Deviation from the Mean

1 What the Mean Means

We have focused on the expectation of a random variable because it indicates the "average value" the random variable will take. But what precisely does this mean?

We know a random variable may never actually equal its expectation. We also know, for example, that if we flip a fair coin 100 times, the chance that we actually flip *exactly* 50 heads is only about 8%. In fact, it gets less and less likely as we continue flipping that the number of heads will exactly equal the expected number, *e.g.*, the chance of exactly 500 heads in 1000 flips is less than 3%, in 1,000,000 flips less than 0.1%, ...

But what is true is that the fraction of heads flipped is likely to be *close* to half of the flips, and the more flips, the closer the fraction is likely to be to 1/2. For example, the chance that the fraction of heads is within 5% of 1/2 is

- more than 24% in 10 flips,
- more than 38% in 100 flips,
- more than 56% in 200 flips, and
- more than 89% in 1000 flips.

These numbers illustrate the single most important phenomenon of probability: the average value from repeated experiments is likely to be close to the expected value of one experiment. And it gets more likely to be closer as the number of experiments increases. This result was first formulated and proved by Jacob D. Bernoulli in his book *Ars Conjectandi* (The Art of Guessing) published posthumously in 1713. In his Introduction, Bernoulli comments that[1]

> even the stupidest man—by some instinct of nature *per se* and by no previous instruction (this is truly amazing)—knows for sure that the more observations . . . that are taken, the less the danger will be of straying from the mark.

But he goes on to argue that this instinct should not be taken for granted:

[1] These quotes are taken from Grinstead & Snell, *Introduction to Probability*, American Mathematical Society, p. 310.

> Something further must be contemplated here which perhaps no one has thought about until now. It certainly remains to be inquired whether after the number of observations has been increased, the probability ... of obtaining the true ratio ... finally exceeds any given degree of certainty; or whether the problem has, so to speak, its own asymptote—that is, whether some degree of certainty is given which one can never exceed.

Here's how to give a technical formulation of the question Bernoulli wants us to contemplate. Repeatedly performing some random experiment corresponds to defining n random variables equal to the results of n trials of the experiment. That is, we let $G_1, \ldots, G_n$ be independent random variables with the same distribution and the same expectation, μ. Now let A_n be the average of the results, that is,

$$A_n ::= \frac{\sum_{i=1}^{n} G_i}{n}.$$

How sure we can be that the average value, A_n, will be close to μ? By letting n grow large enough, can we be as certain as we want that the average will be close, or is there is some irreducible degree of uncertainty that remains no matter how many trials we perform? More precisely, given any positive tolerance, ϵ, how sure can we be that the average, A_n, will be within the tolerance of μ as n grows? In other words, we are asking about the limit

$$\lim_{n\to\infty} \Pr\{|A_n - \mu| < \epsilon\}.$$

Bernuolli asks if we can be sure this limit approaches certainty, that is, equals one, or whether it approaches some number slightly less than one that cannot be increased to one no matter how many times we repeat the experiment. His answer is that the limit is indeed one. This result is now known as the Weak Law of Large Numbers. Bernoulli says of it:

> Therefore, this is the problem which I now set forth and make known after I have pondered over it for twenty years. Both its novelty and its very great usefulness, coupled with its just as great difficulty, can exceed in weight and value all the remaining chapters of this thesis.

With the benefit of three centuries of mathematical development since Bernoulli, it will be a lot easier for us to resolve Bernoulli's questions than it originally was for him.

2 The Weak Law of Large Numbers

The Weak Law of Large Numbers crystallizes, and confirms, the intuition of Bernoulli's "stupidest man" that the average of a large number of independent trials is more and more likely to be within a smaller and smaller tolerance around the expectation as the number of trials grows.

Theorem 2.1. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n, \ldots$ be independent variables with the same distribution and the same expectation, μ. For any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{\sum_{i=1}^{n} G_i}{n} - \mu\right| \leq \epsilon\right\} = 1.$$

This Law gives a high-level description of a fundamental probabilistic phenomenon, but as it stands it does not give enough information to be of practical use. The main problem is that it does not say anything about the *rate* at which the limit is approached. That is, how big must n be to be within a given tolerance of the expected value with a specific desired probability? This information is essential in applications. For example:

- Suppose we want to estimate the number of voters who are registered Republicans. Exactly *how many* randomly selected voters should we poll in order to be sure that 99% of the time, the average number of Republicans in our poll is within 1/2% of the actual percentage in the whole country?
- Suppose we want to estimate the number of fish in a lake. Our procedure will be to catch, tag and release 500 fish caught in randomly selected locations in the lake at random times of day. Then we wait a few days, and catch another 100 random fish. Suppose we discover that 10 of the 100 were previously tagged. Assuming that our 500 tagged fish represent the same proportion of the whole fish population as the ones in our sample of 100, we would estimate that the total fish population was 5000. But how confident can we be of this? Specifically, *how confident* should we be that our estimate of 5000 is within 20% of the actual fish population?
- Suppose we want to estimate the average size of fish in the lake by taking the average of the sizes of the 500 in our initial catch. How confident can we be that this average is within 2% of the average size of all the fish in the lake?

In these Notes we will develop three basic results about this topic of *deviation from the mean*. The first result is Markov's Theorem, which gives a simple but coarse upper bound on the probability that the value of the random variable is more than a certain multiple of its mean. Markov's result holds if we know nothing more than the value of the mean of a random variable. As such, it is very general, but also is much weaker than results which take more information about the random variable into account.

In many situations, we not only know the mean, but also another numerical quantity called the *variance* of the random variable. Our second basic result is Chebyshev's Theorem, which combines Markov's Theorem and information about the variance to give more refined bounds.

The third basic result we call the Pairwise Independent Sampling Theorem. It provides the additional information about rate of convergence we need to calculate numerical answers to questions such as those above. The Sampling Theorem follows from Chebyshev's Theorem and properties of the variance of a sum of independent variables.

Finally, the Weak Law of Large Numbers will be an easy corollary of the Pairwise Independent Sampling Theorem.

2.1 Markov's Theorem

We want to consider the problem of bounding the probability that the value of a random variable is far away from the mean. Our first theorem, Markov's theorem, gives a very rough estimate, based only on the value of the mean.

The idea behind Markov's Theorem can be explained with a simple example of I.Q. measurment. I.Q. was devised so that the average I.Q. measurement would be 100. Now from this fact alone we

can conclude that at most $1/2$ the population can have an I.Q. of 200 or more, because if more than half had an I.Q. of 200, then the average would have to be more than $(1/2)200 = 100$, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an I.Q. of 200 or more is at most $1/2$. Of course this is not a very strong conclusion; in fact no I.Q. of over 200 has ever been recorded. But by the same logic, we can also conclude that at most $2/3$ of the population can have an I.Q. of 150 or more. I.Q.'s of over 150 have certainly been recorded, though again, a much smaller fraction of the population actually has an I.Q. that high.

But although these conclusions about I.Q. are weak, they are actually the *strongest possible* general conclusions that can be reached about a random variable using *only* the fact that its mean is 100. For example, if we choose a random variable equal to 200 with probability $1/2$, and 0 with probability $1/2$, then its mean is 100, and the probability of a value of 200 or more is really $1/2$. So we can't hope to get a upper better bound on the probability of 200 than $1/2$.

Theorem 2.2 (Markov's Theorem). *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr\{R \geq x\} \leq \frac{\mathrm{E}[R]}{x}.$$

Proof. We will show that $\mathrm{E}[R] \geq x \Pr\{R \geq x\}$. Dividing both sides by x gives the desired result.

So let I_x be the indicator variable for the event $[R \geq x]$, and consider the random variable xI_x. Note that $R \geq xI_x$, because if $R(w) \geq x$ then $xI_x(w) = x \cdot 1 = x$, and if $R(w) < x$ then $xI_x(w) = x \cdot 0 = 0$. Therefore,

$$\begin{aligned} \mathrm{E}[R] &\geq \mathrm{E}[xI_x] && (R \geq xI_x) \\ &= x\,\mathrm{E}[I_x] && \text{(linearity of expectation)} \\ &= x \Pr\{R \geq x\}. && (\mathrm{E}[I_x] = \Pr\{I_x = 1\}) \end{aligned}$$

□

Markov's Theorem is often expressed in an alternative form, stated below as a corollary.

Corollary 2.3. *If R is a nonnegative random variable, then for all $c > 0$*

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{1}{c}.$$

Proof. In Markov's Theorem, set $x = c \cdot \mathrm{E}[R]$. This gives:

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{\mathrm{E}[R]}{c \cdot \mathrm{E}[R]} = \frac{1}{c}.$$

□

2.1.1 Examples of Markov's Theorem

Suppose that N men go to a dinner party and check their hats. At the end of the night, each man is given his own hat back with probability $1/N$. What is the probability that x or more men get the right hat?

We can compute an upper bound with Markov's Theorem. Let the random variable, R, be the number of men that get the right hat. In previous notes, we used linearity of expectation to show that $\mathrm{E}[R] = 1$. By Markov's Theorem, the probability that x or more men get the right hat is:

$$\Pr\{R \geq x\} \leq \frac{\mathrm{E}[R]}{x} = \frac{1}{x}.$$

For example, there is no better than a 20% chance that 5 men get the right hat, regardless of the number of people at the dinner party.

The Chinese Appetizer problem is very similar. In this case, N people are eating Chinese appetizers arranged on a circular, rotating tray. Someone then spins the tray so that each person receives a random appetizer. What is the probability that everyone gets the same appetizer as before?

There are N equally likely orientations for the tray after it stops spinning. Everyone gets the right appetizer in just one of these N orientations. Therefore, the correct answer is $1/N$.

But what probability do we get from Markov's Theorem? Let the random variable, R, be the number of people that get the right appetizer. We showed in previous notes that $\mathrm{E}[R] = 1$. Applying Markov's Theorem, we find:

$$\Pr\{R \geq N\} \leq \frac{\mathrm{E}[R]}{N} = \frac{1}{N}.$$

In this case, Markov's Theorem is tight!

On the other hand, Markov's Theorem gives the same $1/N$ bound for the probability everyone gets their hat in the hat check problem. But in reality, the probability of this event is $1/N!$. So Markov's Theorem in this case gives probability bounds that are way off.

2.1.2 Why *R* Must be Nonnegative

The proof of Markov's Theorem requires that the random variable, R, be nonnegative. The following example shows that the theorem is false if this restriction is removed. Let R be -10 with probability $1/2$ and 10 with probability $1/2$. Then we have:

$$\mathrm{E}[R] = -10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 0$$

Suppose that we now tried to compute $\Pr\{R \geq 5\}$ using Markov's Theorem:

$$\Pr\{R \geq 5\} \leq \frac{\mathrm{E}[R]}{5} = \frac{0}{5} = 0.$$

This is the wrong answer! Obviously, R is at least 5 with probability $1/2$. Remember that Markov's Theorem applies only to nonnegative random variables!

On the other hand, we can still apply Markov's Theorem to bound the probability that an arbitrary variable like R is 5 more. Namely, given any random variable, R with expectation 0 and values ≥ -10, we can conclude that $\Pr\{R \geq 5\} \leq 2/3$.

Proof. Let $T ::= R + 10$. Now T is a nonnegative random variable with expectation $\mathrm{E}[R + 10] = \mathrm{E}[R] + 10 = 10$, so Markov's Theorem applies and tells us that $\Pr\{T \geq 15\} \leq 10/15 = 2/3$. But $T \geq 15$ iff $R \geq 5$, so $\Pr\{R \geq 5\} \leq 2/3$, as claimed. □

2.1.3 Deviation Below the Mean

Markov's Theorem says that a random variable is unlikely to greatly exceed the mean. Correspondingly, there is a theorem that says a random variable is unlikely to be much smaller than its mean.

Theorem 2.4. *Let L be a real number and let R be a random variable such that $R \leq L$. For all $x < L$, we have:*

$$\Pr\{R \leq x\} \leq \frac{L - \mathrm{E}[R]}{L - x}.$$

Proof. The event that $R \leq x$ is the same as the event that $L - R \geq L - x$. Therefore:

$$\begin{aligned} \Pr\{R \leq x\} &= \Pr\{L - R \geq L - x\} \\ &\leq \frac{\mathrm{E}[L - R]}{L - x}. \qquad \text{(by Markov' Theorem)} \qquad (1) \end{aligned}$$

Applying Markov's Theorem in line (1) is permissible since $L - R$ is a nonnegative random variable and $L - x > 0$. □

For example, suppose that the class average on the 6.042 midterm was 75/100. What fraction of the class scored below 50?

There is not enough information here to answer the question exactly, but Theorem 2.4 gives an upper bound. Let R be the score of a random student. Since 100 is the highest possible score, we can set $L = 100$ to meet the condition in the theorem that $R \leq L$. Applying Theorem 2.4, we find:

$$\Pr\{R \leq 50\} \leq \frac{100 - 75}{100 - 50} = \frac{1}{2}.$$

That is, at most half of the class scored 50 or worse. This makes sense; if more than half of the class scored 50 or worse, then the class average could not be 75, even if everyone else scored 100. As with Markov's Theorem, Theorem 2.4 often gives weak results. In fact, based on the data given, the entire class could have scored above 50.

2.1.4 Using Markov To Analyze Non-Random Events [Optional]

[Optional]

In the previous examples, we used a theorem about a random variable to conclude facts about non-random data. For example, we concluded that if the average score on a test is 75, then at most $1/2$ the class scored 50 or worse. There is no randomness in this problem, so how can we apply Theorem 2.4 to reach this conclusion?

The explanation is not difficult. For any set of scores $S = \{s_1, s_2, \ldots, s_n\}$, we introduce a random variable, R, such that

$$\Pr\{R = s_i\} = \frac{(\text{\# of students with score } s_i)}{n}$$

We then use Theorem 2.4 to conclude that $\Pr\{R \leq 50\} \leq 1/2$. To see why this means (with certainty) that at most $1/2$ of the students scored 50 or less, we observe that

$$\begin{aligned} \Pr\{R \leq 50\} &= \sum_{s_i \leq 50} \Pr\{R = s_i\} \\ &= \sum_{s_i \leq 50} \frac{(\text{\# of students with score } s_i)}{n} \\ &= \frac{1}{n}(\text{\# of students with score 50 or less}). \end{aligned}$$

So, if $\Pr\{R \leq 50\} \leq 1/2$, then the number of students with score 50 or less is at most $n/2$.

3 Chebyshev's Theorem

We have versions of Markov's Theorem for deviations above and below the mean, but often we want bounds that apply in both directions, that is, bounds on the probability that $|R - \mathrm{E}[R]|$ is large.

It is a bit messy to use Markov's inequality directly to bound the probabilty that $|R - \mathrm{E}[R]| \geq x$, since we then would have to compute $\mathrm{E}[|R - \mathrm{E}[R]|]$. However, since $|R|$ and hence $|R|^k$ are nonnegative variables for any R, Markov's inequality also applies to the event $[|R|^k \geq x^k]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:

Corollary 3.1. *For any random variable R, any positive integer k, and any $x > 0$,*

$$\Pr\{|R| \geq x\} \leq \frac{\mathrm{E}\left[|R|^k\right]}{x^k}.$$

The special case of this corollary when $k = 2$ can be applied to bound the random variable, $R - \mathrm{E}[R]$, that measures R's deviation from its mean. Namely

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} = \Pr\left\{(R - \mathrm{E}[R])^2 \geq x^2\right\} \leq \frac{\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]}{x^2}, \tag{2}$$

where the inequality (2) follows from Corollary 3.1 applied to the random variable, $R - \mathrm{E}[R]$. So we can bound the probability that the random variable R deviates from its mean by more than x by an expression decreasing as $1/x^2$ multiplied by the constant $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$. This constant is called the *variance of R*.

Definition 3.2. The *variance*, $\mathrm{Var}[R]$, of a random variable, R, is:

$$\mathrm{Var}[R] ::= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right].$$

So we can restate (2) as

Theorem 3.3 (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}.$$

The expression $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \mathrm{E}[R]$, is precisely the deviation of R from the mean. Squaring this, we obtain, $(R - \mathrm{E}[R])^2$. This is a random variable that is near 0 when R is close to the mean and is a large positive number when R deviates far above or below the mean. The variance is just the average of this random variable, $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$. Therefore, intuitively, if R is always close to the mean, then the variance will be small. If R is often far from the mean, then the variance will be large. For this reason, variance is useful in studying the probability that a random variable deviates far from the mean.

3.1 Example: Gambling Games

The relevance of variance is apparent when we compare the following two gambling games.

Game A: We win \$2 with probability $2/3$ and lose \$1 with probability $1/3$.

Game B: We win \$1002 with probability $2/3$ and lose \$2001 with probability $1/3$.

Which game is better financially? We have the same probability, 2/3, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example, A is 2 with probability 2/3 and -1 with probability $1/3$. We can compute the expected payoff for each game as follows:

$$\begin{aligned} \mathrm{E}[A] &= 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1, \\ \mathrm{E}[B] &= 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1. \end{aligned}$$

The expected payoff is the same for both games, but they are obviously very different! This difference is hidden by expected value, but captured by variance. We can compute the $\mathrm{Var}[A]$ by working "from the inside out" as follows:

$$\begin{aligned} A - \mathrm{E}[A] &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases} \\ (A - \mathrm{E}[A])^2 &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases} \\ \mathrm{E}\left[(A - \mathrm{E}[A])^2\right] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\ \mathrm{Var}[A] &= 2. \end{aligned}$$

Similarly, we have for $\mathrm{Var}[B]$:

$$\begin{aligned} B - \mathrm{E}[B] &= \begin{cases} 1001 & \text{with probability } \frac{2}{3} \\ -2002 & \text{with probability } \frac{1}{3} \end{cases} \\ (B - \mathrm{E}[B])^2 &= \begin{cases} 1,002,001 & \text{with probability } \frac{2}{3} \\ 4,008,004 & \text{with probability } \frac{1}{3} \end{cases} \\ \mathrm{E}\left[(B - \mathrm{E}[B])^2\right] &= 1,002,001 \cdot \frac{2}{3} + 4,008,004 \cdot \frac{1}{3} \\ \mathrm{Var}[B] &= 2,004,002. \end{aligned}$$

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff in Game A is usually close to the expected value of \$1, but the payoff in Game B can deviate very far from this expected value.

High variance is often associated with high risk. For example, in ten rounds of Game A, we expect to make \$10, but could conceivably lose \$10 instead. On the other hand, in ten rounds of game B, we also expect to make \$10, but could actually lose more than \$20,000!

4 Properties of Variance

4.1 Why Variance?

The definition of variance of R as $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ may seem rather arbitrary. The variance is the average *of the square* of the deviation from the mean. For this reason, variance is sometimes called the "mean squared deviation." But why bother squaring? Why not simply compute the average deviation from the mean? That is, why not define variance to be $\mathrm{E}[R - \mathrm{E}[R]]$?

The problem with this definition is that the positive and negative deviations from the mean exactly cancel. By linearity of expectation, we have:

$$\mathrm{E}[R - \mathrm{E}[R]] = \mathrm{E}[R] - \mathrm{E}[\mathrm{E}[R]].$$

Since $\mathrm{E}[R]$ is a constant, its expected value is itself. Therefore

$$\mathrm{E}[R - \mathrm{E}[R]] = \mathrm{E}[R] - \mathrm{E}[R] = 0.$$

By this definition, every random variable has zero variance. That is not useful! Because of the square in the conventional definition, both positive and negative deviations from the mean increase the variance; positive and negative deviations do not cancel.

Of course, we could also prevent positive and negative deviations from cancelling by taking an absolute value. That is, we could define variance to be $\mathrm{E}[|R - \mathrm{E}[R]|]$. There is no great reason not to use this definition. However, the conventional version of variance has some pleasant mathematical properties that the absolute value variant does not. For example, for independent random variables, the variance of a sum is the sum of the variances; that is, $\mathrm{Var}[R_1 + R_2] = \mathrm{Var}[R_1] + \mathrm{Var}[R_2]$. We will prove this fact below.

4.2 Standard Deviation

Due to squaring, the variance of a random variable may be very far from a typical deviation from the mean. For example, in Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a whopping 2,004,002. From a dimensional analysis viewpoint, the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using standard deviation instead of variance.

Definition 4.1. The *standard deviation* of a random variable R is denoted σ_R and defined as the square root of the variance:

$$\sigma_R ::= \sqrt{\mathrm{Var}[R]} = \sqrt{\mathrm{E}[(R - \mathrm{E}[R])^2]}.$$

So the standard deviation is the square root of the mean of the square of the deviation, or the "root mean square" for short. It has the same units—dollars in our example—as the original random variable and as the mean. Intuitively, it measures the "expected (average) deviation from the mean," since we can think of the square root on the outside as cancelling the square on the inside.

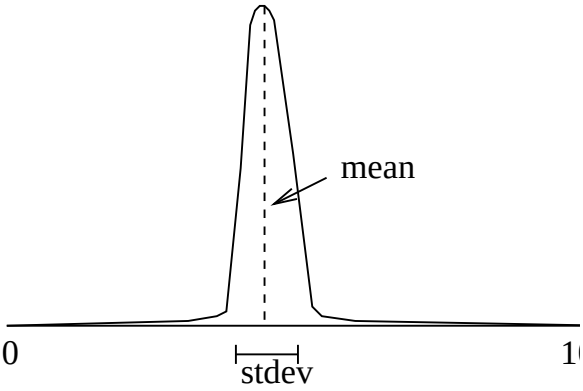

Figure 1: The standard deviation of a distribution says how wide the "main" part of it is.

Example 4.2. The standard deviation of the payoff in Game B is:

$$\sigma_B = \sqrt{\text{Var}\left[B\right]} = \sqrt{2,004,002} \approx 1416.$$

The random variable B actually deviates from the mean by either positive 1001 or negative 2002; therefore, the standard deviation of 1416 describes the situations reasonably well.

As can be seen in Figure 1, the standard deviation measures the "width" of the main part of the distribution graph.

4.3 An Alternative Definition of Variance

There is an equivalent way to define the variance of a random variable that is less intuitive, but is often easier to use in calculations and proofs:

Theorem 4.3.

$$\text{Var}\left[R\right] = \text{E}\left[R^2\right] - \text{E}^2\left[R\right],$$

for any random variable, R.

Here we use the notation $\text{E}^2\left[R\right]$ as shorthand for $(\text{E}\left[R\right])^2$.

Remember that $\text{E}\left[R^2\right]$ is generally not equal to $\text{E}^2\left[R\right]$. We know the expected value of a product is the product of the expected values for independent variables, but not in general. And R is not independent of itself unless it is constant.

Proof. Let $\mu = \text{E}\left[R\right]$. Then

$$\begin{aligned}
\text{Var}\left[R\right] &= \text{E}\left[(R - \text{E}\left[R\right])^2\right] && \text{(Def. 3.2 of variance)}\\
&= \text{E}\left[(R-\mu)^2\right] && \text{(Def. of } \mu\text{)}\\
&= \text{E}\left[R^2 - 2\mu R + \mu^2\right]\\
&= \text{E}\left[R^2\right] - 2\mu\,\text{E}\left[R\right] + \mu^2 && \text{(linearity of expectation)}\\
&= \text{E}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(definition of } \mu\text{)}\\
&= \text{E}\left[R^2\right] - \mu^2\\
&= \text{E}\left[R^2\right] - \text{E}^2\left[R\right]. && \text{(definition of } \mu\text{)}
\end{aligned}$$

□

[Optional]

Theorem 4.3 gives a convenient way to compute the variance of a random variable: find the expected value of the square and subtract the square of the expected value. For example, we can compute the variance of the outcome of a fair die as follows:

$$\mathrm{E}\left[R^2\right] = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6},$$
$$\mathrm{E}^2[R] = \left(3\frac{1}{2}\right)^2 = \frac{49}{4},$$
$$\mathrm{Var}[R] = \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R] = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

This result is particularly useful when we want to estimate the variance of a random variable from a sequence $x_1, x_2, \ldots, x_n$, of sample values of the variable.

Definition 4.4. For any sequence of real numbers $x_1, x_2, \ldots, x_n$, define the *sample mean*, μ_n, and the *sample variance*, v_n, of the sequence to be:

$$\mu_n ::= \frac{\sum_{i=1}^n x_i}{n},$$
$$v_n ::= \frac{\sum_{i=1}^n (x_i - \mu_n)^2}{n}.$$

Notice that if we define a random variable, R, which is equally likely to take each of the values in the sequence, that is $\Pr\{R = x_i\} = 1/n$ for $i = 1, \ldots, n$, then $\mu_n = \mathrm{E}[R]$ and $v_n = \mathrm{Var}[R]$. So Theorem 4.3 applies to R and lets us conclude that

$$v_n = \frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2. \tag{3}$$

This leads to a simple procedure for computing the sample mean and variance while reading the sequence $x_1, \ldots, x_n$ from left to right. Namely, maintain a sum of all numbers seen and also maintain a sum of the squares of all numbers seen. That is, we store two values, starting with the values x_1 and x_1^2. Then, as we get to the next number, x_i, we add it to the first sum and add its square, x_i^2, to the second sum. After a single pass through the sequence $x_1, \ldots, x_n$, we wind up with the values of the two sums $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$. Then we just plug these two values into (3) to find the sample variance.

4.3.1 Expectation Squared [Optional]

[Optional]

The alternate definition of variance given in Theorem 4.3 has a cute implication:

Corollary 4.5. *If R is a random variable, then* $\mathrm{E}\left[R^2\right] \geq \mathrm{E}^2[R]$.

Proof. We first defined $\mathrm{Var}[R]$ as an average of a squared expression, so $\mathrm{Var}[R]$ is nonnegative. Then we proved that $\mathrm{Var}[R] = \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]$. This implies that $\mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]$ is nonnegative. Therefore, $\mathrm{E}\left[R^2\right] \geq \mathrm{E}^2[R]$. □

In words, the expectation of a square is at least the square of the expectation. The two are equal exactly when the variance is zero:

$$\mathrm{E}\left[R^2\right] = \mathrm{E}^2[R] \text{ iff } \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R] = 0 \text{ iff } \mathrm{Var}[R] = 0.$$

4.3.2 Zero Variance

When does a random variable, R, have zero variance? . . . when the random variable *never* deviates from the mean!

Lemma 4.6. *The variance of a random variable, R, is zero if and only if $R = \mathrm{E}[R]$ for all outcomes with positive probability.*

The final phrase is a technicality; for an outcome with zero probability, R can take on any value without affecting the variance.

Proof. By the definition of variance, $\mathrm{Var}[R] = 0$ is equivalent to the condition $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right] = 0$.

The inner expression, $(R - \mathrm{E}[R])^2$, is always nonnegative because of the square. As a result, $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right] = 0$ if an only if $(R - \mathrm{E}[R])^2 = 0$ for all outcomes with positive probability. Now, the conditions $(R - \mathrm{E}[R])^2 = 0$ and $R = \mathrm{E}[R]$ are also equivalent. Therefore, $\mathrm{Var}[R] = 0$ iff $R = \mathrm{E}[R]$ for all outcomes with positive probability. □

4.3.3 Dealing with Constants

The following theorem describes how the variance of a random variable changes when it is scaled or shifted by a constant.

Theorem 4.7. *Let R be a random variable, and let a and b be constants. Then*

$$\mathrm{Var}[aR + b] = a^2 \mathrm{Var}[R]. \tag{4}$$

This theorem makes two points. First, adding a constant b to a random variable does not affect the variance. Second, multiplying a random variable by a constant changes the variance by a *square factor*.

Proof. We will transform the left side of (4) into the right side. The first step is to expand $\mathrm{Var}[aR + b]$ using the alternate definition of variance.

$$\mathrm{Var}[aR + b] = \mathrm{E}\left[(aR + b)^2\right] - \mathrm{E}^2[aR + b].$$

We will work on the first term and then the second term. For the first term, note that by linearity of expectation,

$$\mathrm{E}\left[(aR + b)^2\right] = \mathrm{E}\left[a^2R^2 + 2abR + b^2\right] = a^2 \mathrm{E}\left[R^2\right] + 2ab \mathrm{E}[R] + b^2. \tag{5}$$

Similarly for the second term:

$$\mathrm{E}^2[aR + b] = (a \mathrm{E}[R] + b)^2 = a^2\mathrm{E}^2[R] + 2ab \mathrm{E}[R] + b^2. \tag{6}$$

Finally, we substract the expanded second term from the first.

$$\begin{aligned}
\operatorname{Var}[aR+b] &= \operatorname{E}\left[(aR+b)^2\right] - \operatorname{E}^2[aR+b] && \text{(Theorem 4.3)}\\
&= a^2\operatorname{E}\left[R^2\right] + 2ab\operatorname{E}[R] + b^2 - \\
&\quad (a^2\operatorname{E}^2[R] + 2ab\operatorname{E}[R] + b^2) && \text{(by (5) and (6))}\\
&= a^2\operatorname{E}\left[R^2\right] - a^2\operatorname{E}^2[R]\\
&= a^2(\operatorname{E}\left[R^2\right] - \operatorname{E}^2[R])\\
&= a^2\operatorname{Var}[R] && \text{(Theorem 4.3)}
\end{aligned}$$

□

A similar rule holds for the standard deviation when a random variable is adjusted by a constant. Recall that standard deviation is the square root of variance. Therefore, adding a constant b to a random variable does not change the standard deviation. Multiplying a random variable by a constant a multiplies the standard deviation by a. So we have

Corollary 4.8. *The standard deviation of $aR+b$ equals a times the standard deviation of R.*

4.4 Variance of a Sum

Earlier, we claimed that for independent random variables, the variance of a sum is the sum of the variances. An independence condition is necessary. If we ignored independence, then we would conclude that $\operatorname{Var}[R+R] = \operatorname{Var}[R] + \operatorname{Var}[R]$. However, by Theorem 4.7, the left side is equal to $4\operatorname{Var}[R]$, whereas the right side is $2\operatorname{Var}[R]$. This implies that $\operatorname{Var}[R] = 0$, which, by Lemma 4.6, holds only if R is constant.

However, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations involving variables that are pairwise independent but not mutually independent. Matching birthdays is an example of this kind, as we shall see below.

Theorem 4.9. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\operatorname{Var}[R_1 + R_2 + \ldots + R_n] = \operatorname{Var}[R_1] + \operatorname{Var}[R_2] + \cdots + \operatorname{Var}[R_n].$$

Proof. By linearity of expectation, we have

$$\begin{aligned}
\operatorname{E}\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] &= \operatorname{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} R_i R_j\right]\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{E}[R_i R_j] && \text{(linearity)}\\
&= \sum_{1\le i\ne j\le n} \operatorname{E}[R_i]\operatorname{E}[R_j] + \sum_{i=1}^{n}\operatorname{E}\left[R_i^2\right]. && \text{(pairwise independence)} \qquad (7)
\end{aligned}$$

In (7), we use the fact from previous Notes that the expectation of the product of two independent variables is the product of their expectations.

Also,

$$\begin{aligned}
\mathrm{E}^2\left[\sum_{i=1}^{n} R_i\right] &= \left(\mathrm{E}\left[\sum_{i=1}^{n} R_i\right]\right)^2 \\
&= \left(\sum_{i=1}^{n} \mathrm{E}\left[R_i\right]\right)^2 && \text{(linearity)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] \\
&= \sum_{1\le i\ne j\le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right]. && (8)
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{Var}\left[\left(\sum_{i=1}^{n} R_i\right)\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] - \mathrm{E}^2\left[\sum_{i=1}^{n} R_i\right] && \text{(Theorem 4.3)} \\
&= \sum_{1\le i\ne j\le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}\left[R_i^2\right] - \\
&\quad \left(\sum_{1\le i\ne j\le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right]\right) && \text{(by (7) and (8))} \\
&= \sum_{i=1}^{n} \mathrm{E}\left[R_i^2\right] - \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right] \\
&= \sum_{i=1}^{n} (\mathrm{E}\left[R_i^2\right] - \mathrm{E}^2\left[R_i\right]) && \text{(reordering the sums)} \\
&= \sum_{i=1}^{n} \mathrm{Var}\left[R_i\right]. && \text{(Theorem 4.3)}
\end{aligned}$$

□

4.5 Variance of a Binomial Distribution

We now have enough tools to find the variance of a binomial distribution. Recall that if a random variable, R, has a binomial distribution, then

$$\Pr\{R = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

where n and p are parameters such that $n \ge 1$ and $0 < p < 1$.

We can think of R as the sum of n independent Bernoulli variables. For example, we can regard R as the number of heads that come up when we toss n independent coins, where each coin comes up heads with probability p. Formally, we can write $R = R_1 + R_2 + \cdots + R_n$ where

$$R_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

Now we can compute the variance of the binomially distributed variable R.

$$\begin{aligned}
\operatorname{Var}[R] &= \operatorname{Var}[R_1] + \operatorname{Var}[R_2] + \ldots + \operatorname{Var}[R_n] && \text{(Theorem 4.9)} \\
&= n \operatorname{Var}[R_1] && (\operatorname{Var}[R_i] = \operatorname{Var}[R_j]) \\
&= n(\mathrm{E}\left[R_1^2\right] - \mathrm{E}^2[R_1]) && \text{(Theorem 4.3)} \\
&= n(\mathrm{E}[R_1] - \mathrm{E}^2[R_1]) && (R_1^2 = R_1) \\
&= n(p - p^2). && (\mathrm{E}[R_1] = \Pr\{R_1 = 1\} = p)
\end{aligned}$$

This shows that the binomial distribution has variance $p(1-p)n$ and standard deviation $\sqrt{p(1-p)n}$. In the special case of an unbiased binomial distribution ($p = 1/2$), the variance is $n/4$ and the standard deviation is $\sqrt{n}/2$.

5 Applications of Chebyshev's Theorem

There is a nice reformulation of Chebyshev's Theorem in terms of standard deviation.

Corollary 5.1. *Let R be a random variable, and let c be a positive real number.*

$$\Pr\{|R - \mathrm{E}[R]| \geq c\sigma_R\} \leq \frac{1}{c^2}.$$

Here we see explicitly how the "likely" values of R are clustered in an $O(\sigma_R)$-sized region around $\mathrm{E}[R]$, confirming that the standard deviation measures how spread out the distribution of R is around its mean.

Proof. Substituting $x = c\sigma_R$ in Chebyshev's Theorem gives:

$$\Pr\{|R - \mathrm{E}[R]| \geq c\sigma_R\} \leq \frac{\operatorname{Var}[R]}{(c\sigma_R)^2} = \frac{\sigma_R^2}{(c\sigma_R)^2} = \frac{1}{c^2}.$$

□

5.1 I.Q. Example

Suppose that, in addition to the average I.Q. being 100, we also know the standard deviation of I.Q.'s is 10. How rare is an I.Q. of 200 or more?

Let the random variable, R, be the I.Q. of a random person. So we are supposing that $\mathrm{E}[R] = 100$, $\sigma_R = 10$, and R is nonnegative. We want to compute $\Pr\{R \geq 200\}$.

We have already seen that Markov's Theorem 2.2 gives a coarse bound, namely,

$$\Pr\{R \geq 200\} \leq \frac{1}{2}.$$

Now we apply Chebyshev's Theorem to the same problem:

$$\Pr\{R \geq 200\} = \Pr\{|R - 100| \geq 100\} \leq \frac{\text{Var}[R]}{100^2} = \frac{10^2}{100^2} = \frac{1}{100}.$$

The purpose of the first step is to express the desired probability in the form required by Chebyshev's Theorem; the equality holds because R is nonnegative. Chebyshev's Theorem then yields the inequality.

So Chebyshev's Theorem implies that at most one person in a hundred has an I.Q. of 200 or more. We have gotten a much tighter bound using the additional information, namely the variance of R, than we could get knowing only the expectation.

5.2 A One-Sided Bound

Chebyshev's Theorem gives a "two-sided bound". That is, it bounds the probability that a random variable deviates *above or below* the mean by some amount. What if we want only a one-sided bound? For example, what is the probability that a random variable deviates *above* the mean by some amount?

This question is often answered incorrectly. The erroneous argument runs as follows. By Chebyshev's Theorem, R deviates above or below the mean by some amount with probability p. Therefore, R deviates above the mean by this amount with probability $p/2$, and R deviates below the mean by this amount with probability $p/2$.

While this argument is correct for a probability distribution function that is symmetric about the mean, it is not correct for random variables that are more likely to deviate above the mean than below. For example, in the I.Q. question, some people deviate 100 points above the mean; that is, there are people with I.Q. greater than 200. However, by assumption everyone has a positive I.Q.; no one deviates more than 100 points below the mean. For this reason it turns out we could actually improve the bound of Section 5.1 slightly—from 1 in 100 to 1 in 101. In general, there is a Chebyshev bound for the one-sided case that slightly improves our two-sided bound, but we don't need to go into it.

6 Deviation of Repeated Trials

Using Chebyshev's Theorem and the facts about variance and expectation, we are finally in a position to be show how the average of many trials approaches the mean.

6.1 Estimation from Repeated Trials

For example, suppose we want to estimate the fraction of the U.S. voting population who would favor Al Gore over George Bush in the year 2004 presidential election. Let p be this unknown

fraction. Let's suppose we have some random process—say throwing darts at voter registration lists—which will select each voter with equal probability. Now we can define a Bernoulli variable, G, by the rule that $G = 1$ if a random voter most prefers Gore, and $G = 0$ otherwise. In this case, $G = G^2$, so

$$\mathrm{E}\left[G^2\right] = \mathrm{E}\left[G\right] = \Pr\left\{G = 1\right\} = p,$$

and

$$\mathrm{Var}\left[G\right] = \mathrm{E}\left[G^2\right] - \mathrm{E}^2\left[G\right] = p - p^2 = p(1-p).$$

To estimate p, we take a large number, n, of sample voters and count the fraction who favor Gore. We can describe this estimation as taking independent Bernoulli variables $G_1, G_2, \ldots, G_n$, each with the same expectation as G, computing their sum

$$S_n ::= \sum_{i=1}^{n} G_i, \tag{9}$$

and then using the average, S_n/n, as our estimate of p.

More generally, we can consider *any* set of random variables $G_1, G_2, \ldots, G_n$, with the same mean, μ, and likewise use the average, S_n/n, to estimate μ. One of the properties of S_n/n that is critical for this purpose is that S_n/n has the same expectation as the G_i's, namely,

$$\mathrm{E}\left[\frac{S_n}{n}\right] = \mu, \tag{10}$$

Proof.

$$\begin{aligned}
\mathrm{E}\left[\frac{S_n}{n}\right] &= \mathrm{E}\left[\frac{\sum_{i=1}^{n} G_i}{n}\right] && \text{(by def. (9) of } S_n\text{)} \\
&= \frac{\sum_{i=1}^{n} \mathrm{E}\left[G_i\right]}{n} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^{n} \mu}{n} \\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

□

Note that the random variables G_i need not be Bernoulli or even independent for (10) to hold, because linearity of expectation always holds.

Now suppose the G_i's also have the same deviation, σ. The second critical property of S_n/n is that

$$\mathrm{Var}\left[\frac{S_n}{n}\right] = \frac{\sigma^2}{n}. \tag{11}$$

This follows as long as the variance of S_n is the sum of the variances of the G_i's. For example, by Theorem 4.9, the variances can be summed if the G_i's are pairwise independent. Then we calculate:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \frac{1}{n^2}\,\text{Var}\left[S_n\right] && \text{(Theorem 4.7)}\\
&= \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def (9) of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[G_i\right] && \text{(variances assumed to add)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}$$

This is enough to apply Chebyshev's Bound and conclude:

Theorem 6.1. *[Pairwise Independent Sampling] Let*

$$S_n ::= \sum_{i=1}^{n} G_i$$

where $G_1, \ldots, G_n$ are pairwise independent variables with the same mean, μ, and deviation, σ. Then

$$\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2. \tag{12}$$

Proof.

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} &\leq \frac{\text{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound, Theorem 3.3)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (11))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

□

Theorem 6.1 finally provides a precise statement about how the average of independent samples of a random variable approaches the mean. It generalizes to many cases when S_n is the sum of independent variables whose mean and deviation are not necessarily all the same, though we shall not develop such generalizations here.

6.2 Birthdays again

We observed in lecture that the expected number of matching birthday pairs among n people was $\binom{n}{2}/365$. But how close to this expected value can we expect a typical sample to be? We can apply the Pairwise Independent Sampling Theorem to answer this question.

Now having matching birthdays for different pairs of students are not mutually independent events. For example, knowing that Alice and Bob have matching birthdays, and also that Ted and Alice have matching birthdays obviously implies that Bob and Ted have matching birthdays. On the other hand, knowing that Alice and Bob have matching birthdays tells us nothing about whether Alice and Carol have matching birthdays, *viz.*, these two events really are independent. We already already observed this phenomenon in Notes 11-12, §4.3.1, for the case of matching pairs among three coins. So even though the events that a pair of students have matching birthdays are not mutually independent, indeed not even three-way independent, they are *pairwise* independent.

This allows us to apply the Sampling Theorem. Let $B_1, B_2, \ldots, B_n$ be the birthdays of the n people, let $E_{i,j}$ be the indicator variable for the event $[B_i = B_j]$. For $i \neq j$, the probability that $B_i = B_j$ is $1/365$, so $\mathrm{E}[E_{i,j}] = 1/365$.

Now let M_n be the number of matching pairs, *i.e.*,

$$M_n ::= \sum_{1 \le i < j \le n} E_{i,j}. \tag{13}$$

So by linearity of expectation

$$\mathrm{E}[M_n] = \mathrm{E}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] = \sum_{1 \le i < j \le n} \mathrm{E}[E_{i,j}] = \binom{n}{2}\frac{1}{365},$$

as we noted above. Also, by linearity of variance for pairwise independent variables

$$\mathrm{Var}[M_n] = \mathrm{Var}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] = \sum_{1 \le i < j \le n} \mathrm{Var}[E_{i,j}] = \binom{n}{2}\frac{1}{365}\left(1 - \frac{1}{365}\right).$$

Now for our 6.042 class of 146 students, we have $\mathrm{E}[M_{146}] = 29$ and $\mathrm{Var}[M_{146}] = 29(1 - 1/365) < 29$. So by Theorem 6.1,

$$\Pr\{|M_{146} - 29| \ge x\} < \frac{29}{x^2}.$$

Letting $x = 8$, we conclude that there is a better than 50% chance that in a class of 146 students, the number of pairs of students with the same birthday will be between 21 and 37. In our class, we actually found that there were 17 matching pairs and 2 triples, for a total of 23 matching pairs.

6.3 Size of a Poll

Theorem 6.1 allows us to calculate poll size. How many people should we poll to get a reliable estimate of voters' preference?

Suppose, in particular, we want to know within tolerance $x ::= 0.02$ what fraction of the voters favor Gore. By choosing n large enough in Theorem 6.1 that we can reduce the probability that our estimate is off by more than x to as close to zero as we please.

For example, ninety-five per cent "confidence level" is a standard used in many statistical applications. So let's suppose we want our estimate of p to be within the tolerance 95% of the time, that is, with probability 0.95. Then we choose n so that $(1/n)(\sigma/x)^2 \leq 1 - 0.95$. That is, we want

$$n \geq \frac{\sigma^2}{(0.02)^2(1-0.95))} = \frac{p(1-p)}{0.00002} = 50,000p(1-p).$$

Solving for the sample size n in terms of the unknown p that we are trying to estimate in the first place may not seem to be making progress, but it's easy to see that the maximum value of $p(1-p)$ in the interval $0 \leq p \leq 1$ occurs at $p = 1/2$. So we conclude that if we sample

$$n \geq 50,000(1-1/2)1/2 = 12,500$$

voters, we can say that 95% of the time, our estimate $S_{12,500}/12,500$ will be within 0.02 of the fraction of voters who favor Gore.

Note that this bound on poll size holds regardless of how large the total voting population may be—whether we are trying to determine the preferences of a few tens of thousands of voters in a small city like Cambridge, or of the tens of millions of voters in a large nation like the U.S., the same poll size is adequate.

6.4 Confidence Levels

Now suppose a pollster dutifully checks with 12,500 randomly chosen voters and finds that 6,300 prefer Gore. It's tempting, but sloppy, to say that this means "With probability 0.95, the fraction, p, of voters who prefer Gore is $6300/12,500 = 0.504 \pm 0.02$."

What's objectionable about this statement is that it talks about the probability of a real world fact, namely the actual value of the fraction p. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.53; then it's nonsense to ask about the probability that it is within 0.02 of 0.504—it simply isn't.

A more careful summary of what we have accomplished goes this way: we have described a probabilistic procedure for estimating the actual value of the fraction p. The probability that *our estimation procedure* will yield a value within 0.02 of p is 0.95. This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that "At the 95% *confidence level*, the fraction of voters who prefer Gore is 0.504 ± 0.02."

Actually, polling 12,500 voters is excessive. We derived this bound on poll size solely by applying Chebyshev's Theorem to the value of the variance of S_n/n. But in fact we know the exact distribution of S_n, namely, it has a binomial distribution with parameters n, p. In the next section we do a more detailed calculation of probabilities of deviation from the mean specifically for the binomial distribution; we can show that the poll size need only be about $1/5$ of the size derived from the Chebyshev bound.

6.5 Better Polling

Let ϵ be the acceptable error tolerance of our poll. In the previous section we chose $\epsilon = 0.02$. We can define δ, the probability that our poll is off by more than ϵ as follows:

$$\begin{aligned}\delta \quad ::= \quad & \underbrace{\Pr\left\{\frac{S_n}{n} < p - \epsilon\right\}}_{\substack{\text{too many in sample} \\ \text{say "Bush"}}} \quad + \quad \underbrace{\Pr\left\{\frac{S_n}{n} > p + \epsilon\right\}}_{\substack{\text{too many in sample} \\ \text{say "Gore"}}} \\ = \quad & \Pr\{S_n < (p-\epsilon)n\} + \Pr\{S_n > (p+\epsilon)n\}.\end{aligned}$$

Since S_n has the binomial distribution with parameters n and p, the two terms in the definition of δ can be bounded using the bound (14) on $F_{n,p}$ from Notes 10:

Lemma.

$$F_{n,p}(\alpha n) \le \left(\frac{1-\alpha}{1-\alpha/p}\right) f_{n,p}(\alpha n) \tag{14}$$

for $\alpha < p$.

To ensure that $\alpha < p$, we observe that

$$\Pr\left\{\frac{S_n}{n} > p + \epsilon\right\} = \Pr\left\{\frac{n - S_n}{n} < 1 - p - \epsilon\right\},$$

where $(n - S_n)/n$ is the fraction of people polled who say that they prefer Bush, and $1 - p$ is the fraction of all Americans who prefer Bush. This gives

$$\delta \le F_{n,p}((p-\epsilon)n) + F_{n,1-p}((1-p-\epsilon)n). \tag{15}$$

As in the previous section, the bound (15) contains p, the fraction of Americans that favor Gore, which is the number we are trying to determine by polling. But as before, the worst case for the bound is when $p = 1/2$, though we shall not prove this. So we get

$$\delta \le F_{n,\frac{1}{2}}((\frac{1}{2} - \epsilon)n) + F_{n,1-\frac{1}{2}}((1 - \frac{1}{2} - \epsilon)n) = 2F_{n,\frac{1}{2}}((\frac{1}{2} - \epsilon)n). \tag{16}$$

Now plugging in $\epsilon = 0.02$ into (16) gives:

$$\begin{aligned}\delta \le 2F_{n,\frac{1}{2}}(0.48n) &\le 2 \cdot \frac{1-\alpha}{1-2\alpha} f_{n,1/2}(0.48n) \\ &\approx 2 \cdot 13 \cdot 2^{-n(1-H(0.48))} / \sqrt{2\pi \cdot 0.48(1-0.48)n} \\ &= 26 \cdot \frac{2^{-0.00115n}}{1.2523\sqrt{n}}.\end{aligned}$$

We want to poll enough people so that δ is less than 0.05. The easiest way is to plug in values for n, the number of people polled:

n = people polled	upper bound on probability poll is wrong
1000	29.4%
2000	9.3%
3000	3.4%
2500	5.6%
2750	4.4%
2616	5.004%
2617	4.999%

So polling 2617 people is sufficient to determine public opinion to within 2% with confidence of 95%. Again, the remarkable point is that the population of the country has no effect on the poll size. Whether there are ten thousand people or a billion in a country, polling 2617 people is sufficient!

Here we got a much better estimate of probable deviation from the mean using the fact that the samples were independent—and hence that the sampling distribution was binomial—than in the previous section using the Pairwise Independent Sampling Theorem 6.1. This should not be surprising, since the Sampling Theorem is based on Chebyshev's bound, and we've already seen that the Chebyshev bound can be much weaker than bounds derived using more information about a density function than simply its variance.

However, there are situations—matching birthdays is a good example—where mutual independence of the samples doesn't hold, but pairwise independence does, and that's where the Pairwise Independent Sampling Theorem becomes our main handle on predicting sample deviations.

7 Proof of the Weak Law

An equivalent way to state the conclusion of the Weak Law of Large Numbers, Theorem 2.1, is that the probability that the average *differs* from the expectation by more than any given tolerance approaches zero.

Theorem 7.1. *[Weak Law of Large Numbers] Let*

$$S_n ::= \sum_{i=1}^{n} G_i,$$

where $G_1, \ldots, G_n, \ldots$ are pairwise independent variables with the same expectation, μ and standard deviation, σ. For any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right\} = 0.$$

Proof. Choose x in Theorem 6.1 to be ϵ. Then, given any $\delta > 0$, choose n large enough to make $(\sigma/x)^2/n < \delta$. By Theorem 6.1,

$$\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right\} < \delta.$$

So the limiting probability must equal zero. □

Notice that this version of the Weak Law is slightly different from the version we first stated in Theorem 2.1. Theorem 7.1 requires that the G_i's have finite variance but Theorem 2.1 only requires finite expectation. On the other hand, the original version 2.1 requires mutual independence, while Theorem 7.1 requires only pairwise independence. The case when the variance may be infinite is not important to us, and we will not try to prove it.

A weakness of both the Weak Law as well as our Pairwise Independence Sampling Theorem 6.1 is that neither provides any information about the way the average value of the observations may be expected to *oscillate* in the course of repeated experiments. In later Notes we will briefly consider a *Strong* Law of Large Numbers which deals with the oscillations. Such oscillations may not be important in our example of polling about Gore's popularity or of birthday matches, but they are critical in gambling situations, where large oscillations can bankrupt a player, even though the player's average winnings are assured in the long run. As the famous economist Keynes is alleged to have remarked, the problem is that "In the long run, we are all dead."

8 Random Walks and Gamblers' Ruin

Random Walks nicely model many natural phenomena in which a person, or particle, or process takes steps in a randomly chosen sequence of directions. For example in Physics, three-dimensional random walks are used to model Brownian motion and gas diffusion. In Computer Science, the Google search engine uses random walks through the graph of world-wide web links to determine the relative importance of websites. In Finance Theory, there is continuing debate about the degree to which one-dimensional random walks can explain the moment-to-moment or day-to-day fluctuations of market prices. In these Notes we consider 1-dimensional random walks: walks along a straight line. Our knowledge of expectation and deviation will make 1-dimensional walks easy to analyze, but even these simple walks exhibit probabilistic behavior that can be astonishing.

In the Mathematical literature, random walks are for some reason traditionally discussed in the context of some social vice. A one-dimensional random walk is often described as the path of a drunkard who randomly staggers left or right at each step. In the rest of these Notes, we examine one-dimensional random walks using the language of gambling. In this case, a position during the walk is a gambler's cash-on-hand or *capital*, and steps on the walk are bets whose random outcomes increase or decrease his capital. We will be interested in two main questions:

1. What is the probability that the gambler wins?

2. How long must the gambler expect to wait for the walk to end?

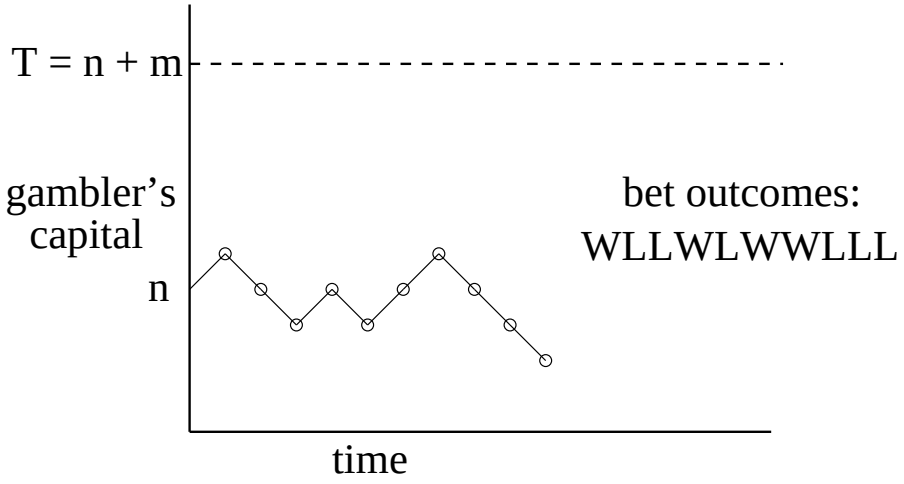

Figure 2: *This is a graph of the gambler's capital versus time for one possible sequence of bet outcomes. At each time step, the graph goes up with probability p and down with probability $1-p$. The gambler continues betting until the graph reaches either 0 or $T = n + m$.*

In particular, we suppose a gambler starts with n dollars. He makes a sequence of \$1 bets. If he wins an individual bet, he gets his money back plus another \$1. If he loses, he loses the \$1. In each bet, he wins with probability $p > 0$ and loses with probability $q ::= 1 - p > 0$. The gambler plays until either he is bankrupt or increases his capital to a goal amount of T dollars. If he reaches his goal, then he is called an overall *winner*, and his *profit* will be $m ::= T - n$ dollars. If his capital reaches zero dollars before reaching his gaol, then we say that he is "ruined" or *goes broke*.

The gambler's situation as he proceeds with his \$1 bets is illustrated in Figure 2. The random walk has boundaries at 0 and T. If the random walk ever reaches either of these boundary values, then it terminates. We want to determine the probability, w, that the walk terminates at boundary T, namely, the probability that the gambler is a winner.

In a fair game, $p = q = 1/2$. The corresponding random walk is called *unbiased*. The gambler is more likely to win if $p > 1/2$ and less likely to win if $p < 1/2$; the corresponding random walks are called *biased*.

Example 8.1. Suppose that the gambler is flipping a coin, winning \$1 on Heads and losing \$1 on Tails. Also, the gambler's starting capital is $n = 500$ dollars, and he wants to make $m = 100$ dollars. That is, he plays until he goes broke or reaches a goal of $T = n + m = \$600$. What is the probability that he is a winner? We will show that in this case the probability $w = 5/6$. So his chances of winning are really very good, namely, 5 chances out of 6.

Now suppose instead, that the gambler chooses to play roulette in an American casino, always betting \$1 on red. A roulette wheel has 18 black numbers, 18 red numbers, and 2 green numbers. In this game, the probability of winning a single bet is $p = 18/38 \approx 0.47$. It's the two green numbers that slightly bias the bets and give the casino an edge. Still, the bets are almost fair, and you might expect that the gambler has a reasonable chance of reaching his goal—the $5/6$ probability of winning in the unbiased game surely gets reduced, but perhaps not too drastically. Not so! His odds of winning against the "slightly" unfair roulette wheel are less than 1 in 37,000. If that seems surprising, listen to this: *no matter how much money* the gambler has to start, *e.g.*, \$5000, \$50,000, $\$5 \cdot 10^{12}$, his odds are still less than 1 in 37,000 of winning a mere 100 dollars!

Moral: Don't play!

The theory of random walks is filled with such fascinating and counter-intuitive conclusions.

9 The Probability Space

Each random-walk game corresponds to a path like the one in Figure 2 that starts at the point $(n, 0)$. A winning path never touches the x axis and ends when it first touches the line $y = T$. Likewise, a losing path never touches the line $y = T$ and ends when it first touches the x axis.

Any length k path can be characterized by the history of wins and losses on individual \$1 bets, so we use a length k string of W's and L's to model a path, and assign probability $p^r q^{k-r}$ to a string that contains r W's. The *outcomes* in our sample space will be precisely those string corresponding to winning or losing walks.

What about the infinite walks in which the gambler plays forever, neither reaching his goal nor going bankrupt? We saw in an in-class problem that the probability of playing forever is zero, so we don't need to include any such outcomes in our sample space.

As a sanity check on this definition of the probability space, we should verify that the sum of the outcome probabilities is one, but we omit this calculation.

10 The Probability of Winning

10.1 The Unbiased Game

Let's begin by considering the case of a fair coin, that is, $p = 1/2$, and determine the probability, w, that the gambler wins. We can handle this case by considering the expectation of the random variable G equal to the gambler's dollar gain. That is, $G = T - n$ if the gambler wins, and $G = -n$ if the gambler loses, so

$$\mathrm{E}[G] = w(T - n) - (1 - w)n = wT - n.$$

Notice that we're using the fact that the only outcomes are those in which the gambler wins or loses—there are no infinite games—so the probability of losing is $1 - w$.

Now let G_i be the amount the gambler gains on the ith flip: $G_i = 1$ if the gambler wins the flip, $G_i = -1$ if the gambler loses the flip, and $G_i = 0$ if the game has ended before the ith flip. Since the coin is fair, $\mathrm{E}[G_i] = 0$.

The random variable G is the sum of all the G_i's, so by linearity of expectation[2]

$$wT - n = E(G) = \sum_{i=1}^{\infty} E(G_i) = 0,$$

which proves

Theorem 10.1. *In the unbiased Gambler's Ruin game with probability $p = 1/2$ of winning each individual bet, with initial capital, n, and goal, T,*

$$\Pr\{\textit{the gambler is a winner}\} = \frac{n}{T}. \tag{17}$$

Example 10.2. Suppose we have \$100 and we start flipping a fair coin, betting \$1 with the aim of winning \$100. Then the probability of reaching the \$200 goal is $100/200 = 1/2$—the same as the probability of going bankrupt. In general, if $T = 2n$, then the probability of doubling your money or losing all your money is the same. This is about what we would expect.

Example 10.3. Suppose we have \$500 and we start flipping a fair coin, betting \$1 with the aim of winning \$100. So $n = 500, T = 600$, and $\Pr\{\text{win}\} = 500/600 = 5/6$, as we claimed at the outset.

Example 10.4. Suppose Albert starts with \$100, and Radhi starts with \$10. They flip a fair coin, and every time a Head appears, Albert wins \$1 from Radhi, and vice versa for Tails. They play this game until one person goes bankrupt. What is the probability of Albert winning?

This problem is identical to the Gambler's Ruin problem with $n = 100$ and $T = 100+10 = 110$. The probability of Albert winning is $100/110 = 10/11$, namely, the ratio of his wealth to the combined wealth. Radhi's chances of winnning are $1/11$.

Note that although Albert will win most of the time, the game is still fair. When Albert wins, he only wins \$10; when he loses, he loses big: \$100. Albert's—and Radhi's—expected win is zero dollars.

Another intuitive idea is confirmed by this analysis: the larger the gambler's initial stake, the larger the probability that he will win a fixed amount.

Example 10.5. If the gambler started with one million dollars instead of 500, but aimed to win the same 100 dollars as in the Example 10.3, the probability of winning would increase to $1M/(1M + 100) > .9999$.

[2]We've been stung by paradoxes in this kind of situation, so we should be careful to check that the condition for infinite linearity of expectation is satisfied. Namely, we have to check that $\sum_{i=1}^{\infty} \mathrm{E}[|G_i|]$ converges.

In this case, $|G_i| = 1$ iff the walk is of length at least i, and $|G_i| = 0$ otherwise. So

$$\mathrm{E}[|G_i|] = \Pr\{\text{the walk is of length } \geq i\}.$$

But we show in an in-class problem that there is a constant $r < 1$ such that

$$\Pr\{\text{the walk is of length } \geq i\} \leq \Theta(r^i).$$

So the $\sum_{i=1}^{\infty} \mathrm{E}[|G_i|]$ is bounded term-by-term by a convergent geometric series, and therefore it also converges.

10.2 A Recurrence for the Probability of Winning

To handle the case of a biased game we need a more general approach. We consider the probability of the gambler winning as a function of his initial capital. That is, let p and T be fixed, and let w_n be the gambler's probabiliity of winning when his initial capital is n dollars. For example, w_0 is the probability that the gambler will win given that he starts off broke; clearly, $w_0 = 0$. Likewise, $w_T = 1$.

Otherwise, the gambler starts with n dollars, where $0 < n < T$. Consider the outcome of his first bet. The gambler wins the first bet with probability p. In this case, he is left with $n+1$ dollars and becomes a winner with probability w_{n+1}. On the other hand, he loses the first bet with probability $1-p$. Now he is left with $n-1$ dollars and becomes a winner with probability w_{n-1}. Overall, he is a winner with probability $w_n = pw_{n+1} + qw_{n-1}$. Solving for w_{n+1} we have

$$w_{n+1} = \frac{w_n}{p} - w_{n-1}\frac{q}{p}. \tag{18}$$

This kind of inductive definition of a quantity w_{n+1} in terms of a linear combination of values w_k for $k < n+1$ is called a *homogeneous linear recurrence*. There is a simple general method for solving such recurrences which we now illustrate. The method is based on a guess that the form of the solution is $w_n = c^n$ for some $c > 0$. It's not obvious why this is a good guess, but we now show how to find the constant c and verify the guess.

Namely, from (18) we have

$$w_{n+1} - \frac{w_n}{p} + w_{n-1}\frac{q}{p} = 0. \tag{19}$$

If our guess is right, then this is equivalent to

$$c^{n+1} - \frac{c^n}{p} + c^{n-1}\frac{q}{p} = 0.$$

Now factoring out c^{n-1} gives

$$c^2 - \frac{c}{p} + \frac{q}{p} = 0.$$

Solving this quadratic equation in c yields two roots, $(1-p)/p$ and 1. So if we define $w_n ::= ((1-p)/p)^n = (q/p)^n$, then (19), and hence (18) is satisifed. We can also define $w_n ::= 1^n$ and satisfy (19). Since the lefthand side of (19) is zero using either definition, it follows that any definition of the form

$$w_n ::= A\left(\frac{q}{p}\right)^n + B \cdot 1^n$$

will also satisfy (19). Now our boundary conditions, namely the values of w_0 and w_T, let us solve for A and B:

$$\begin{aligned} 0 &= w_0 = A + B, \\ 1 &= w_T = A\left(\frac{q}{p}\right)^T + B, \end{aligned}$$

so

$$A = \frac{1}{(q/p)^T - 1}, \qquad B = -A, \tag{20}$$

and therefore

$$w_n = \frac{(q/p)^n - 1}{(q/p)^T - 1}. \tag{21}$$

Now we could verify our guess work and prove (21) by a routine induction on n which we omit.

The solution (21) only applies to biased walks, since we require $p \neq q$ so the denominator is not zero. That's ok, since we already worked out that the case when $p = q$ in Theorem 10.1. So we have shown:

Theorem 10.6. *In the biased Gambler's Ruin game with probability, $p \neq 1/2$, of winning each bet, with initial capital, n, and goal, T,*

$$\Pr\{\textit{the gambler is a winner}\} = \frac{(q/p)^n - 1}{(q/p)^T - 1}. \tag{22}$$

The expression (22) for the probability that the Gambler wins in the biased game is a little hard to interpret. There is a simpler upper bound which is nearly tight when the gambler's starting capital is large.

Suppose that $p < 1/2$; that is, the game is biased *against* the gambler. Then both the numerator and denominator in the quotient in (22) are positive, and the quotient is less than one. So adding 1 to both the numerator and denominator increases the quotient[3], and the bound (22) simplifies to $(q/p)^n/(q/p)^T = (p/q)^{T-n}$, which proves

Corollary 10.7. *In the Gambler's Ruin game biased against the Gambler, that is, with probability $p < 1/2$ of winning each bet, with initial capital, n, and goal, T,*

$$\Pr\{\textit{the gambler is a winner}\} < \left(\frac{p}{q}\right)^m, \tag{23}$$

where $m ::= T - n$.

The amount $m = T - n$ is called the Gambler's *intended profit*. So the gambler gains his intended profit, m, before going broke with probability at most $(p/q)^m$. Notice that this upper bound does not depend on the gambler's starting capital, but only on his intended profit. The consequences of this are amazing:

[3] If $0 < a < b$, then

$$\frac{a}{b} < \frac{a+1}{b+1},$$

because

$$\frac{a}{b} = \frac{a(1+1/b)}{b(1+1/b)} = \frac{a + a/b}{b+1} < \frac{a+1}{b+1}.$$

Example 10.8. Suppose that the gambler starts with \$500 aiming to profit \$100, this time by making \$1 bets on red in roulette. By (23), the probability, w_n, that he is a winner is less than

$$\left(\frac{18/38}{20/38}\right)^{100} = \left(\frac{9}{10}\right)^{100} < \frac{1}{37,648}.$$

This is a dramatic contrast to the unbiased game, where we saw in Example 10.3 that his probability of winning was 5/6.

Example 10.9. We also observed that with \$1,000,000 to start in the unbiased game, he was almost certain to win \$100. But betting against the "slightly" unfair roulette wheel, even starting with \$1,000,000, his chance of winning \$100 remains less than 1 in 37,648! He will almost surely lose all his \$1,000,000. In fact, because the bound (23) depends only on his intended profit, his chance of going up a mere \$100 is less than 1 in 37,648 *no matter how much money he starts with*!

The bound (23) is exponential in m. So, for example, doubling his intended profit will square his probability of winning.

Example 10.10. The probability that the gambler's stake goes up 200 dollars before he goes broke playing roulette is at most

$$(9/10)^{200} = ((9/10)^{100})^2 = \left(\frac{1}{37,648}\right)^2,$$

which is about 1 in 70 billion.

The odds of winning a little money are not so bad.

Example 10.11. Applying the exact formula (22), we find that the probability of winning \$10 before losing \$10 is

$$\frac{\left(\frac{20/38}{18/38}\right)^{10} - 1}{\left(\frac{20/38}{18/38}\right)^{20} - 1} = 0.2585\ldots$$

This is somewhat worse than the 1 in 2 chance in the fair game, but not dramatically so.

Thus, in the fair case, it helps a lot to have a large bankroll, whereas in the unfair case, it doesn't help much.

10.3 Intuition

Why is the gambler so unlikely to make money when the game is slightly biased against him? Intuitively, there are two forces at work. First, the gambler's capital has random upward and downward *swings* due to runs of good and bad luck. Second, the gambler's capital will have a steady, downward *drift*, because he has a small, negative expected return on every bet. The situation is shown in Figure 3.

For example, in roulette the gambler wins a dollar with probability $9/19$ and loses a dollar with probability $10/19$. Therefore, his expected return on each bet is $9/10 - 10/19 = -1/19 \approx -0.053$ dollars. That is, on each bet his capital is expect to drift downward by a little over 5 cents.

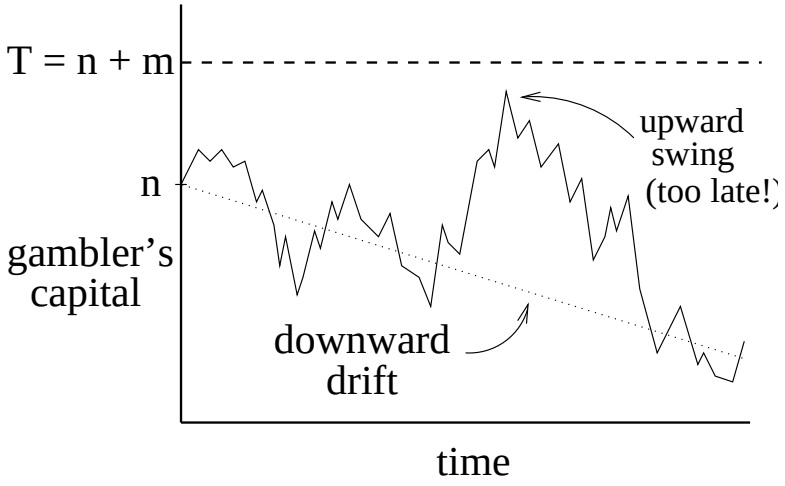

Figure 3: *In an unfair game, the gambler's capital swings randomly up and down, but steadily drifts downward. If the gambler does not have a winning swing early on, then his capital drifts downward, and later upward swings are insufficient to make him a winner.*

Our intuition is that if the gambler starts with a trillion dollars, then he will play for a very long time, so at some point there should be a lucky, upward swing that puts him \$100 ahead. The problem is that his capital is steadily drifting downward. If the gambler does not have a lucky, upward swing early on, then he is doomed. After his capital drifts downward a few hundred dollars, he needs a huge upward swing to save himself. And such a huge swing is extremely improbable. As a rule of thumb, *drift dominates swings* in the long term.

We can quantify these drifts and swings. After k rounds, the number of wins by our player has a binomial distribution with parameters $p < 1/2$ and k. His expected win on any single bet is $p - q = 2p - 1$ dollars, so his expected capital is $n - k(1 - 2p)$. Now to be a winner, his actual number of wins must exceed the expected number by $m + k(1 - 2p)$. But we saw before that the binomial distribution has a standard deviation of only $\sqrt{kp(1-p)}$. So for the gambler to win, he needs his number of wins to deviate by

$$\frac{m + k(1 - 2p)}{\sqrt{kp(1 - 2p)}} = \Theta(\sqrt{k})$$

times its standard deviation. In our study of binomial tails we saw that this was extremely unlikely.

In a fair game, there is no drift; swings are the only effect. In the absence of downward drift, our earlier intuition is correct. If the gambler starts with a trillion dollars then almost certainly there will eventually be a lucky swing that puts him \$100 ahead.

If we start with \$10 and play to win only \$10 more, then the difference between the fair and unfair games is relatively small. We saw that the probability of winning is $1/2$ versus about $1/4$. Since swings of \$10 are relatively common, the game usually ends before the gambler's capital can drift very far. That is, the game does not last long enough for drift to dominate the swings.

11 How Long a Walk?

Now that we know the probability, w_n, that the gambler is a winner in both fair and unfair games, we consider how many bets he needs on average to either win or go broke.

11.1 Duration of an Biased Walk

Let Q be the number of bets the gambler makes until the game ends. Since the gambler's expected win on any bet is $2p-1$, Wald's Theorem should tell us that his game winnings, G, will have expectation $\mathrm{E}[Q](2p-1)$. That is,

$$\mathrm{E}[G] = (2p-1)\,\mathrm{E}[Q], \tag{24}$$

In an unbiased game (24) is trivially true because both $2p-1$ and the expected overall winnings, $\mathrm{E}[G]$, are zero. On the other hand, in the unfair case, $2p-1 \neq 0$. Also, we know that

$$\mathrm{E}[G] = w_n(T-n) - (1-w_n)n = w_nT - n.$$

So assuming (24), we conclude

Theorem 11.1. *In the biased Gambler's Ruin game with initial capital, n, goal, T, and probability, $p \neq 1/2$, of winning each bet,*

$$\mathrm{E}[\textit{number of bets till game ends}] = \frac{\Pr\{\textit{gambler is a winner}\}\,T-n}{2p-1}. \tag{25}$$

The only problem is that (24) is not a special case of Wald's Theorem because $G = \sum_{i=1}^{Q} G_i$ is not a sum of *nonnegative* variables: when the gambler loses the ith bet, the random variable G_i equals -1. However, this is easily dealt with.[4]

Example 11.2. If the gambler aims to profit \$100 playing roulette with n dollars to start, he can expect to make $((n+100)/37, 648-n)/(2(18/38)-1) \approx 19n$ bets before the game ends. So he can enjoy playing for a good while before almost surely going broke.

[4]The random variable G_i+1 is nonnegative, and $\mathrm{E}[G_i+1 \mid Q \geq i] = \mathrm{E}[G_i \mid Q \geq i]+1 = 2p$, so by Wald's Theorem

$$\mathrm{E}\left[\sum_{i=1}^{Q}(G_i+1)\right] = 2p\,\mathrm{E}[Q]. \tag{26}$$

But

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{Q}(G_i+1)\right] &= \mathrm{E}\left[\sum_{i=1}^{Q} G_i + \sum_{i=1}^{Q} 1\right] \\
&= \mathrm{E}\left[(\sum_{i=1}^{Q} G_i) + Q\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{Q} G_i\right] + \mathrm{E}[Q] \\
&= \mathrm{E}[G] + \mathrm{E}[Q].
\end{aligned} \tag{27}$$

Now combining (26) and (27) confirms the truth of our assumption (24).

11.2 Duration of an Unbiased Walk

This time, we need the more general approach of recurrences to handle the unbiased case. We consider the expected number of bets as a function of the gambler's initial capital. That is, for fixed p and T, let e_n be the expected number of bets until the game ends when the gambler's initial capital is n dollars. Since the game is over in no steps if $n = 0$ or T, the boundary conditions this time are $e_0 = e_T = 0$.

Otherwise, the gambler starts with n dollars, where $0 < n < T$. Now by the conditional expectation rule, the expected number of steps can be broken down into the expected number of steps given the outcome of the first bet weighted by the probability of that outcome. That is,

$$e_n = p\,\mathrm{E}\left[Q \mid \text{gambler wins first bet}\right] + q\,\mathrm{E}\left[Q \mid \text{gambler loses first bet}\right].$$

But after the gambler wins the first bet, his capital is $n + 1$, so he can expect to make another e_{n+1} bets. That is,

$$\mathrm{E}\left[Q \mid \text{gambler wins first bet}\right] = 1 + e_{n+1},$$

and similarly,

$$\mathrm{E}\left[Q \mid \text{gambler loses first bet}\right] = 1 + e_{n-1}.$$

So we have

$$e_n = p(1 + e_{n+1}) + q(1 + e_{n-1}) = pe_{n+1} + qe_{n-1} + 1,$$

which yields the linear recurrence

$$e_{n+1} = \frac{e_n}{p} - \frac{q}{p}e_{n-1} - \frac{1}{p}.$$

For $p = q = 1/2$, this equation simplifies to

$$e_{n+1} = 2e_n - e_{n-1} - 2. \tag{28}$$

There is a general theory for solving linear recurrences like (28) in which the value at $n + 1$ is a linear combination of values at some arguments $k < n + 1$ plus another simple term—in this case plus the constant -2. This theory implies that

$$e_n = (T - n)n. \tag{29}$$

Fortunately, we don't need the general theory to *verify* this solution. Equation (29) can be verified routinely from the boundary conditions and (28) using strong induction on n.

So we have shown

Theorem 11.3. *In the unbiased Gambler's Ruin game with initial capital, n, and goal, T, and probability, $p = 1/2$, of winning each bet,*

$$\mathrm{E}\left[\textit{number of bets till game ends}\right] = n(T - n). \tag{30}$$

Another way to phrase Theorem 11.3 is

$$\mathrm{E}\left[\text{number of bets till game ends}\right] = \text{initial capital} \cdot \text{intended profit}. \tag{31}$$

Now for example, we can conclude that if the gambler starts with \$10 dollars and plays until he is broke or ahead \$10, then $10 \cdot 10 = 100$ bets are required on average. If he starts with \$500 and plays until he is broke or ahead \$100, then the expected number of bets until the game is over is $500 \times 100 = 50,000$.

Notice that (31) is a very simple answer that cries out for an intuitive proof, but we have not found one.

12 Quit While You Are Ahead

Suppose that the gambler never quits while he is ahead. That is, he starts with $n > 0$ dollars, ignores any goal T, but plays until he is flat broke. Then it turns out that if the game is not favorable, *i.e.*, $p \leq 1/2$, the gambler is sure to go broke. In particular, he is even sure to go broke in a "fair" game with $p = 1/2$. [5]

Lemma 12.1. *If the gambler starts with one or more dollars and plays a fair game until he is broke, then he will go broke with probability 1.*

Proof. If the gambler has initial capital n and goes broke in a game without reaching a goal T, then he would also go broke if he were playing and ignored the goal. So the probability that he will lose if he keeps playing without stopping at any goal T must be at least as large as the probability that he loses when he has a goal $T > n$.

But we know that in a fair game, the probability that he loses is $1 - n/T$. This number can be made arbitrarily close to 1 by choosing a sufficiently large value of T. Hence, the probability of his losing while playing without any goal has a lower bound arbitrarily close to 1, which means it must in fact be 1. □

So even if the gambler starts with a million dollars and plays a perfectly fair game, he will eventually lose it all with probability 1. In fact, if the game is unfavorable, then Theorem 11.1 and Corollary 10.7 imply that his expected time to go broke is essentially proportional to his initial capital, *i.e.*, $\Theta(n)$.

But there is good news: if the game is fair, he can "expect" to play for a very long time before going broke; in fact, he can expect to play forever!

Lemma 12.2. *If the gambler starts with one or more dollars and plays a fair game until he goes broke, then his expected number of plays is infinite.*

Proof. Consider the gambler's ruin game where the gambler starts with initial capital n, and let u_n be the expected number of bets for the *unbounded* game to end. Also, choose any $T \geq n$, and as above, let e_n be the expected number of bets for the game to end when the gambler's goal is T.

[5] If the game is favorable to the gambler, *i.e.*, $p > 1/2$, then we could show that there is a positive probability that the gambler will play forever, but we won't examine this case in these Notes.

The unbounded game will have a larger expected number of bets compared to the bounded game because, in addition to the possibility that the gambler goes broke, in the bounded game there is also the possibility that the game will end when the gambler reaches his goal, T. That is,

$$u_n \geq e_n.$$

So by (29),

$$u_n \geq n(T-n).$$

But $n \geq 1$, and T can be any number greater than or equal to n, so this lower bound on u_n can be arbitrarily large. This implies that u_n must be infinite.

Now by Lemma 12.1, with probability 1, the unbounded game ends when the gambler goes broke. So the expected time for the unbounded game to *end* is the *same* as the expected time for the gambler to *go broke*. Therefore, the expected time to go broke is infinite. □

In particular, even if the gambler starts with just one dollar, his expected number of plays before going broke is infinite! Of course, this does not mean that it is likely he will play for long. For example, there is a 50% chance he will lose the very first bet and go broke right away.

Lemma 12.2 says that the gambler can "expect" to play forever, while Lemma 12.1 says that with probability 1 he will go broke. These Lemmas sound contradictory, but our analysis showed that they are not.

13 Infinite Expectation

So what are we to make of such a random variable with infinite expectation? For example, suppose we repeated the experiment of having the gambler make fair bets with initial stake one dollar until he went broke, and we kept a record of the average number of bets per experiment. Our theorems about deviation from the mean only apply to random variables with finite expectation, so they don't seem relevant to this situation. But in fact they are.

For example, let Q be the number of bets required for the gambler to go broke in a fair game starting with one dollar. We could use some of our combinatorial techniques to show that

$$\Pr\{Q = m\} = \Theta(m^{-3/2}). \tag{32}$$

This implies that

$$\mathrm{E}[Q] = \Theta\left(\sum_{m=1}^{\infty} m \cdot m^{-3/2}\right) = \Theta\left(\sum_{m=1}^{\infty} m^{-1/2}\right).$$

We know this last series is divergent, so we have another proof that Q has infinite expectation.

But suppose we let $R ::= Q^{1/5}$. Then the estimate (32) also lets us conclude that

$$\mathrm{E}[R] = \Theta\left(\sum_{m=1}^{\infty} m^{-13/10}\right)$$

and

$$\mathrm{E}\left[R^2\right] = \Theta\left(\sum_{m=1}^{\infty} m^{-11/10}\right).$$

Since both these series are convergent, we can conclude that $\mathrm{Var}\left[R\right]$ is finite. Now our theorems about deviation can be applied to tell us that the average *fifth root* of the number of bets to go broke is very likely to converge to a finite expected value.

We won't go further into the details, but the moral of this discussion is that our results about deviation from a finite mean can still be applied to natural models like random walks where variables with infinite expectation may play an important role.

14 The Chernoff Bound

The Chernoff bound applies to a sum of independent random variables that satisfy conditions that lie between the conditions needed for the Pairwise Independent Sampling Theorem of Notes 11-12 and conditions that imply the sum has a binomial distribution. When it applies, the Chernoff bound gives nearly as good a bound as our estimates for the binomial distribution in Notes 11-12. In particular, the Chernoff bound is exponentially smaller than bound given by the Pairwise Independent Sampling Theorem.

The Chernoff bound plays a larger role in Computer Science than the more traditional Central Limit Theorem (which will be briefly considered in later Notes). Both theorems give bounds on deviation from the mean, but the Chernoff bound gives better estimates on the probability of deviating from the mean by many standard deviations.

For example, suppose we are designing a system whose components may occasionally fail, but we want the system as a whole to be very reliable. The Chernoff bound can provide good estimates for the number of failures the system should be designed to survive in order to meet the specified high level of reliability. That is, the system will only fail only if a the number of component failures exceeds a designated threshold, but the Chernoff bound tells us that this threshold is very unlikely to be exceeded.

Another typical application is in designing probabilistic algorithms. We expect that such algorithms might give a wrong answer, but will do so only if the number of mistaken probabilistic "guesses" it makes is much larger than should be expected. The likelihood of this unusually large number of mistakes can often be estimated well using the Chernoff bound.

15 The Probability of at Least One Event

Let $A_1, A_2, \ldots, A_n$ be a sequence of events, and let T be the number of these events that occur. What is $\Pr\{T \geq 1\}$, the probability that at least 1 event occurs? Note that the event $[T \geq 1]$ is precisely the same as the event $\bigcup A_i$. In Notes 10, §5.2, and in Class Problems 10W, Problem 1, we verified the general bounds

$$\max_{1\leq i\leq n} \Pr\{A_i\} \leq \bigcup A_i \leq \sum_{i=1}^{n} \Pr\{A_i\}, \tag{33}$$

and described situations in which each of these bounds were achieved. So in general, we cannot improve the bounds given in (33).

On the other hand, if the events A_i are mutually independent, we can be much more precise about the probability that one or more of them occur. In fact, we will show that if we expect several events to occur, then almost certainly at least one event will occur. Another way to say this is that if we expect more than one event to occur, then the probability that no event occurs is practically zero. Specifically, we have:

Theorem 15.1. *Let $A_1, A_2, \ldots A_n$ be independent events, and let T be the number of these events that occur. The probability that none of the events occur is at most $e^{-\mathrm{E}[T]}$.*

Interestingly, Theorem 15.1 does not depend on n, the number of events. It gives the same bound whether there are 100 events each with probability 0.1 or 1000 events each with probability 0.01. In both cases, the expected number of events is 10, and so the probability of no event occurring is at most e^{-10} or about 1 in 22,000. Note that the actual probabilities are somewhat different in these two cases, indicating that the given bound is not always tight.

Theorem 15.1 can be interpreted as a sort of "Murphy's Law": if we expect some things to go wrong, then something probably will. For example, suppose that we are building a microprocessor, and the fabrication process is such that each transistor is faulty mutually independently with a probability of one in a million. This sounds good. However, microprocessors now contain about ten million transistors, so the expected number of faulty transistors is 10 per chip. Since we expect some things to go wrong, something probably will. In fact, Theorem 15.1 implies that the probability of a defect-free a chip is less than 1 in 22,000!

In proving Theorem 15.1, we first note that

$$T = T_1 + T_2 + \cdots + T_n, \tag{34}$$

where T_i is the indicator variable for the event A_i. We also use the fact that

$$1 + x \leq e^x \tag{35}$$

for all x, which follows from the Taylor expansion

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

Proof.

$$\begin{aligned}
\Pr\{T=0\} &= \overline{A_1 \cup A_2 \cup \cdots \cup A_n} && \text{(def. of } T\text{)} \\
&= \Pr\left\{\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}\right\} && \text{(De Morgan's law)} \\
&= \prod_{i=1}^{n} \Pr\left\{\overline{A_i}\right\} && \text{(mutual independence of } A_i\text{'s)} \\
&= \prod_{i=1}^{n} 1 - \Pr\{A_i\} && \text{(complement rule)} \\
&\leq \prod_{i=1}^{n} e^{-\Pr\{A_i\}} && \text{(by (35))} \\
&= e^{-\sum_{i=1}^{n} \Pr\{A_i\}} && \text{(exponent algebra)} \\
&= e^{-\sum_{i=1}^{n} \mathrm{E}[T_i]} && \text{(expectation of indicator variable)} \\
&= e^{-\mathrm{E}[T]}. && \text{((34) \& linearity of expectation)}
\end{aligned}$$

□

Two special cases of Theorem 15.1 are worth singling out because they come up all the time.

Corollary 15.2. *Suppose an event has probability $1/m$. Then the probability that the event will occur at least once in m independent trials is at least approximately $1 - 1/e \approx 63\%$. There is at least 50% chance the event will occur in $n = m \log 2 \approx 0.69m$ trials.*

16 Chernoff Bounds

16.1 Probability of at least *k* events

Now we consider the more general question than the probability that one event occurs, namely, the probability that k events occur, still assuming mutual independence of the events A_i. In other words, what is $\Pr\{T \geq k\}$, given that the events A_i are mutually independent?

For example, suppose we want to know the probability that at least k heads come up in N tosses of a coin. Here A_i is the event that the coin is heads on the ith toss, T is the total number of heads, and $\Pr\{T \geq k\}$ is the probability that at least k heads come up.

As a second example, suppose that we want the probability of a student answering at least k questions correctly on an exam with N questions. In this case, A_i is the event that the student answers the ith question correctly, T is the total number of questions answered correctly, and $\Pr\{T \geq k\}$ is the probability that the student answers at least k questions correctly.

There is an important difference between these two examples. In the first example, all events A_i have equal probability, *i.e.*, the coin is as likely to come up heads on one flip as on another. So T has a binomial distribution whose tail bounds we have already characterized in Notes 11-12.

In the second example, however, some exam questions might be more difficult than others. If Question 1 is easier than Question 2, then the probability of event A_1 is greater than the probability

of event A_2. In this section we develop a method to handle this more general situation in which the events A_i may have different probabilities.

We will prove that the number of events that occur is almost never much greater than the expectation. This result is called the Chernoff Bound. For example, if we toss N coins, the expected number of heads is $N/2$ heads. The Chernoff Bound implies that for sufficiently large N, the number of heads is almost always not much greater than $N/2$.

A nice feature of the Chernoff Bound is that we do not even need to know the probability of each event A_i or even the number of events N; rather, we need only the expected number of events that occur and the fact that the events are mutually independent.

16.2 Statement of the Bound

We state Chernoff's Theorem in terms of Bernoulli variables instead of events. However, we can regard T_i as an indicator for the event A_i.

Theorem 16.1 (Chernoff Bound). *Let $T_1, T_2, \ldots, T_n$ be mutually independent Bernoulli variables, and let $T = T_1 + T_2 + \cdots + T_n$. Then for all $c \geq 1$, we have*

$$\Pr\{T \geq c\,\mathrm{E}[T]\} \leq e^{-(c\ln c - c + 1)\,\mathrm{E}[T]}. \tag{36}$$

The formula for the exponent in the bound is a little awkward. The situation is simpler when $c = e = 2.718\ldots$. In this case, $c\ln c - c + 1 = e\ln e - e + 1 = e\cdot 1 - e + 1 = 1$, so we have as an immediate corollary of Theorem 16.1:

Corollary 16.2. *Let $T_1, T_2, \ldots, T_n$ be mutually independent Bernoulli variables, and let $T = T_1 + T_2 + \cdots + T_n$. Then*

$$\Pr\{T \geq e\,\mathrm{E}[T]\} \leq e^{-\mathrm{E}[T]}.$$

We will prove the Chernoff Bound shortly. First, let's see an example of how it is used.

16.3 Example: Pick 4

There is a lottery game called Pick 4. In this game, each player picks 4 digits, defining a number in the range 0 to 9999. A winning number is drawn each week. The players who picked the winning number win some cash. A million people play the lottery, so the expected number of winners each week is

$$\frac{1}{10,000}\cdot 1,000,000 = 100.$$

However, on some lucky day thousands of people might all pick the winning number, costing the lottery operators loads of money. How likely is this?

Assume that all players pick numbers uniformly and mutually independently. Let T_i be an indicator variable for the event that the ith player picks the winning number. Let $T = T_1 + T_2 + \cdots + T_n$. Then T is the total number of players that pick the winning number. As noted above, an average

of 100 people win each week, so $\mathrm{E}[T] = 100$. We can use Corollary 16.2 to bound the probability that number of winners is greater than 272 as follows:

$$\Pr\{T \geq 272\} \leq \Pr\{T \geq e\,\mathrm{E}[T]\} \leq e^{-Ex(T)} = e^{-100}.$$

The probability of 272 or more people winning is absurdly small! It appears that the lottery operators should not worry that ten thousand people will pick correctly one day!

But there is a catch. The assumption that people choose Pick 4 numbers uniformly and mutually independently is empirically false; people choose certain "favorite" numbers far more frequently than others.

Chernoff used this fact in devising a scheme to actually make money on the lottery. In this case, a fraction of all money taken in by the lottery was divided up equally among the winners. A bad strategy would be to pick a popular number. Then, even if you pick the winning number, you must share the cash with many other players. A better strategy is to pick a lot of unpopular numbers. You are just as likely to win with an unpopular number, but will not have to share with anyone. Chernoff found that peoples' picks were so highly correlated that he could actually turn a 7% profit by picking unpopular numbers!

16.4 The constant in the exponent

For general c, what can we say about the factor $c\ln c - c + 1$ in the exponent? First, note that when $c = 1$, the exponent factor equals $1 \cdot 0 - 1 + 1 = 0$. This means that the Chernoff bound cannot say anything useful about the probability simply of exceeding the mean. However, the exponent factor increases with c for $c > 1$. This follows because its derivative respect to c is positive:

$$\frac{d(c\ln c - c + 1)}{dc} = (\frac{c}{c} + \ln c) - 1 = \ln c > 0$$

when $c > 1$. In particular, for any $c > 1$, the factor $(c\ln c - c + 1)$ in the exponent is positive.

Let's consider the case of c close to 1, say $c = 1 + \epsilon$. Then a Taylor expansion gives:

$$\begin{aligned} c\ln c - c + 1 &= (1+\epsilon)\ln(1+\epsilon) - (1+\epsilon) + 1 \\ &= (1+\epsilon)(\epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} - \frac{\epsilon^4}{4} + \cdots) - \epsilon \\ &= (\epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} - \frac{\epsilon^4}{4} + \cdots) + (\epsilon^2 - \frac{\epsilon^3}{2} + \frac{\epsilon^4}{3} - \frac{\epsilon^5}{4} + \cdots) - \epsilon \\ &= \frac{\epsilon^2}{2} - \frac{\epsilon^3}{2\cdot 3} + \frac{\epsilon^4}{3\cdot 4} - \frac{\epsilon^5}{4\cdot 5} + \cdots. \end{aligned}$$

In particular, for very small ϵ, we have $c\ln c - c + 1 \approx \epsilon^2/2$. In fact one can prove the following:

Lemma 16.3. *For any* $0 < \epsilon < 1$, $\Pr\{T \geq (1+\epsilon)\,\mathrm{E}[T]\} < e^{-\epsilon^2\,\mathrm{E}[T]/3}$.

In other words, the probability of deviation above the mean by a fraction ϵ decays *exponentially* in the expected value, for any $\epsilon > 0$.

Another useful observation is that the Chernoff bound starts to "kick in" when $\epsilon \approx 1/\sqrt{\mathrm{E}[T]}$. In other words, the typical deviation of our random variable is going to be about the square root of its expectation. This is in line with our analysis of binomial random variables: the number of heads in n unbiased coin flips has expectation $n/2$, but has standard deviation $\sqrt{n}/2$, or roughly the square root of the expectation.

16.5 Proof of the Bound

The Chernoff Bound uses an ingenious trick along the lines of the way we derived Chebyshev's Bounds:

$$\Pr\{T \geq c\,\mathrm{E}[T]\} = \Pr\left\{c^T \geq c^{c\,\mathrm{E}[T]}\right\} \leq \frac{\mathrm{E}\left[c^T\right]}{c^{c\,\mathrm{E}[T]}} \tag{37}$$

The first step may be a shocker; we exponentiate both sides of the inequality in the probability by c. Since the new inequality describes the same event as the old, the probabilities are the same. The second step uses Markov's Theorem.

Recall that Markov's Theorem sometimes gives weak bounds and sometimes gives tight bounds. The motivation for the first step is to alter the distribution of the random variable T to hit the "sweet spot" of Markov's Theorem. That is, Markov's Theorem gives a tighter bound on the random variable c^T than on the original variable T. We used the same trick in Chebyshev's theorem: we looked at the expectation of T^2 instead of that of T, because that gave us more powerful results.

All that remains is to evaluate $\mathrm{E}\left[c^T\right]$. To do this we need a Lemma:

Lemma 16.4. *If R and S are independent random variables, and f and g are* any *real-valued functions on the reals, then $f(R)$ and $g(S)$ are independent random variables.*

We leave the proof of Lemma 16.4 as a routine exercise.

We begin by calculating $\mathrm{E}\left[c^{T_i}\right]$:

$$\begin{aligned}
\mathrm{E}\left[c^{T_i}\right] &::= c^1 \Pr\{T_i = 1\} + c^0 \Pr\{T_i = 0\} \\
&= c \Pr\{T_i = 1\} + (1 - \Pr\{T_i = 1\}) && \text{(complement rule, since } T_i = 0 \text{ iff } T_i \neq 1) \\
&= 1 + (c-1)\Pr\{T_i = 1\} \\
&\leq e^{(c-1)\Pr\{T_i=1\}} && (1 + x \leq e^x) \\
&= e^{(c-1)\,\mathrm{E}[T_i]} && \text{(expectation of indicator variable).}
\end{aligned} \tag{38}$$

So now we have

$$\begin{aligned}
\mathrm{E}\left[c^T\right] &= \mathrm{E}\left[c^{T_1+T_2+\cdots+T_n}\right] && \text{(def of } T) \\
&= \mathrm{E}\left[c^{T_1} \cdot c^{T_2} \cdots c^{T_n}\right] \\
&= \mathrm{E}\left[c^{T_1}\right] \cdot \mathrm{E}\left[c^{T_2}\right] \cdots \mathrm{E}\left[c^{T_n}\right] && \text{(independence of } T_i\text{'s and Lemma 16.4)} \\
&\leq e^{(c-1)\,\mathrm{E}[T_1]} \cdot e^{(c-1)\,\mathrm{E}[T_2]} \cdots e^{(c-1)\,\mathrm{E}[T_n]} && \text{(by (38))} \\
&= e^{(c-1)\,\mathrm{E}[T_1]+(c-1)\,\mathrm{E}[T_2]+\cdots+(c-1)\,\mathrm{E}[T_n]} \\
&= e^{(c-1)\,\mathrm{E}[T_1+\cdots+T_n]} && \text{(linearity of expectation)} \\
&= e^{(c-1)\,\mathrm{E}[T]} && \text{(def of } T).
\end{aligned} \tag{39}$$

Now we can substitute into the Markov inequality we started with to complete the proof of the

Chernoff bound (36):

$$\begin{aligned}
\Pr\{T \geq c\,\mathrm{E}[T]\} &\leq \frac{\mathrm{E}\left[c^T\right]}{c^{c\,\mathrm{E}[T]}} && \text{(by (37))}\\
&\leq \frac{e^{(c-1)\,\mathrm{E}[T]}}{c^{c\,\mathrm{E}[T]}} && \text{(by (39))}\\
&= \frac{e^{(c-1)\,\mathrm{E}[T]}}{(e^{\ln c})^{c\,\mathrm{E}[T]}}\\
&= e^{(c-1)\,\mathrm{E}[T]-c\ln c\,\mathrm{E}[T]}\\
&= e^{-(c\ln c-c+1)\,\mathrm{E}[T]}.
\end{aligned}$$

16.6 Example: A Phone Network Problem

Suppose that there is a phone network that handles a billion calls a day. Some of these call are routed through a certain switch. The exact number of calls passing through this switch is somewhat random and fluctuates over time, but on average the switch handles a million calls a day.

Our problem is to set the capacity of the switch; that is, we must determine the number of calls that a switch is able to handle in a day. If we make the capacity too small, then some phone calls will not go through. On the other hand, if we make the capacity too large, then we are wasting money. Of course, we cannot rule out a freak situation in which a huge number of calls are all coincidentally routed through the same switch, thus overloading it. However, we would like to guarantee that a switch is rarely overloaded.

Assume that each call has some probability of passing through a particular switch. In particular, let T_i be an indicator variable for the event that the ith call passes through the switch. That is, $T_i = 1$ if the call is routed through the switch, and $T_i = 0$ if the call does not pass through the switch. Then the total call load on the switch is $T = T_1 + T_2 + \cdots + T_n$. We do not know the exact probability that the switch handles the ith call, but we are given that the switch handles an average of a million calls a day; that is, $\mathrm{E}[T] = 1,000,000$.

We will make the crucial assumption that the random variables T_i are mutually independent; that is, calls do or do not pass through the switch mutually independently.

16.6.1 How to Build One Switch

We can now compute the probability that the load on a switch fluctuates upwards by 1% due to the randomness of calling patterns. Substituting $c = 1.01$ and $\mathrm{E}[T] = 1,000,000$ into the Chernoff Bound gives:

$$\begin{aligned}
\Pr\{\text{particular switch overloaded}\} &= \Pr\{T \geq 1.01 \cdot 1,000,000\}\\
&\leq e^{-(1.01\ln 1.01-1.01+1)\cdot 1,000,000}\\
&< e^{-1.01(0.00004934)\cdot 1,000,000}\\
&< 2.3\cdot 10^{-22}.
\end{aligned}$$

The probability that the load on the switch ever rises by even 1% is unbelievably small! (A June blizzard during an earthquake in Cambridge is far more likely.) If we build the switch with capacity only 1% above the average load, then the switch will (almost) never be overloaded.

The strength of this result relies on the huge expected number of calls. For example, suppose that the average number of calls through the switch were 100 per day instead of 1,000,000. Then every million in the above calculation would be replaced by a hundred; no other numbers change. The final probability of overloading the switch would then be bounded above not by $2.3 \cdot 10^{-22}$, but by 0.995! If the switch handles only 100 calls on an average day, then the call load can very often fluctuate upward by 1% to 101 or more.

16.6.2 How to Build the Network

We now know that building 1% excess capacity into a switch ensures that it is effectively never overloaded. The next problem is to guarantee that no switch in the entire network is overloaded. Suppose that are 1000 switches, and every switch handles an average of a million calls a day.

Previously, we saw that the probability that some event occurs is at most the sum of the event probabilities. In particular, the probability that some switch is overloaded is at most the sum of the probabilities that each of the 1000 switches is overloaded. Therefore, we have:

$$\Pr\{\text{some switch overloaded}\} \leq 1000 \cdot \Pr\{\text{particular switch overloaded}\} < 2.3 \cdot 10^{-19}.$$

This means that building 1% excess capacity into every switch is sufficient to ensure that no switch is ever overloaded.

The above results are of limited practical value, because calls typically do not pass through a switch mutually independently. For example, after an earthquake on Albany Street, everyone would call through the eastern Cambridge switchboard to check on friends and family. Furthermore, there are many more phone calls on certain dates like Mother's Day. On such occasions, 1% excess capacity is insufficient.

17 A Generalized Chernoff Bound

Chernoff's Theorem applies only to sums of Bernoulli (0-1-valued) random variables. It can, however, be extended to apply to sums of random variables with considerably more arbitrary distributions. We state one such generalization in Theorem 40 below. We omit its proof, noting only that the proof is similar to that of the Chernoff bound of Theorem 16.1. The bound of Theorem 40 is not quite as good as the Chernoff bound, but it is more general in what it covers.

Theorem 17.1. *Let $R_1, \ldots, R_n$ be independent random variables with $0 \leq R_i \leq 1$. Let $R = \sum_{i=1}^{n} R_i$. Then*

$$\Pr\left\{R - \mathrm{E}[R] \geq c\sqrt{n}\right\} \leq e^{-c^2/2}. \tag{40}$$

Example 17.2. Load balancing.

A set of n jobs have to be scheduled on a set of m equally fast processing machines. The length (processing time) of the ith job is some number L_i in the range $[0, 1]$. We would like to schedule the jobs on the machines so that the loadtime on the machines is reasonably balanced. This means we would like the loadtime on every machine to be not much more than the average loadtime

$L ::= \sum_{i=1}^{n} L_i/m$ per machine. Finding an optimally balanced assignment is a notoriously time-consuming task even when we know all the processing times. But commonly, successive jobs have to be assigned to machines without knowing how long the later jobs will take, and in that case it is impossible to guarantee a balanced load.

We will approach this problem of load balancing using the simplest random strategy: we independently assign each job to a randomly selected machine, with each machine equally likely to be selected. It turns out that for many job scheduling problems, this strategy is almost certain to do very well, even though it does not even take into account the number of jobs nor their processing times.

To see why the simple random strategy does well, notice that, since each job has probability $1/m$ of being assigned to any given machine, the expected loadtime on a machine is precisely $1/m$th of the total time required by the jobs. That is, the *expected* loadtime per machine is precisely the *average* loadtime, L. Now consider any machine, M, and define R_i to be the time M spends processing the ith job. That is, $R_i = L_i$ if the ith job gets assigned to machine M, otherwise $R_i = 0$. So the total loadtime on machine M is $\sum_{i=1}^{n} R_i$. From the generalized Chernoff bound (40), we conclude

$$\Pr\left\{(\text{loadtime on machine } M) - L \geq c\sqrt{n}\right\} \leq e^{-c^2/2}.$$

Now by Boole's inequality,

$$\begin{aligned} &\Pr\left\{\text{the loadtime on some machine is } \geq L + c\sqrt{n}\right\} \\ &\leq \sum_{k=1}^{m} \Pr\left\{(\text{loadtime on machine } k) \geq L + c\sqrt{n}\right\} \\ &\leq me^{-c^2/2}. \end{aligned}$$

If we choose $c = \sqrt{2\ln m} + 6$, say, so that $c^2/2 \geq \ln m + 18$, then

$$\begin{aligned} &\Pr\left\{(\text{loadtime on each machine}) \leq L + c\sqrt{n}\right\} \\ &\geq 1 - me^{-\ln m - 18} \\ &= 1 - e^{\ln m}e^{-\ln m - 18} \\ &= 1 - e^{-18} = 0.99999998\ldots. \end{aligned}$$

Hence, we can be 99.999998% sure that every machine will have load at most $L + (\sqrt{2\ln m} + 6)\sqrt{n}$. For many values of n and m this is comes very close to balancing the loads on all the machines. For example, if $m = 10$ machines, $n = 5000$ jobs, and average load length $L = 300$ per machine, the maximum load on any machine will almost surely not exceed 332. (In fact it is likely to be even less—we have been fairly crude in our bounds.)

18 Review of Markov, Chebyshev, Chernoff and Binomial bounds

Let us review the methods we have for bounding deviation from the mean via the following example. Assume that I.Q. is made up of thinkatons; each thinkaton fires independently with a 10% chance. We have 1000 thinkatons in all, and I.Q. is the number of thinkatons that fire. What is the probability of having Marilyn's IQ of 228?

So the I.Q. is a Binomial distribution with $n = 1000, p = 0.1$. Hence, $\mathrm{E}[\text{I.Q.}] = 100, \sigma_{\text{I.Q.}} = \sqrt{0.09 \times 1000} = 9.48$.

An I.Q. of 228 is $128/9.48 > 13.5$ standard deviations away.

Let us compare the methods we have for bounding the probability of this I.Q..

1. Markov:
$$\Pr\{\text{I.Q.} \geq 228\} \leq \frac{100}{228} < 0.44$$
2. Chebyshev:
$$\Pr\{\text{I.Q.} - 100 \geq 128\} \leq \frac{1}{13.5^2 + 1} < \frac{1}{183}$$
3. Chernoff:
$$\Pr\{\text{I.Q.} \geq 2.28 \times 100\} \leq e^{-(2.28 \ln 2.28 - 2.28 + 1)100} \leq e^{-59.9}$$
4. Binomial tails:
$$\Pr\{\text{I.Q.} \geq 228\} = \Pr\{1000 - \text{I.Q.} \leq 772\} = F_{0.9,1000}(772) \leq e^{-72.5}$$
   Here we used the formula for the binomial tail from Notes 11-12 with $p = 0.9, n = 1000, \alpha = 0.772$.

Note that the more we know about the distribution the better are the bounds we can obtain on the probability.