- Suppose we can do infinite rounds of betting. Can you formulate a betting strategy?
    - This is done using the Kelly criterion, a very useful trade sizing framework in industry. Feel free to search up Kelly criterion on Wikipedia; the core ideas are simple to remember, and although quants will very rarely expect you to know Kelly criterion in interviews, it's a good trick to have up your sleeve.
    - The Kelly criterion tells you what fraction of your investing wealth you should bet when you know the expected return. For a simple betting situation where you lose a fraction a of your original investment with probability q and gain a fraction b of your original investment with probability p, you should invest a fraction of your total wealth equal to $f = \frac{p}{a} - \frac{q}{b}$. Applied to the sock game, we have $a = b = 1$, $p = 2/3$, $q = 1/3$, so $f = 1/3$, i.e. we should bet $1/3$ of our total wealth each round to maximize our expected wealth after all rounds we play.

7.10 SIG by Ravi

- Paint the outside of a $3 * 3 * 3$ cube red. Cut it into 27 smaller $1 * 1 * 1$ cubes. Randomly pick a cube and roll it. The face is red. What is the probability that the cube we rolled was from the corner of the bigger cube?
  - There are 54 total red faces (total space of our conditional probability) and 8 corner cubes * 3 red faces on each corner cube = 24 red faces corresponding to the event that the cube is a corner. The answer is 24/54 = 4/9.
- Opponent has 2 dollars, you have 1 dollar. You play a game where you each wager 1 dollar. Opponent has 1/3 chance of winning each game, you have 2/3 chance of winning each game. You play until someone is bankrupt, and then they lose. What is the probability you win the game?
  - Let $P[x]$ = probability you win given you have x dollars right now. We have four possible states $P[0], P[1], P[2], P[3]$ and clearly $P[0] = 1$, $P[3] = 0$. For the rest we get a system of equations $P[1] = \frac{1}{3} * P[2] + \frac{2}{3} * P[0]$ and $P[2] = \frac{1}{3} * P[3] + \frac{2}{3} * P[1]$. Solving the system gives $P[1] = 4/7$.
- Analyst report says that stock will be taken over with probability 50%. If it's not taken over, it'll move to 5 dollars with 40% chance, and to 0 dollars with 60% chance. If it's taken over, it'll move to 15 dollars with 80% chance, and x with 20% chance. The stock is trading at 10 dollars. How much would you pay, assuming no time value of money, for the option to buy the stock at 22 dollars after everything settles? This means you get to see if it's taken over or not, and then get to see where it goes after that.
  - $0.80