MATH 341: TAKEAWAYS

STEVEN J. MILLER

ABSTRACT. Below we summarize some items to take away from the class (as well as previous classes!). In particular, what are one time tricks and methods, and what are general techniques to solve a variety of problems, as well as what have we used from various classes. Comments and additions welcome!

1. CALCULUS I AND II (MATH 103 AND 104)

We used a variety of results and techniques from 103 and 104:

(1) **Standard integration theory:** For us, the most important technique is integration by parts; one of many places we used this was in computing the moments of the Gaussian. Integration by parts is a very powerful technique, and is frequently used. While most of the time it is clear how to choose the functions u and dv, sometimes we need to be a bit clever. For example, consider the second moment of the standard normal: $(2\pi)^{-1/2}\int_{-\infty}^{\infty} x^2 \exp(-x^2/2)dx$. The natural choices are to take $u = x^2$ or $u = \exp(-x^2/2)$, but neither of these work as they lead to choices for dv that do not have a closed form integral. What we need to do is split the two 'natural' functions up, and let $u = x$ and $dv = \exp(-x^2/2)xdx$. The reason is that while there is no closed form expression for the anti-derivative of the standard normal, once we have xdx instead of dx then we can obtain nice integrals. One final remark on integrating by parts: it is a key ingredient in the 'Bring it over' method (which will be discussed below).

(2) **Definition of the derivative:** Recall

$$f'(x) \ = \ \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}.$$

In upper level classes, the definition of the derivative is particularly useful when there is a split in the definition of a function. For example, consider

$$f(x) \ = \ \begin{cases} \exp(-1/x^2) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

This function has all derivatives zero at $x = 0$, but is non-zero for $x \neq 0$. Thus the Taylor series does not converge in a neighborhood of positive length containing the origin. This function shows how different real analysis is from complex analysis. Explicitly, here we have an infinitely differentiable function which is not equal to its Taylor series in a neighborhood of $x = 0$; if a complex function is differentiable once it is infinitely differentiable and it equals its derivative in a neighborhood of that point.

Date: December 20, 2009.

(3) **Ratio, root and comparison tests:** These are used to determine if a series or integral converges. We frequently used the geometric series formula $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ if $|x| < 1$.

(4) **Taylor series:** Taylor expansions are very useful, allowing us to replace complicated functions (locally) by simpler ones. The moment generating function of a random variable is a Taylor series whose coefficients are the moments of the distribution. Another instance where we used this is in proving the Central Limit Theorem. The moment generating function of a sum of independent random variables is the product of the moment generating functions. To study a product, we summify it (we'll discuss this technique in much greater detail below). Thus we need to expand $\log\left(M_X(t)^n\right) = n \log M_X(t)$. As $M_X(0) = 1$, we for small t we just need to understand the expansion of $\log(1+u)$.

Taylor's Theorem: *If f is differentiable at least $n+1$ times on $[a,b]$, then for all $x \in [a,b]$, $f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$ plus an error that is at most $\max_{a \le c \le x} |f^{(n+1)}(c)| \cdot |x-a|^{n+1}$.*

(5) **L'Hopital's Rule:** This is one of the most useful ways to compare growth rates of different functions. It works for ratios of differentiable functions such that either both tend to zero or both tend to $\pm\infty$. We used this in class to see that, as $x \to \infty$, $(\log x)^A \ll x^B \le e^x$ for any $A, B > 0$. (Recall $f(x) \ll g(x)$ means there is some C such that for all x sufficiently large, $|f(x)| \le Cg(x)$.) We also used L'Hopital to take the derivatives of the troublesome function $h(x) = \exp(-1/x^2)$ for $x \neq 0$ and 0 otherwise (this function is the key to why real analysis is so much harder than complex analysis).

2. Multivariable Calculus (Math 105/106)

(1) **Fubini Theorem (or Fubini-Tonelli):** Frequently we want to / need to justify interchanging two integrals (or an integral and a sum). Doing such interchanges is one of the most frequent tricks in mathematics; whenever you see a double sum, a double integral, or a sum and an integral you should consider this. While we cannot always interchange orders, we can if the double sum (or double integral) of the absolute value of the summand (or the integrand) is finite. For example,

$$\begin{aligned}
\int_{y=0}^{1} \left[\int_{x=0}^{1} e^{-xy} x dx \right] dy &= \int_{x=0}^{1} \left[\int_{y=0}^{1} e^{-xy} x dy \right] dx \\
&= \int_{x=0}^{1} e^{-xy} \Big|_{1}^{0} dx \\
&= \int_{x=0}^{1} \left(1 - e^{-x}\right) dx \ = \ 2 - e^{-x}. \qquad (2.1)
\end{aligned}$$

Note how much easier it is when we integrate with respect to y first – we bypass having to use Integration by Parts. For completeness, we state:

Fubini's Theorem: *Assume f is continuous and*

$$\int_a^b \int_c^d |f(x,y)| dx dy \ < \ \infty. \tag{2.2}$$

Then

$$\int_a^b \left[\int_c^d f(x,y) dy\right] dx \ = \ \int_c^d \left[\int_a^b f(x,y) dx\right] dy. \tag{2.3}$$

Similar statements hold if we instead have

$$\sum_{n=N_0}^{N_1} \int_c^d f(x_n, y) dy, \qquad \sum_{n=N_0}^{N_1} \sum_{m=M_0}^{M_1} f(x_n, y_m). \tag{2.4}$$

(2) ***Whenever you have a theorem, you should always explore what happens if you remove a condition. Frequently (though not always) the claim no longer holds; sometimes the claim is still true but the proof is harder. Rarely, but it can happen, removing a condition causes you to look at a problem in a new light, and find a simpler proof.*** We apply this principle to Fubini's theorem; specifically, we remove the finiteness condition and construct a counter-example.

For simplicity, we give a sequence a_{mn} such that $\sum_m(\sum_n a_{m,n}) \neq \sum_n(\sum_m a_{m,n})$. For $m, n \geq 0$ let

$$a_{m,n} \ = \ \begin{cases} 1 & \text{if } n = m \\ -1 & \text{if } n = m+1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

We can show that the two different orders of summation yield different answers; if we sum over the columns first we get 0 for each column, and then doing the sum of the column sums gives 0; however, if we do the row sums first, than all the row sums vanish but the first (which is 1), and hence the sum of the row sums is 1, *not* 0. The reason for this difference is that the sum of the absolute value of the terms diverges.

(3) **Interchanging derivatives and sums:** It is frequently useful to interchange a derivative and an infinite sum. The first place this is met is in proving the derivative of e^x is e^x; using the series expansion for e^x, it is trivial to find the derivative *if* we can differentiate term by term and then add.

Interchanging differentiation and integration: *Let $f(x,t)$ and $\partial f(x,t)/\partial x$ be continuous on a rectangle $[x_0, x_1] \times [t_0, t_1]$ with $[a,b] \subset [t_0, t_1]$. Then*

$$\frac{d}{dx} \int_{t=a}^b f(x,t) dt \ = \ \int_{t=a}^b \frac{\partial f}{\partial x}(x,t) dt. \tag{2.6}$$

Frequently one wants to interchange differentiation and summation; this leads to the method of differentiating identities, which is extremely useful in computing moments of probability distributions. For example, consider the identity

$$(p+q)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}. \tag{2.7}$$

Applying the operator $p\frac{d}{dp}$ to both sides we find

$$p \cdot n(p+q)^{n-1} \;=\; \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}. \tag{2.8}$$

Setting $q = 1-p$ yields the mean of a binomial random variable:

$$np \;=\; \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.9}$$

It is very important that initially p and q are distinct, free variables, and only at the end do we set $q = 1-p$.

(4) **Dangers when interchanging:** One has to be very careful in interchanging operations. Consider, for example, the family of probability densities $f_n(x)$, where f_n is a triangular density on $[1/n, 3/n]$ with midpoint (i.e., maximum value) n. While each f_n is continuous (as is the limit $f(x)$, which is identically 0), each f_n is a probability density (as each integrates to 1); however, the limit density is identically 0, and thus not a density! We can easily modify our example so that the limit is not continuous:

$$g_n(x) \;=\; \begin{cases} n|x| & \text{if } 0 \le |x| \le 1/n \\ 1 & \text{if } 1/n \le |x| \le 1/2 \\ n\left(\frac{1}{2} + \frac{1}{n} - |x|\right) & \text{if } 1/2 \le x \le 1/2 + 1/n \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Note that $g_n(0) = 0$ for all n, but as we approach 0 from above or below, in the limit we get 1.

(5) **Change of Variables Theorem:** *Let V and W be bounded open sets in $\mathbb{R}^n$. Let $h : V \to W$ be a 1-1 and onto map, given by*

$$h(u_1, \dots, u_n) \;=\; (h_1(u_1, \dots, u_n), \dots, h_n(u_1, \dots, u_n)). \tag{2.11}$$

Let $f : W \to \mathbb{R}$ be a continuous, bounded function. Then

$$\begin{aligned} &\int \cdots \int_W f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &\quad = \int \cdots \int_V f\left(h(u_1, \dots, u_n)\right) |J(u_1, \dots, u_v)| du_1 \cdots du_n, \end{aligned} \tag{2.12}$$

where J is the ***Jacobian***

$$J \;=\; \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \cdots & \frac{\partial h_n}{\partial u_n} \end{vmatrix}. \tag{2.13}$$

We used this result to simplify the algebra in many problems by passing to an easier set of variables.

3. Differential Equations (Math 209)

(1) **The method of Divine Inspiration and Difference Equations:** Difference equations, such as the Fibonacci equation $a_{n+1} = a_{n+1} + a_n$, arise throughout nature. There is a rich theory when we have linear recurrence relations. To find a solution, we 'guess' that $a_n = r^n$ and take linear combinations.

Specifically, let k be a fixed integer and $c_1, \dots, c_k$ given real numbers. Then the general solution of the difference equation

$$a_{n+1} \ = \ c_1 a_n + c_2 a_{n-1} + c_3 a_{n-2} + \cdots + c_k a_{n-k+1}$$

is

$$a_n \ = \ \gamma_1 r_1^n + \cdots + \gamma_k r_k^n$$

if the characteristic polynomial

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - \cdots - c_k \ = \ 0$$

has k distinct roots. Here the $\gamma_1, \dots, \gamma_k$ are any k real numbers; if initial conditions are given, these conditions determine these γ_i's. If there are repeated roots, we add terms such as $nr^n, \dots, n^{m-1}r^n$, where m is the multiplicity of the root r.

For example, consider the equation $a_{n+1} = 5a_n - 6a_{n-1}$. In this case $k = 2$ and we find the characteristic polynomial is $r^2 - 5r + 6 = (r-2)(r-3)$, which clearly has roots $r_1 = 2$ and $r_2 = 3$. Thus the general solution is $a_n = \gamma_1 2^n + \gamma_2 3^n$. If we are given $a_0 = 1$ and $a_1 = 2$, this leads to the system of equations $1 = \gamma_1 + \gamma_2$ and $2 = \gamma_1 \cdot 2 + \gamma_2 \cdot 3$, which has the solution $\gamma_1 = 1$ and $\gamma_2 = 0$.

Applications include population growth (such as the Fibonacci equation) and why double-plus-one is a bad strategy in roulette.

4. Analysis (Math 301)

(1) **Continuity:** General continuity properties, in particular some of the $\epsilon - \delta$ arguments to bound quantities, are frequently used to prove results. Often we use these to study moments or other properties of densities. Most important, however, was probably when we can interchange operations, typically interchanging integrals, sums, or an infinite sum and a derivative. For the derivative of the geometric series, this can be done by noting the tail is another geometric series; in general this is proved by estimating the contribution from the tail of the sum). See the multivariable calculus section for more comments on these subjects.

(2) **Proofs by Induction:** Induction is a terrific way to prove formulas for general n *if* we have a conjecture as to what the answer should be. Assume for each positive integer n we have a statement $P(n)$ which we desire to show is true for all n. $P(n)$ is true for all positive integers n if the following two statements hold: (i) **Basis Step:** $P(1)$ is true; (ii) **Inductive Step**: whenever $P(n)$ is true, $P(n+1)$ is true. Such proofs are called proofs by induction or induction (or inductive) proofs.

The standard examples are to show results such as $\sum_{k=0}^n k = \frac{n(n+1)}{2}$. It turns out that $\sum_{k=0}^n k^m$ is a polynomial in n of degree $m+1$ with leading coefficient $1/(m+1)$ (one can see that this is reasonable by using the integral test to replace the sum with an integral); however, the remaining coefficients of the polynomial are harder to find, and without them it is quite hard to run the induction argument for say $m = 2009$.

(3) **Dirichlet's Pigeonhole principle:** Let $A_1, A_2, \dots, A_n$ be a collection of sets with the property that $A_1 \cup \cdots \cup A_n$ has at least $n+1$ elements. Then at least one of the sets A_i has at least two elements. We frequently use the Pigeonhole principle to ensure that some event happens.

5. Probablity Theory (Math 341)

5.1. Combinatorics.

(1) **Combinatorics:** There are several items to remember for combinatorial problems. The first is to be careful and avoid double counting. The second is that frequently a difficult sum can be interpreted two different ways; one of the interpretations is what we want, while the other is something we can do. We have seen many examples of this. One is that

$$\sum_{k=0}^n \binom{n}{k}^2 = \sum_{k=0}^n \binom{n}{k}\binom{n}{n-k}$$

is the middle coefficient of $(x+y)^{2n}$, and thus equals $\binom{2n}{n}$.

(2) **'Auxiliary lines':** In geometry, one frequently encounters proofs where the authors add an auxiliary line not originally in the picture; once the line is added things are clear, but it is often a bit of a mystery as to how someone would think of adding a line in that place. In combinatorics we have an analogue of this. Consider the classic cookie problem: we wish to divide 10 identical cookies among 5 distinct people. One simple way to do this is to imagine we have 14 ($14 = 10 + 5 - 1$) cookies, and eat 4 of them. This partitions the remaining cookies into 5 sets, with the first set going to the first person and so on.

For example, if we have 10 cookies and 5 people, say we choose cookies 3, 4, 7 and 13 of the $10 + 5 - 1$ cookies:

$$\odot \ \odot \ \otimes \ \otimes \ \odot \ \odot \ \otimes \ \odot \ \odot \ \odot \ \odot \ \odot \ \otimes \ \odot$$

This corresponds to person 1 receiving two cookies, person 2 receiving zero, person 3 receiving two, person 4 receiving five and person 5 receiving one cookie.

This implies that the answer to our problem is $\binom{10+5-1}{5-1}$, or in general $\binom{C+P-1}{P-1}$.

(3) **Find an interpretation:** Consider the following sum: $\sum_{c=0}^C \binom{c+P-1}{P-1}$. By the arguments above, we are summing the number of ways of dividing c cookies among P people for $c \in \{0, \dots, C\}$ (or we divide C cookies among P people, but we do not assume each cookie is given). A nice way to solve this is to imagine that there is a $P + 1^{\text{st}}$ person who receives $C - c$ cookies, in which case this sum is now the same as counting the number of ways of dividing C cookies among $P + 1$ people where each cookie must be assigned to a person, or $\binom{C+P}{P}$. (See also the 'tell a story' entry in §5.2 and the 'convolution' entry in

§5.3.)

(4) **Inclusion - Exclusion Principle:** Suppose $A_1, A_2, \dots, A_n$ is a collection of sets. Then the *Inclusion-Exclusion Principle* asserts that

$$\left|\bigcup_{i=1}^{n} A_i\right| \ = \ \sum_i |A_i| - \sum_{i,j} |A_i \cap A_j| + \sum_{i,j,k} |A_i \cap A_j \cap A_k| - \cdots .$$

This has many uses for counting probabilities. We used it to determine the probability of a generic integer is square-free, as well as the probability a random permutation of $\{1, \dots, n\}$ returns at least one element to its initial location.

(5) **Binomial Theorem:** We have

$$(x+y)^n \ = \ \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \ = \ \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k;$$

in probability we usually take $x = p$ and $y = 1-p$. The coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ have the interpretation as counting the number of ways of choosing k objects from n when order does not matter. A better definition of this coefficient is

$$\binom{n}{k} \ = \ \frac{n(n-1)\cdots(n-(k-1))}{k(k-1)\cdots 1}.$$

The reason this definition is superior is that $\binom{3}{5}$ makes sense with this definition, and is just zero. One can easily show $\binom{n}{k} = 0$ whenever $k > n$, which makes sense with our combinatorial interpretation: there is no way to choose k objects from n when $n < k$, regardless of whether or not order matters.

5.2. General Techniques of Probability.

(1) **Differentiating Identities:** Equalities are the bread and butter of mathematics; differentiating identities allows us to generate infinitely many more from one, which is a very good deal! For example, consider the identity

$$(p+q)^n \ = \ \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}. \tag{5.1}$$

Applying the operator $p\frac{d}{dp}$ to both sides we find

$$p \cdot n(p+q)^{n-1} \ = \ \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k}. \tag{5.2}$$

Setting $q = 1-p$ yields the mean of a binomial random variable:

$$np \ = \ \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k}. \tag{5.3}$$

It is very important that initially p and q are distinct, free variables, and only at the end do we set $q = 1-p$. Another example is differentiating $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ by applying the operator $x\frac{d}{dx}$ gives $\sum_{n=0}^{\infty} nx^n \ = \ x/(1-x)^2$. While we can prove the $2m^{\text{th}}$ moment of the standard normal is $(2m-1)!!$ by induction, we can also do this with differentiating identities.

(2) **Law of Total Probability:** This is perhaps one of the most useful observations: $\text{Prob}(A^c) = 1 - \text{Prob}(A)$, where A^c is the complementary event. It is frequently easier to compute the probability that something does not happen than the probability it does. Standard examples include hands of bridge or other card games.

(3) **Fundamental Theorem of Calculus (cumulative distribution functions and densities):** One of the most important uses of the Fundament Theorem of Calculus is the relationship between the cumulative distribution function F_X of a random variable X and its density f_X. We have

$$F_X(x) \ = \ \text{Prob}(X \le x) \ = \ \int_{-\infty}^{x} f_X(t)dt.$$

In particular, the Fundamental Theorem of Calculus implies that $F_X'(x) = f_X(x)$. This means that if we know the cumulative distribution function, we can essentially deduce the density. For example, let X have the standard exponential density (so $f_X(x) = e^{-x}$ for $x \ge 0$ and 0 otherwise) and set $Y = X^2$. Then for $y \ge 0$ we have

$$F_Y(y) \ = \ \text{Prob}(Y \le y) \ = \ \text{Prob}(X^2 \le y) \ = \ \text{Prob}(X \le \sqrt{y}) \ = \ F_X(\sqrt{y}).$$

We now differentiate, using the Fundamental Theorem of Calculus and the Chain Rule, and find that for $y \ge 0$

$$f_Y(y) \ = \ F_X'(\sqrt{y}) \cdot \frac{d}{dy}(\sqrt{y}) \ = \ f_x(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \ = \ \frac{e^{-\sqrt{y}}}{2\sqrt{y}}.$$

(4) **Binary (or indicator) random variables:** For many problems, it is convenient to define a random variable to be 1 if the event of interest happens and 0 otherwise. This frequently allows us to reduce a complicated problem to many simpler problems. For example, consider a binomial process with parameters n and p. We may view this as flipping a coin with probability p of heads a total of n times, and recording the number of heads. We may let $X_i = 1$ if the i^{th} toss is heads and 0 otherwise; then the total number of heads is $X = X_1 + \cdots + X_n$. In other words, we have represented a binomial random variable with parameters n and p as a sum of n independent Bernoulli random variables. This facilitates calculating quantities such as the mean or variance, as we now have $\mathbb{E}[X] = n\mathbb{E}[X_i] = np$ and $\text{Var}(X) = n\text{Var}(X_i) = np(1-p)$. Explicitly, to compute the mean we need to evaluate $\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p)$ and then multiply by n; this is significantly easier than directly evaluating the mean of the binomial random variable, which requires us to determine $\sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}$.

(5) **Linearity of Expectation:** One of the worst complications in probability is that random variables might not be independent. This greatly complicates the analysis in a variety of cases; however, if all we care about is the expected value, these difficulties can vanish! The reason is that the expected values of a sum is the sum of the expected values; explicitly, if $X = X_1 + \cdots + X_n$ then $\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$. One great example of this was in the coupon or prize problem. Imagine we have c different prizes, and each day we are randomly given one and only of the c prizes. We assume the choice of prize is independent

of what we have, with each prize being chosen with probability $1/c$. How long will it take to have one of each prize? If we let X_i denote the random variable which is how long we must wait, given $i-1$ prizes, until we obtain the next new prize, then X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{c}$ and expected value $\frac{1}{p_i} = \frac{c}{c-(i-1)}$. Thus the expected number of days we must wait until we have one of each prize is simply

$$\mathbb{E}[X] \ = \ \sum_{i=1}^{c-1} \mathbb{E}[X_i] \ = \ \sum_{i=1}^{c-1} \frac{c}{c-(i-1)} \ = \ c\sum_{i=1}^{c} \frac{1}{i} \ = \ cH_c,$$

where $H_c = 1/1 + 1/2 + \cdots + 1/c$ is the c^{th} harmonic number (and $H_c \approx \log c$ for c large). Note we do not need to consider elaborate combinations or how the prizes are awarded. Of course, if we want to compute the variance or the median, it's a different story and we can't just use linearity of expectation.

(6) **Bring it Over:** We have seen two different applications of this method. One is in evaluating integrals. Let I be a complicated integral. What often happens is that, after some number of integration by parts, we obtain an expression of the form $I = a + bI$; so long as $b \neq 1$ we can rewrite this as $(1-b)I = a$ and then solve for I ($I = \frac{a}{1-b}$). This frequently occurs for integrals involving sines and cosines, as two derivatives (or integrals) basically returns us to our starting point. We also saw applications of this in memoryless games, to be described below.

(7) **Memoryless games / processes:** There are many situations where to analyze future behavior, we do not need to know how we got to a given state or configuration, but rather just what the current game state is. A terrific example is playing basketball, with the first person to make a basket winning. Say A shoots first and always gets a basket with probability p, and B shoots second and always makes a basket with probability q. A and B keep shooting, A then B then A then B and so on, until someone makes a basket. What is the probability A wins? The long was is to note that the probability A wins on her n^{th} shot is $((1-p)(1-q))^{n-1}\, p$, and thus

$$\text{Prob}(A \text{ wins}) \ = \ \sum_{n=0}^{\infty} ((1-p)(1-q))^{n-1}\, p;$$

while we can evaluate this with the geometric series, there is an easier way. How can A win? She can win by making her first basket, which happens with probability p. If she misses, then to win she needs B to miss as well. At this point, it is A's turn to shoot again, and it is as if we've just started the game. It does not matter that both have missed! Thus

$$\text{Prob}(A \text{ wins}) \ = \ p + (1-p)(1-q)\text{Prob}(A \text{ wins}).$$

Note this is exactly the set-up for using 'Bring it over', and we find

$$\text{Prob}(A \text{ wins}) \ = \ \frac{p}{1-(1-p)(1-q)};$$

in fact, we can use this to provide a proof of the geometric series formula! The key idea here is that once both miss, it is as if we've just started the game. This is a very fruitful way of looking at many problems.

(8) **Standardization:** Given a random variable X with finite mean and variance, it is almost always a good idea to consider the standardized random variable $Y = (X - \mathbb{E}[X])/\mathrm{StDev}(X)$, especially if X is a sum of independent random variables. The reason is that Y now has mean 0 and variance 1, and this sets us up to compare quantities on the same scale. Equivalently, when we discuss the Central Limit Theorem everything will converge to the same distribution, a standard normal. We thus will only need to tabulate the probabilities for one normal, and not a plethora or even an infinitude. The situation is similar to logarithm tables. We only need to know logarithms in one base to know them in all, as the Change of Base formula gives $\log_c x = \log_b x / \log_b c$ (and thus if we know logarithms in base b, we know then in base c).

(9) **Tell a story:** One of our exam questions was whether or not $f(n) = \binom{n+k-1}{n}(1-p)^n p^k$ for $n \in \{0, 1, 2, \dots\}$, $p \in (0,1)$ is a probability mass function. One way to approach a problem like this is to try and tell a story. How should we interpret the factors? Well, let's make p the probability of getting a head when we toss a coin, or we could let it denote the probability of a success. Then $(1-p)^n p^k$ is the probability of a string with exactly n failures and k successes. There are $\binom{n}{k}$ ways to choose which n of $n+k$ places to be the failures; however, we have $\binom{n+k-1}{n}$. What's going on? The difference is that we are not considering all possible strings, but only strings where the *last* event is a success. Thus we must have exactly n failures (or exactly $k-1$ successes) in the first $n+k-1$ tosses followed by a success on trial $n+k$. By finding a story like this, we know it is a probability mass function; it is possible to directly sum this, but that is significantly harder. (See also the 'find an interpretation' entry in §5.1 and the 'convolution' entry in §5.3.)

(10) **Probabilistic Models:** We can often gain intuition about complex but deterministic phenomena by employing a random model. For example, the Prime Number Theorem tells us that there are about $x/\log x$ primes at most x, leading to the estimation that any n is prime with probability about $1/\log n$ (this is known as the Cramer model). Using this, we can estimate various number theoretic quantities. For example, let X_n be a random binary indicator variable which is 1 with probability $\frac{1}{\log n}$ and 0 with probability $1 - \frac{1}{\log n}$. If we want to estimate how many numbers up to x start a twin prime pair (i.e., n and $n+2$ are both prime) then the answer would be given by the random variable $X = X_2X_4 + X_3X_5 + \cdots + X_{n-2}X_n$. As everything is independent and $\mathbb{E}[X_k] = \frac{1}{\log k}$, we have

$$\mathbb{E}[X] \;=\; \sum_{k=2}^{n-2} \mathbb{E}[X_k]\mathbb{E}[X_{k+2}] \;=\; \sum_{k=2}^{n-2} \frac{1}{\log(k)\log(k+2)} \;\approx\; \int_2^{n-2} \frac{dt}{\log^2 t} \;\approx\; \frac{x}{\log^2 x}.$$

The actual (conjectured!) answer is about $C_2 x/\log^2 x$, where

$$C_2 \;=\; \prod_{\substack{p \ge 3 \\ p \text{ prime}}} \frac{p(p-2)}{(p-1)^2} \;\approx\; .66016.$$

What's important is to note that the simple heuristic *did* capture the correct x dependence, namely a constant times $x/\log^2 x$. Of course, one must be very careful about how far one pushes and trusts these models. For example, it would predict there are about $C_3 x/\log^3 x$ prime triples $(n, n+2, n+4)$ up to x for some non-zero C_3, whereas in actuality there is only the triple $(3, 5, 7)$! The problem is this model misses arithmetic, and in any three consecutive odd numbers exactly one of them is divisible by 3.

(11) **Simplifying sums:** Often we encounter a sum which is related to a standard sum; this is particularly true in trying to evaluate moment generation functions. Some of the more common (and important) identities are

$$\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!} \\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots = \sum_{n=0}^{\infty} x^n \\
\frac{1}{(1-x)^2} &= 1 + 2x + 3x^3 + 4x^3 = \sum_{n=0}^{\infty} \binom{n}{1} x^{n-1} \\
\frac{1}{(1-x)^k} &= \sum_{n=0}^{\infty} \binom{n}{k} x^{n-k} \\
(x+y)^n &= x^n + nx^{n-1}y + \frac{n(n-1)}{2} x^{n-2} y^2 \\
&= \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.
\end{aligned}$$

The goal is to 'see' a complicated expression is one of the above (for a special choice of x). For example, let X be a Poisson with parameter λ; thus $f_X(n) = x\lambda^n e^{-n}/n!$ if $n \in \{0, 1, 2, \dots\}$ and 0 otherwise. Then

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} e^{tn} \cdot \frac{\lambda^n e^{-\lambda}}{n!}.$$

Fortunately, this looks like one of the expressions above, namely the one for e^x. Rearranging a bit gives

$$M_X(t) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} \cdot \exp\left(\lambda e^t\right) = \exp\left(\lambda e^t - \lambda\right).$$

5.3. Moments.

(1) **Convolution:** Let X and Y be independent random variables with densities f_X and f_Y. Then the density of $X + Y$ is

$$f_{X+Y}(u) = (f_X * f_Y)(u) := \int_{-\infty}^{\infty} f_X(u) f_Y(t-u) du;$$

we call $f_X * f_Y$ the convolution of X and Y. While we can prove by brute force that $f_X * f_Y = f_Y * f_X$, a faster interpretation is obtained by noting that since addition is commutative, $X + Y = Y + X$ and hence $f_{X+Y} = f_{Y+X}$, which implies convolution is commutative. Convolutions give us a handle on the density for sums of independent random variables, and is a key ingredient in the proof of the Central Limit Theorem.

(2) **Generating Functions:** Given a sequence $\{a_n\}_{n=0}^{\infty}$, we define its generating function by

$$G_a(s) \ = \ \sum_{n=0}^{\infty} a_n s^n$$

for all s where the sum converges. For discrete random variables that take on values at the non-negative integers, an excellent choice is to take $a_n = \text{Prob}(X = n)$, and the result is called the generating function of the random variable X. Using convolutions, we find that if X_1 and X_2 be *independent* discrete random variables taking on non-negative integer values, with corresponding probability generating functions $G_{X_1}(s)$ and $G_{X_2}(s)$, then $G_{X_1+X_2}(s) = G_{X_1}(s)G_{X_2}(s)$.

(3) **Moment Generating Functions:** For many probability problems, the moment generating function $M_X(t)$ is more convenient to study than the generating function. It is defined by $M_X(t) = \mathbb{E}[e^{tX}]$, which implies (if everything converges!) that

$$M_X(t) \ = \ 1 + \mu_1' t + \frac{\mu_2' t^2}{2!} + \frac{\mu_3' t^3}{3!} + \cdots,$$

where $\mu_k' \ = \ d^k M_X(t)/dt^k \Big|_{t=0}$ is the k^{th} moment of X. Key properties of the moment generating function are: (i) Let α and β be constants. Then

$$M_{\alpha X+\beta}(t) \ = \ e^{\beta t} M_X(\alpha t).$$

(ii) if $X_1, \dots, X_N$ are independent random variables with moment generating functions $M_{X_i}(t)$ which converge for $|t| < \delta$, then

$$M_{X_1+\cdots+X_N}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_N}(t).$$

If the random variables all have the same moment generating function $M_X(t)$, then the right hand side becomes $M_X(t)^N$. Unfortunately the moment generating function does not always exist in a neighborhood of the origin (this can be seen by considering the Cauchy distribution); this is rectified by studying the characteristic function, $\mathbb{E}[e^{itX}]$, which is essentially the Fourier transform of the density (that is $\mathbb{E}[e^{-2\pi itX}]$).

(4) **Moment Problem:** When does a sequence of moments uniquely determine a probability density? If our distribution is discrete and takes on only finitely many (for definiteness, say N) values, then only finitely many moments are needed. If the density is continuous, however, infinitely many might not be enough. Consider

$$\begin{aligned} f_1(x) \ &= \ \frac{1}{\sqrt{2\pi x^2}}\, e^{-(\log^2 x)/2} \\ f_2(x) \ &= \ f_1(x)\left[1 + \sin(2\pi \log x)\right]. \end{aligned}$$

These two densities have the same integral moments (their k^{th} moments are $e^{k^2/2}$ for k a non-negative integer); while they also have the same half-integral moments, all other moments differ (thus there is no sequence of moments where they agree which has an accumulation point; see §6). Thus it is possible for two densities to have the same integral moments but differ.

5.4. Approximations and Estimations.

(1) **Cauchy-Schwarz inequality:** For complex-valued functions f and g,

$$\int_0^1 |f(x)g(x)|dx \ \le \ \left(\int_0^1 |f(x)|^2 dx\right)^{\frac{1}{2}} \cdot \left(\int_0^1 |g(x)|^2 dx\right)^{\frac{1}{2}}.$$

One of my favorite applications of this was proving the absolute value of the covariance of X and Y is at most the product of the square-roots of the variances. The key step in the proof was writing the joint density $f_{X,Y}(x,y)$ as $\sqrt{f_{X,Y}(x,y)} \cdot \sqrt{f_{X,Y}(x,y)}$ and putting one factor with $|x - \mu_X|$ and one with $|y - \mu_Y|$. The reason we do this is we cannot directly integrate x^2 or $|x - \mu_X|^2$; we need to hit it with a probability density in order to have a chance of getting a finite value. This explains why we write the density as a product of its square root with its square root; it allows us to use Cauchy-Schwarz.

(2) **Stirling's Formula:** Almost any combinatorial problem involves factorials, either directly or through binomial coefficients. It is essential to be able to estimate $n!$ for large n. Stirling's formula says

$$n! \ = \ n^n e^{-n} \sqrt{2\pi n} \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} + \cdots\right);$$

thus for n large, $n! \approx (n/e)^2 \sqrt{2\pi n}$. There are many ways to prove this, the most common being complex analysis or stationary phase. We can get a ballpark estimate by 'summifying'. We have $n! = \exp(\log n!)$, and

$$\log n! = \sum_{k=1}^n \log k \ \approx \ \int_1^n \log t dt.$$

As the anti-derivative of $\log t$ is $t \log t$, we find $\log n! \approx n \log n - n$, so $n! \approx e^{n \log n - n} = n^n e^{-n}$, which is off by a factor of $\sqrt{2\pi n}$ (while this is a large number, it is small relative to $n^n e^{-e}$. If we wanted, using the integral test and a better job of estimate upper and lower sums (the Euler-Maclaurin formula), we could get a better approximation for $n!$.

(3) **Chebyshev's Theorem:** Chebyshev's theorem (or inequality) is a mixed blessing; it is terrific in the sense that it works for any density that has finite mean and variance; however, in many applications its estimates are far from the truth. The reason is that it works for *all* such densities, and thus cannot exploit any specific properties of the density to get decay. (This is similar to the difference between using Divide and Conquer or Newton's Method to find a zero of a function; Newton's method is magnitudes faster because it assumes more about the function, namely differentiability, and thus it exploits that to get better estimates.) Chebyshev's theorem states

$$\text{Prob}(|X - \mu| \ge k\sigma) \ \le \ \frac{1}{k^2}.$$

Note the event $|X - \mu| \ge k\sigma$ is a very natural event to consider: we are seeing how far X is from its expected value, and measuring this difference in terms of the natural units, the standard deviation. The assumptions for Chebyshev's theorem are a little weaker than those for the Central Limit Theorem, and there are situations where crude bounds suffice (for example, some of the problems we studied in additive number theory).

(4) **The Central Limit Theorem:** The Central Limit Theorem (CLT) states that if $X_1, \dots, X_n$ are independent, identically distributed random variables with mean μ and variance σ^2, then in many instances we have

$$Z_n \ := \ \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \ = \ \frac{\frac{X_1+\cdots+X_n}{n} - \mu}{\sigma/\sqrt{n}}$$

converges to having the standard normal distribution as $n \to \infty$. If the moment generating function exists in a neighborhood containing the origin, that suffices for the CLT to hold (though with additional work we the conclusion holds under weaker assumptions about the X_i's). In practice one often uses the normal approximation once $n \geq 30$. One application is to use the CLT to estimate sums of random variables. Another is for hypothesis testing; there key thresholds are that if Z has the standard normal distribution, the $\text{Prob}(|Z| \leq 1) \approx 68.3\%$, $\text{Prob}(|Z| \leq 1.96) \approx 95.0\%$ and $\text{Prob}(|Z| \leq 2.575) \approx 99.0\%$.

(5) **Taylor Series:** See the section from Calculus I and II. For us, particularly important Taylor series are

$$\begin{aligned}
\log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
\log(1-x) &= -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots\right) \\
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \\
e^{-x} &= 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots = \lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n \\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots .
\end{aligned}$$

5.5. Applications.

(1) **Benford's Law:**
(2) **Additive Number Theory:**
(3) **Economics:**
(4) **Gambling:**
(5) **Sabermetrics:**
(6) **Monte Carlo Integration:**

6. Complex Analysis and Fourier Analysis

(1) **Integral transforms:** If $K(s,t)$ and $g(t)$ are nice functions, we define the integral transform of g with kernel K to be $\int_{-\infty}^{\infty} g(t)K(s,t)dt$. What this does is, given a function as input, generates a new function. Two particularly useful transforms are the Fourier transform ($\widehat{f}(y) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx$) and the Laplace transform ($(\mathcal{L}f)(s) = \int_0^{\infty} f(t)e^{-st}dt$). Depending on the problem, it may be worthwhile to take a transform of both sides, as often

the transformed quantity is easier to analyze. For example, if X and Y are independent random variables with densities f_X and f_Y, then the density of their sum is the convolution

$$f_{X+Y}(t) \ = \ (f_X * f_Y)(t) \ = \ \int_{-\infty}^{\infty} f_X(u) f_Y(t-u) du.$$

As the Fourier transform of a convolution is the pointwise product of the Fourier transforms, we have

$$\widehat{f_{X+Y}}(t) \ = \ \widehat{f_X}(t) \cdot \widehat{f_Y}(t);$$

thus the convolution integral has been replaced with standard multiplication (the integration has not vanished – we must take the Fourier transforms of f_X and f_Y, and then we must take the inverse Fourier transform to recover f_{X+Y}; however, this is still often progress). There are many other nice properties of the Fourier transform. For example, let p be a probability density. Then

$$\widehat{p}(y) \ = \ \int_{-\infty}^{\infty} p(x) e^{-2\pi i x y} dx.$$

Taking the derivative yields

$$\widehat{p}'(y) \ = \ \int_{-\infty}^{\infty} p(x) \cdot (-2\pi i x) e^{-2\pi i x y} dx,$$

and then setting $y = 0$ yields

$$\widehat{p}'(0) \ = \ -2\pi i \int_{-\infty}^{\infty} x p(x) dx \ = \ -2\pi i \mathbb{E}[X].$$

We note two important items: the Fourier transform of $-2\pi i x$ times the function p is the derivative of the Fourier transform of p, and the derivative of the Fourier transform at 0 is a simple multiple of the mean (and a generalization holds for higher moments).

(2) **Complex differentiability:** A function of a complex variable is said to be complex differentiable at z if

$$\lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

exists as $h \neq 0$ tends to 0 along any path. Functions such as the polynomials $\sum_{k=0}^{n} c_k z^k$ are differentiable, while functions such as $\overline{z}^k$ are not (remember $\overline{z} = x - iy$ if $z = x + iy$). If a complex function is differentiable once, it is infinitely differentiable and it equals its Taylor series; this is remarkably different than real analysis (remember the function $f(x) = \exp(-1/x^2)$ for $x \neq 0$ and 0 for $x = 0$; this function is infinitely differentiable, but only equals its Taylor series at $x = 0$ (which is not impressive, as by definition all functions equal their Taylor series at the expansion point!).

(3) **Analytic continuation:** Given a function f defined in some subset of the complex plane, its analytic continuation is a new function which agrees with the old in the original region, but makes sense elsewhere. The standard example is the geometric series formula: $\sum_{n=0}^{\infty} x^n = 1/(1-x)$; the right hand side makes sense for all values of $x \neq 1$, while the left hand side is only defined if $|x| < 1$. This leads to the interpretation that $1 + 2 + 4 + 8 + 16 + \cdots = -1$!

(4) **Accumulation points:** Let f be a complex differentiable function defined on an open set U; assume $f(z_n) = 0$ for some sequence of points $\{z_n\}_{n=1}^{\infty}$ that has an accumulation point in U (i.e., there is some z^* such that a subsequence of the z_n's converge to z^*). Then a beautiful result from complex analysis says that f is identically zero! Again, this is very different than real analysis: the function $f(x) = x^3 \sin(1/x)$ for $x \neq 0$ and 0 for $x = 0$ is zero whenever $x = 1/n\pi$, and is zero at $x = 0$; however, clearly this function is not identically zero even near the origin (just consider $x = 2/n\pi$ for n odd). In probability, this result is used to study the moment problem, namely, how many moments are needed to uniquely determine a probability density.

(5) **Poisson summation:** for nice functions, $\sum f(n) = \sum \widehat{f}(n)$. Often this allows us to replace a long sum of slowly decaying terms with a short sum of rapidly decaying terms. We used this in obtaining very good estimates on the probability of being far from the mean for normal random variables.

7. General Techniques

(1) **Reduction to integration:** Many problems can be reduced to the determination of an integral, such as the equidistribution of $n^k\alpha \bmod 1$ (for $n \in \{1, \dots, N\}$ and $\alpha \notin \mathbb{Q}$). That's the good news. The bad news is that we must evaluate these integrals as a function of some parameter, say N, and typically it is *not* possible to write down a closed form answer. Thus we need to develop techniques to approximate these integrals and determine some good control of the N dependence.

(2) **Being algebraically lazy:** Another common theme is that we try to do as little work as possible to get as good of an estimate as needed. For example, we computed the moment generating function of the standard normal by completing the square, and found $M_X(t) = \mathbb{E}[e^{tX}] = e^{t^2/2}$. Later we needed to Fourier transform of the standard normal; while we could attack the integral which arises, it is far easier to note the Fourier transform at y is the same as the moment generating function at $-2\pi i y$. While we need to use some results from complex analysis to justify this argument, we now get the Fourier transform.

(3) **Problem formulation and blinders:** We've also seen on a few problems how the way the problem is formulated can influence how one attempts to solve it. For example, recall the function $x^3 \sin(1/x)$. The oscillation is bounded by two cubics; however, if we just look at the part above the x-axis, the plot looks like a parabola. It is thus a good idea, if you're stuck, to try and think of alternative ways of looking at a problem.

(4) **Choosing approaches.** Certain functions become natural choices in studying certain problems. For example, for $n^k\alpha \bmod 1$ we use the exponential function. The reason this is so useful is that $\exp(2\pi i n^k \alpha) = \exp(2\pi i (n^k\alpha \bmod 1))$. Thus we may drop the difficult modulo 1 condition and sum more easily. Depending on the problem, different functions and expansions will be more useful than others. The ease at which the exponential function handles the modulo 1 condition suggests the usefulness of applying Fourier analysis.

(5) **Adding zero / multiplying by one:** This is perhaps **the** most important technique to learn, though it is one of the hardest to master. The difficult part of these methods is figuring out how to 'do nothing' in an intelligent way. The first example you might remember is proving the product rule from calculus. Let $A(x) = f(x)g(x)$. Then

$$\begin{aligned} A'(x) &= \lim_{h\to 0} \frac{A(x+h) - A(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)g(x+h)\textbf{-f(x)g(x+h) + f(x)g(x+h)} - f(x)g(x)}{h} \\ &= \lim_{h\to 0} \left[\frac{f(x+h)g(x+h) - f(x)g(x+h)}{h} + \frac{f(x)g(x+h) - f(x)g(x)}{h}\right] \\ &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} g(x+h) + \lim_{h\to 0} f(x) \frac{g(x+h) - g(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \lim_{h\to 0} g(x+h) + f(x) \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} \\ &= f'(x)g(x) + f(x)g'(x). \end{aligned}$$

My favorite example was probably in proving the multinomial distribution is a density.

(6) **Summifying or summification:** We frequently replace $\prod a_n$ with $\exp\left(\log \prod a_n\right)$, as this converts the product to a sum, and we have a much better understanding of sums. Probably the most important use was in proving the Central Limit Theorem, where we replaced studying $\prod_i M_{X_i}(t)$ with studying $\sum_i \log M_{X_i}(t)$. We also used it to obtain an approximation for Stirling's formula, replacing $n!$ with $\sum_{\ell \le n} \log \ell$ (which we evaluated by using the integral test).

(7) **Removing conditions:** Whenever you have a theorem, you should always explore what happens if you remove a condition. Frequently (though not always) the claim no longer holds; sometimes the claim is still true but the proof is harder. Rarely, but it can happen, removing a condition causes you to look at a problem in a new light, and find a simpler proof.

(8) **Efficient algebra:** It is frequently worthwhile to think about whether or not we can approach a tedious algebra problem another way. Some examples from previous courses: to compute A^n for n large, diagonalize A if possible, say $A = S\Lambda S^{-1}$ with Λ the diagonal matrix of eigenvalues. Then $A^n = S\Lambda^n S^{-1}$, and Λ^n is readily computed. Another example is telescoping series, $(a_1 - a_0) + (a_2 - a_1) + \cdots + (a_n - a_{n-1}) = a_n - a_0$; this is a key ingredient in many proofs of the Fundamental Theorem of Calculus. Frequently in probability we combine these approaches with recognizing and exploiting an identity; for example, if we had to evaluate $\binom{n}{2}2^2 + \binom{n}{3}2^3 + \cdots + \binom{n}{n}2^n$, we might notice that this is almost the binomial expansion of $(1+2)^n$; it would be, but we're missing the first two terms. The

solution is to add zero by adding and subtracting those terms, which gives

$$\begin{aligned}\binom{n}{2}2^2 + \binom{n}{3}2^3 + \cdots + \binom{n}{n}2^n &= \sum_{k=0}^{n}\binom{n}{k}1^{n-k}2^k\Bigg) - \Bigg(\binom{n}{0} + \binom{n}{1}2\Bigg) \\ &= (1+2)^n - (n + n(n-1)) \;=\; 3^n - n^2;\end{aligned}$$

note we included the factor 1^{n-k} to make this match the standard binomial theorem expansion.

(9) **Illuminating algebra:** It is very easy to obtain complicated expressions involving the parameters of interest; while the answer is correct, the final product is not illuminating. It is worthwhile to see if the answer can be simplified. For example, consider the sabermetrics (baseball math) problem where we had Team X scores runs from a geometric distribution with parameter p (in this case, $\text{Prob}(X = m) = (1-p)p^m$ for $m \in \{0, 1, 2, \dots\}$ and allows runs to Team Y with a geometric distribution with parameter q; we assume the two random variables are independent. The mean number of runs Team X scores is denoted RS, and equals $\text{RS} = \frac{p}{1-p}$ which implies $p = \frac{\text{RS}}{\text{RS}+1}$; we let RA denote the runs allowed, and $\text{RA} = \frac{q}{1-q}$ which implies $q = \frac{\text{RA}}{\text{AS}+1}$. After some algebra we found the probability Team X wins is

$$\frac{p(1-q)}{p(1-q) + q(1-p)}.$$

No one, however, things in terms of the decay probability from scoring m to scoring $m+1$ runs; we want a formula in terms of runs scored RS and runs allowed RA. Substituting for p and q yields

$$\frac{\left(1 - \frac{\text{RA}}{1+\text{RA}}\right)\text{RS}}{(1+\text{RS})\left(\frac{\left(1-\frac{\text{RA}}{1+\text{RA}}\right)\text{RS}}{1+\text{RS}} + \frac{\text{RA}\left(1-\frac{\text{RS}}{1+\text{RS}}\right)}{1+\text{RA}}\right)},$$

a most unilluminating formula! With some work, we can simplify this to the nice answer we'll describe below; however, what is important about this problem (for us – major league baseball would beg to differ!) is not the result, but how to reach it efficiently. We know that $\frac{p}{1-p}$ is a nice expression, namely RS, and similarly for $\frac{q}{1-q}$. Thus we should take our expression and multiply by 1 in the form $(1/(1-p)(1-q)) \Big/ (1/(1-p)(1-q))$. Doing so yields

$$\frac{p(1-q)}{p(1-q) + q(1-p)} \cdot \frac{\frac{1}{(1-p)(1-q)}}{\frac{1}{(1-p)(1-q)}} \;=\; \frac{\frac{p}{1-p}}{\frac{p}{1-p} + \frac{q}{1-q}} \;=\; \frac{\text{RS}}{\text{RS} + \text{RA}}.$$

Note we obtain a very nice formula very quickly.

(10) **Numerical exploration:** When given a problem, one can frequently build intuition by running numerical experiments. For example, one of our problems concerned a person who made 40% of all their shots. We wanted to know the probability that the number of shots required to make 341 baskets was within 35 of the mean number of shots required. We came up with an answer by seeing that this was equivalent to the sum of 341 independent

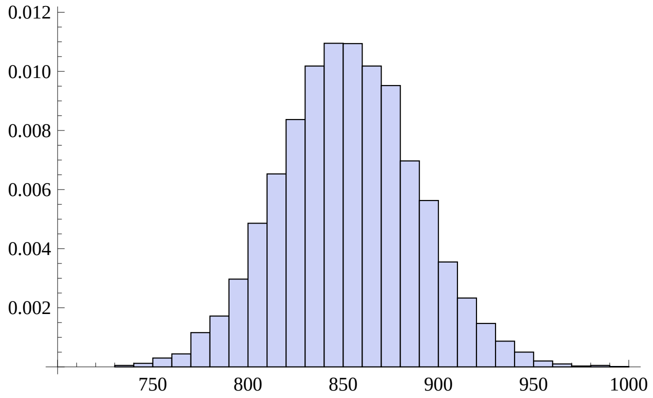

FIGURE 1. Histogram plot of number of shots to make 341 baskets given a 40% chance of making a shot. The data was obtained by playing the game 10,000 times and recording how long it took. The sample mean is 852.058 (which is quite close to the predicted 852.5), the sample standard deviation is 35.8092 (quite close to the predicted 35.7596), and 67.4% of the time the number of shots was within 35 of 852.5 (quite close to our prediction).

geometric random variables with parameter $p = .4$, and thus the Central Limit Theorem is applicable to estimate the probability.

To test our predictions, consider the person shooting until they get 341 baskets a staggering 10,000 times (see Figure 1). Note the numerical data is quite close to theory. If you can program in some environment, you can quickly gather numerical data to help elucidate the answer. The Mathematica code for this problem is:

```
tester[num_]:=Module[{},
count = {};
prob = 0;
mean = 852.5;
For[n = 1, n ≤ num, n++,
{
numfound = 0;
counter = 0;
While[numfound < 341,
{
counter = counter + 1;
If[Random[] ≤ .4, numfound = numfound + 1];
}];
count = AppendTo[count, counter];
If[Abs[counter − mean] ≤ 35, prob = prob + 1];
}];
Print[Histogram[count, {750, 950, 10}, Probability]];
Print[prob100.0/num];
];
```

Of course, sometimes we are fortunate enough that, instead of settling for numerical answers, programs like Mathematica can find the exact answer. For example, consider the following difference equation, which arises in a problem related to a random walk with boundaries:

$$T_{i+1} \ = \ \frac{1}{p}T_i - \frac{1-p}{p}T_{i-1} - \frac{1}{p}.$$

Typing

Simplify[RSolve[{T[i] == p (T[i + 1] + 1) + (1 - p) (T[i - 1] + 1), T[0] == 0, T[M] == 0}, T[i], i]]

into Mathematica yields

$$T_i \ = \ \frac{i + M\left(\left(\frac{1-p}{p}\right)^i - 1\right) - i\left(\frac{1-p}{p}\right)^M}{\left(\left(\frac{1-p}{p}\right)^M - 1\right)(2p-1)}.$$

(11) **Test functions:** You should always consider testing the limits of a theorem, conjecture or intuition. Does it hold for the standard normal? For the Cauchy? How important is the finiteness of moments? Usually a result is false if you remove a condition; however, when you are trying to figure out what the conditions should be in a theorem, you're in a different mindset. In this case, it is worthwhile to play with various functions and see what happens.

(12) **Check for reasonableness:** Whenever we have a formula, it is a very good idea to check special cases to see if it is reasonable. For example, consider the sabermetrics formula from the previous point: if a team scores on average RS runs per game and allows on average RA per game (with RS and RA independent geometric random variables with respective means RS and RA), then its probability of winning is $\mathrm{RS}/(\mathrm{RS}+\mathrm{RA})$. Is this formula reasonable? There are many checks we can do. The first is that we always get a number between 0 and 1 (which is a must for a probability!). Further, if RS is zero or if RA tends to infinity than we have no chance of winning, exactly as we would expect. If we score on average more runs than we allow, our winning percentage is greater than 50%, while if we score and allow the same number on average than the winning percentage is 50%, again quite reasonable.

For another example, imagine we flip a fair coin with probability p of heads and $1-p$ of tails n times, and we ask how many runs (alterations between heads and tails) there are; for example, if the outcome were HHTTHTHTTTTTHTHHHH then there were 18 tosses, 9 heads and 9 tails and 9 runs, the shortest being a run of length 1 and the longest being a run of length 5. The expected number of runs is $1 + (n-1)2p(1-p)$. Is this formula reasonable? Note that if $p = 0$ or $p = 1$ then because of the factor $p(1-p)$ the expected number of runs is 1; we should be shocked if this is not the case, as if the coin always lands on heads, how could there ever be an alteration? A little calculus shows that the maximum expected value is when $p = 1/2$, which also seems reasonable. Finally, in the special case $p = 1/2$ the expected number is essentially $n/2$; there are n tosses and each toss has a 50% chance of being different than the previous (and thus starting a run), so again our answer makes sense.

(13) **Check all conditions:** Whenever you want to use a theorem, make sure all the conditions are satisfied. For example, if you are summing the geometric series $1 + x + x^2 + x^3 + \cdots$ then you better have $|x| < 1$. If you are asked whether or not something is a probability distribution, it must satisfy both requirements (non-negative and sums to 1; it is not enough to just sum to one). If you want something to be a group, it must satisfy all four properties (closure, identity, associativity, inverse). Frequently some but not all of the conditions are met.

E-mail address: Steven.J.Miller@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267