Deviations from the Mean

Final exam: Cumulative, coverage of everything we've done, emphasis on latter part of course (probability, structures).

1 Variance and Chebyshev's Theorem

In general, just knowing the range and expected value of a random variable is not enough to give a really good idea of what the distribution looks like. In an effort to get a better handle on the shape and nature of a distribution, mathematicians have defined several other properties of a distribution which help describe it. After the expected value, the next most important measure of a random variable is its variance.

Sometimes Markov is too weak. What to do?

Definition 1.1 The variance of R is

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

The variance gives us the average of the squares of the deviations from the mean. The idea behind variance is that it is large when R is likely to be far from its expectation, that is, when its distribution is "spread out". If R is usually close to its expectation, the variance will be small.

Example 1 Consider the follwing two gambling games, A and B:

- Game A: With probability 2/3, you win 2$. Otherwise, you lose 1$
- Game B: With probability 2/3, you win 1002$. Otherwise, you lose 2001$

Which would you play? First, note that in either game, you win with probability 2/3. Let A, B be the returns from playing games A and B, respectively. Then for game A, $E[A] = 2 \cdot 2/3 - 1 \cdot 1/3 = 1$, and for game B, $E[B] = 1002 \cdot 2/3 - 2001 \cdot 1/3 = 1$. So the expected return is the same for both games, namely 1! Still, the games are very different. This difference is captures by the variances. Let's see what the variance is for each game:

- Game A:

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

- Game B:

$$
\begin{aligned}
B - E[B] &= \begin{cases} 1001 & \text{with probability } 2/3, \\ -2002 & \text{with prob. } 1/3 \end{cases} \\
(B - E[B])^2 &= \begin{cases} 1002001 & \text{with probability } 2/3 \\ 4008004 & \text{with prob. } 1/3 \end{cases} \\
\text{Var}[B] &= 1002001 \cdot 2/3) + 4008004 \cdot 1/3 \\
&= 2004002
\end{aligned}
$$

Hence we se that game B is much more risky than game A. Risk averse people should stay away from game B!!! In short, playing a game with high variance is a serious risk: we can win a lot, but we can loose a lot, too. Playing another game with the same expectation of winnings but less variance we can hope for the same average gains risking less. In investments, many people prefer bonds to stocks because, although the expected earnings in bonds are lower, the variance is much lower too. They are willing to pay for risk-reduction.

Why do we bother to square the deviation before taking expectations? Why not just compute the expected deviation from the mean? Let's consider a random variable R with mean $E[R]$. What we are interested in, then, is the deviation of the random variable from its mean, namely $R - E[R]$. Unfortunately, $E[R - E[R]] = E[R] - E[R] = 0$, so the positive and negative deviations cancel out and give us no information. This problem occurs because the expected value is the center of mass for the distribution, so the positive and negative deviations *must* balance out, by definition of expectation. So, one advantage of squaring is that all deviations become positive, and hence there is no cancelling. Of course, we could also get around this cancellation problem by computing $E[|R - E[R]|]$. However, absolute values are painful in analysis (not differentiable, etc.). We could also take higher powers of R-E[R]. If we take the 4th power, for instance, we get what is called the *kurtosis* of R. It turns out however, for reasons what we will not go into, that squaring is pretty much the best way to go.

People often like to "renormalize" the variance:

Definition 1.2 The *standard deviation* of R is

$$\sigma_R = \sqrt{\mathrm{Var}[R]}.$$

Thus the variance is the square of the standard deviation. It follows that the standard deviation has the same units (days, people, etc.) as the original random variable and its expectation. On an *intuitive* level, the standard deviation measures $E[|R - E[R]|]$ (careful—they aren't really equal!). The standard deviation is a very popular measure in statistics, particularly when fitting curves to experimental data.

Example 2 Returning to game B, we can compute its standard deviation σ_B:

$$\sigma_B = \sqrt{\mathrm{Var}[R]} = \sqrt{2004002} \approx 1416$$

So far, we have given a precise definition of variances of random variables, as well as their standard deviations. The intuition is that these measures tell you how likely a random variable is to be far from its mean. Unfortunately, this is not always true. Just like the expectation of a random variable does not completely specify the distribution, the expectation of the deviations does not completely specify the distribution of deviations. However, the variance and standard deviation of a distribution are still useful tools for analysis.

1.1 The Chebyshev Bound

A variable with small variance is unlikely to be far from its mean. With Markov's Theorem, we can deduce, for example, that given a mean IQ of 100, only 1/2 the people can have IQ bigger than 200. But suppose we know that the standard deviation in IQ is 10. Then intuitively, it should quite rare to have an IQ bigger than 110. We can show that in fact only 1/100 people have IQ bigger than 200.

Chebyshev's theorem formalizes this observation:

Theorem 1.3 For any random variable R,

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

Proof.

$$\begin{aligned}
\Pr(|R - E[R]| \geq x) &= \Pr((R - E[R])^2 \geq x^2 \\
&\leq \frac{E[(R - E[R]^2]}{x^2} \\
&= \frac{\mathrm{Var}[R]}{x^2}
\end{aligned}$$

■

Corollary 1.4 $\Pr(|R - E[R]| > c \cdot \sigma_R) \leq 1/c^2$.

Proof. Set $x = c \cdot \sigma_R$ ■

Chebyshev's theorem is very similar to Markov's Theorem, but uses the added information supplied by the standard deviation to get a better upper bound. In fact, the proof uses Markov's Theorem.

Example 3 Let R be the IQ of a random person. Then, assume $R \geq 0$, $\sigma_R = 10$ and $E[R] = 100$. Therefore, using Chebyshev:

$$\begin{aligned}\Pr(R \geq 200(&= \Pr((R - 100) \geq 100) \\ &= \Pr(|R - E[R]| \geq 10\sigma_R) \\ &\leq 1/100\end{aligned}$$

Had we used Markov's Theorem here, we would have gotten a lower bound of 1/2 on this same probability.

In the previous example, we used $\Pr(|R - E[R]|)$ to give an upper bound on $\Pr(R - E[R])$. Maybe the probability $\Pr(R - E[R])$ is in fact half the probability $\Pr(|R - E[R]|)$? In this case, the answer is no. If the distribution were symmetric about the mean, however, it would be true. We do not, however, have any reason to believe that the distribution is symmetric. In general, if we only want the probability mass of one of the tails, we can do a little better than what we just did, however.

Theorem 1.5 For any random variable R, the following holds:

$$\Pr(R - E[R] \geq x) \leq \frac{\mathrm{Var}[R]}{\mathrm{Var}[R] + x^2}$$

$$\Pr(R - E[R] \leq -x) \leq \frac{\mathrm{Var}[R]}{\mathrm{Var}[R] + x^2}$$

We will not prove this theorem. The following is a useful corollary:

Corollary 1.6 For any random variable R, the following holds:

$$\Pr(R - E[R] \geq c\sigma_R) \leq \frac{\sigma_R^2}{\sigma_R^2 + c^2\sigma_R^2} = \frac{1}{1+c^2}$$

$$\Pr(R - E[R] \leq -c\sigma_R) \leq \frac{\sigma_R^2}{\sigma_R^2 + c^2\sigma_R^2} = \frac{1}{1+c^2}$$

Example 4 Returning to our IQ example, we can now recompute the probability of having an IQ greater than 200;

$$\Pr[R \geq 200] \leq 1/101$$

Example 5 At most 1/4 of the students score more than 2 standard deviations away from the mean, whatever the set of scores is. Indeed, if R is the score of a random student, then

$$|Pr(R - E(R)| \geq 2\sigma_R) \leq \frac{1}{4},$$

which means that at most 1/4 of all the scores can be that far from the average. Moreover, at most 1/5 of the students score more than 2 standard deviations above the mean, whatever the set of scores is. Indeed, if R is the score of a random student, then

$$Pr(R - E(R) \geq 2\sigma_R) \leq \frac{1}{1+2^2} = 1/5,$$

Note that the results we get from Markov's Theorem and Chebyshev's Theorem are only upper bounds on the probability of deviating from the mean. If there is more information available about the distribution, it might be possible to prove better bounds. For example, if we know we are dealing with an unbiased binomial distribution, then it is possible to prove a much better upper bound:

Theorem 1.7 For an unbiased binomially distributed random variable R,

$$\Pr[R - E[R] \geq c\sigma_R] < e^{-c^2}$$

Example 6 Suppose that the test scores are binomially distributed. Then the probability of a randomly selected student having a score greater than or equal to two times the standard deviation is bounded above by $1/e^4$.

1.2 A Different Variance Formula

There is another popular way of defining the variance of a random variable, which we will state as a theorem.

Theorem 1.8 For all R, we have

$$Var(R) = E[R^2] - E[R]^2.$$

Proof.

$$\begin{aligned} Var(R) &= E[(R - E[R])^2] \\ &= E[R^2 - 2RE[R] + E[R]^2] \\ &= E[R^2] - 2E[R]E[R] + E[R]^2 \\ &= E[R^2] - E[R]^2. \end{aligned}$$

■

This formula is often more useful to compute variances than the previous one.

Example 7 Let R be the outcome of rolling a 6-sided die. Then the variance of R is given by

$$\begin{aligned} \mathrm{Var}[R] &= \frac{1}{6}(1 + 4 + 9 + \cdots + 36) - (3.5)^2 \\ &= 15\frac{1}{6} - 12\frac{1}{4} = 2\frac{11}{12} \end{aligned}$$

This theorem is also useful from a computational point of view. Suppose, say, that you are given a stream of data (like test scores), and you are trying to compute the expectation and variance on the fly with only one pass through the data. Computing a running average is easy, we just compute a running sum and divide by the number of data points seen so far. Similarly, to compute the variance, we compute two running sums, one for the scores and one for the square of scores.

The theorem has a nice corollary:

Corollary 1.9 For all R, $E(R^2) \geq E(R)^2$.

Proof. This follows from the previous theorem, since Var(R) is always nonnegative, being the expectation of a square. ■

1.3 Manipulating Variances

Theorem 1.10 If Var(R)=0, then R is a constant scalar.

Proof. Suppose that

$$\mathrm{Var}[R] = E(R - E[R])^2] = 0.$$

Then $(R - E[R])^2$ must be always 0, thus R must always be equal to E[R]. ∎

Theorem 1.11 For any constants a and b, $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$.

Proof.

$$\begin{aligned}
\mathrm{Var}[aX + b] &= E[(aX + b)^2] - (E[aX + b])^2 \\
&= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\
&= a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 \\
&= a^2(E[X^2] - E[X]^2) \\
&= a^2\mathrm{Var}[X]
\end{aligned}$$

∎

A Bernoulli Random variable with probability p has variance

$$p \cdot 1^2 + (1 - p) \cdot 0^2 - p^2 = p(1 - p)$$

What about the binomial distribution? This is messy to do directly, but easy if we figure out how to sum variances:

Theorem 1.12 If X and Y are *independent*, then

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Proof. The assumption that X and Y are independent means that $E[XY] = E[X]E[Y]$. Therefore,

$$\begin{aligned}
\mathrm{Var}[X + Y] &= E[(X + Y)^2] - E[X + Y]^2 \\
&= E[X^2 + 2XY + Y^2] - (\mu_X + \mu_Y)^2 \\
&= E[X^2] + 2E[XY] + E[Y^2] - \mu_X^2 - 2\mu_X\mu_Y - 2\mu_Y^2 \\
&= E[X^2] - \mu_X^2 + E[Y^2] - \mu_Y^2 + 2E[XY] - 2\mu_X\mu_Y \\
&= E[X^2] - \mu_X^2 + E[Y^2] - \mu_Y^2
\end{aligned}$$

(since $E[XY] = \mu_X \mu_Y$ by independence)

$$= \text{Var}[X] + \text{Var}[Y]$$

■

Corollary 1.13 If $X_1, \ldots, X_n$ are *pairwise* independent random variables, then

$$\text{Var}[\sum X_i] = \sum \text{Var}[X_i].$$

Note that mutual independence is *not* required, because the only fact we used that relied on independence was that $E[X_i X_j] = E[X_i]E[X_j]$. This follows from pairwise independence, since it means X_i and X_j are independent.

Therefore, the variance of the binomial distribution on n Bernoulli trials with probability p is $np(1-p)$. It follows that the standard deviation is $\sigma = \sqrt{np(1-p)}$.

2 Exponential Bounds for Independent Events

Markov and Chebyshev bounds apply to a large class of problems, but this means they are rather weak. By restricting attention to the important case where all the events are *independent*, we can prove much stronger bounds.

We'll solve a particular kind of probability problem that arises in all sorts of applications. In the general form of the problem, we have a collection of independent events and we want to get some idea of how many of the events are likely to occur.

Given: mutually independent events $A_1, A_2, \cdots, A_n \subseteq S$.

For example, we can think of A_i as being the event that the ith component of a system fails. Or we can think of A_i as the event that the ith coin is heads when flipping n coins. r anything else–probability student answers the ith question on the final correctly.

What we want to know is, what is the probability that k of the events occur. Let T be the number of events that occur.

The question is easy to answer if each event is equally likely, that is, $\forall i Pr(A_i) = p$. Then $Pr(T = k) = \binom{n}{k} p^k (1-p)^{n-k}$, the binomial distribution. We've already talked about important distribution. Very tightly centered around the mean $= pn$.

Today, we're going to talk about the general case: $Pr(A_i)$ are different.

First thing is to look at the expected number of events to occur. To figure this out, we can use linearity of expectation and indicator variables.

Let T_i be an indicator for the ith event. (1 if A_i occurs, 0 otherwise)

$$\begin{aligned} T &= T_1 + T_2 + \cdots + T_n \\ E[T] &= E[T_1] + E[T_2] + \cdots + E[T_n] \\ &= Pr(A_1) + Pr(A_2) + \cdots + Pr(A_n) \\ &= \sum_{i=1}^{n} Pr(A_i) \end{aligned}$$

So to compute the expected number of events, we just add up the probabilities. We don't even need the A_i independent for this, because we just used linearity of expectation.

Example: A_i is the probability the ith component fails. Then the expected number of failures is the sum of the individual component failures, even if there is dependence.

OK, nothing really new here. But say that we want to know the probability that there is one or more failures. That is, say that any fault is critical, so we don't care how many there are, just if there is one. (Once there is one it is all over, so who cares how many more happen?) So we want $Pr(T \geq 1)$.

First let's get a simple upper bound. By Markov,

$$Pr(T \geq 1) \leq \frac{E[T]}{1} = E[T] = \sum_{i=1}^{n} Pr(A_i)$$

So if the sum of the individual event probabilities is small, then the chance that one or more occurs is also small. Note we haven't used independence yet.

We could show this another way, using Boole's inequality. Let A be the event that 1 or more occurs. So $A = A_1 \cup A_2 \cup \cdots \cup A_n$.

$$Pr(T \geq 1) = Pr(A) \leq \sum_{i=1}^{n} Pr(A_i) = E[T]$$

Note we still didn't need independence.

So if the expected number of events to occur is $<< 1$, then the probability at least one occurs is very small. For example, if you buy 1000 lottery tickets, each of which wins with probability 1 in a million, then the probability you have a winner is $\leq 1000/1000000 = 1/1000$.

Now let's look at the case where the expected number of events to occur is bigger than 1. For example if the expected number is 10, it seems likely that at least one will happen. This is the case if the events are independent. The following theorem captures this fact:

Theorem 2.1 Given independent events $A_1, A_2, \cdots, A_n$, the probability that none occur is $\leq e^{-m}$, where $m = \sum_{i=1}^{n} Pr(A_i)$ = expected number to occur.

So probability none occur is exponentially small in the expected number of events to occur.

Proof. Probability that no event occurs

$$
\begin{aligned}
&= Pr(\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}) \\
&= \prod_{i=1}^{n} Pr(\overline{A_i}) \quad \text{(by independence)} \\
&= \prod_{i=1}^{n} (1 - Pr(A_i)) \\
&\le \prod_{i=1}^{n} e^{-Pr(A_i)} \\
&= e^{-\sum_{i=1}^{n} Pr(A_i)} \\
&= e^{-m}
\end{aligned}
$$

∎

Note independence is critical here. Also, this is another example where Taylor series bounds are important.

This is a really powerful result. For example, we immediately get the following corollary:

Corollary 2.2 Given independence, if we expect 10 or more events to occur, the probability that no event occurs is $\le e^{-10} < 1/22000$

Note there is no dependence on n here. This is a really strong result since it says that if we expect some things to happen then they almost certainly will. I.e, if you expect your system to have some faults, then it surely will–this is sort of a proof of Murphy's Law. If there are a bunch of things that can go wrong, even if none of them is very likely, this result says that almost surely something will go wrong.

Example 8 (Murphy's Law) n components, each fails with probability $p = 10/n$. expected number of failures = 10, probability there is no faults $\le e^{-10}$.

So even though the failure probability of any component is small, and only 10 are expected, probability of none is not $1/10$ but $\le e^{-10}$.

This theorem is in fact so powerful that it will allow me to read your mind.

3 Mind Reading Game

You do the following: Pick a number i from 1 to 9. You then start dealing out a deck of cards one at a time. When you get to card i, you continue counting ahead, this time with the value of that card. 10, J, Q, K, A, and Joker all count as 1. Whatever card you are counting off of when the deck runs out is your card to be mind read.

Example (on a few cards): You pick 3. The sequence of cards in the deck is 4 5 2 3 10 Q 3 7 6 4 2. So the cards you use are 2, 10, Q, 3, 4, and the card to be mind read is the 4.

I get to read your mind because if I do the same thing as you it is likely we will have the same last card. The reason, is that is we ever land on the same card we will have the same last card. Each time one of use hits a card and starts counting again, there is some chance we will land on the others positions. If 1 away, 5/13 chance, if >1 away, 1/13 chance, and there are lots of times that we each jump. On average, we will each jump >10 times, so 20 about jumps total between the two sides, making it likely that we will collide. In particular, A_i is the event that on the ith jump we collide.

Now the intuition I just went thru was pretty fuzzy. In fact, the theorem doesn't really apply in this example because the events are not independent. If we have not yet collided, it tells us something about the numbers we have had, and since we only have a 52 card deck, something about the numbers that are forthcoming, which conditions the probability that we collide. However, we could analyze it if the cards we selected at random from an infinite number of decks. Then the event of a hit would be independent for every jump.

We don't have time in class to analyze this precisely. It turns out that the probability I read you mind is $> 90\%$ for a 52 card deck.

The point of all this can be summed up as follows:

If we expect several events to occur, then the probability that no events occur is exponentially small if the events are independent. If the events are dependent, then the probability of no events occurring could be anything.

4 Chernoff Bound

Now we are going to prove an even stronger result. If we have a collection of independent events and we expect m of them to occur, then with probability very close to 1, close to m will occur. It's similar to the phenomenon with tossing n coins. We expect $n/2$ heads, and with probability close to 1 we will get close to $n/2$ heads.

For simplicity, we will only look at the case of getting much more than the expected value; a similar method gives similar results for the other side.

Theorem 4.1 (Chernoff Bound) Let $T = T_1 + T_2 + \cdots + T_n$, where the T_i are independent Bernoulli random variables. Then $\forall c > 1$ $Pr(T \geq cE[T]) \leq e^{-\alpha c E[T]}$, where $\alpha = \ln c + 1/c - 1 \geq 0$.

In general, this is much stronger than what Markov's theorem gives (1/c); here we get exponential bounds on probability.

Corollary 4.2 With T as above, $Pr(T \geq eE[T]) \leq e^{-E[T]}$.

Proof. $c = e \Rightarrow \alpha = \ln e + 1/e - 1 = 1/e \Rightarrow \alpha c = 1$ ■

Example 9 If $E[T] = 100$, $Pr(T \geq 272) \leq e^{-100}$.

Example 10 1 million people play "Pick 4". Each player picks a random 4 digit decimal number. So $Pr(winning) = 1/10000$, $E[\#winners] = 1M/1000 = 100$. We assume random picks for the sake of argument. If picks are independent, then $Pr(\geq 272 \text{ winners}) \leq e^{-100}$, which is really, really, tiny. Note answer is not dependent on the number of people playing or the individual probability of winning.

Of course, the result is not true if the picks are dependent. If everyone picks the same number, then the $Pr(\geq 272 \text{ winners}) = 1/10000 =$ chance that number comes up. (In effect, we've proved that it is incredibly unlikely that 272 pick the same number if they choose independently.)

Now in truth with lottery, people tend to pick in a structured fashion. (Digression) They pick the date, or something they saw on TV, or yesterdays number, etc. In fact, Chernoff himself (who was once on the faculty here) developed a strategy to make money on the state lottery. He did a statistical study of unpopular numbers and showed that by betting on them you would expect to earn money! This only works in lotteries where payoff is fraction of total pot. If zillions of people all bet on few numbers and you bet on all the rest, you are likely to win their zillions of dollars. Way it turned out made 7 percent return. Now state sets it up to give you random number.

Back to the proof of Chernoff's theorem:

Proof. This uses an exponential transform of Markov's theorem.

$$Pr(T \geq cE[T]) = Pr(c^T \geq c^{cE[T]})$$

This is because it is the same event, just stuffed in an exponent on one side. Then by Markov,

$$\leq \frac{E[c^T]}{c^{cE[T]}}$$

Need to evaluate $E[c^T]$:

$$\begin{aligned}
T &= T_1 + T_2 + \cdots + T_n \\
\Rightarrow c^T &= c^{T_1} c^{T_2} \cdots c^{T_n} \\
\Rightarrow E[c^T] &= E[c^{T_1}]E[c^{T_2}] \cdots E[c^{T_n}] \quad \text{(by product rule (}T_i \textit{ ind.} \Rightarrow c^{T_i} \textit{ ind.))} \\
&= \prod_{i=1}^{n} E[c^{T_i}]
\end{aligned}$$

So just need to evaluate $E[c^{T_i}]$:

$$\begin{aligned}
E[c^{T_i}] &= c^1 Pr(T_i = 1) + c^0 Pr(T_i = 0) \\
&= cPr(T_i = 1) + (1 - Pr(T_i = 1)) \\
&= 1 + (c-1)Pr(T_i = 1) \\
&\le e^{(c-1)Pr(T_i=1)} \quad \text{(since } 1 + x \le e^x\text{)} \\
&= e^{(c-1)E[T_i]} \quad \text{(since indicator variables)}
\end{aligned}$$

Now we can plug in:

$$\begin{aligned}
E[c^T] &\le \prod_{i=1}^{n} e^{(c-1)E[T_i]} \\
&= e^{\sum_{i=1}^{n}(c-1)E[T_i]} \\
&= e^{(c-1)\sum_{i=1}^{n} E[T_i]} \\
&= e^{(c-1)E[T]}
\end{aligned}$$

Now we can go back and plug into the bounds from Markov:

$$\begin{aligned}
Pr(T \ge cE[T]) &\le \frac{e^{(c-1)E[T]}}{c^{cE[T]}} \\
&= e^{(c-1)E[T] - c\ln cE[T]} \\
&= e^{-c(\ln c + 1/c - 1)E[T]} \\
&= e^{-\alpha c E[T]} \;\; \textit{where } \alpha = \ln c + 1/c - 1
\end{aligned}$$

■

This is the deepest result in the class. Don't feel like you'd have to come up with this on your own–it's pretty tricky. But the basic method is not so bad:

$$Pr(A > B) = Pr(f(A) > f(B)) \le \frac{E[F(A)]}{f(B)}$$

and then just plug and chug.

You should try to remember the implications of the result. If you have a sum of 0-1 variables or a collection of independent events, then almost surely the number of events is close to the expected number. "Tails" are exponentially small.

Example 11 Say we have a phone network (or some kind of internet), and the network typically handles 1 billion calls (or messages) per day. In this problem, we are interested in the volume of traffic that is routed thru a particular switch. In particular, we are interested in the fluctuations in volume - want to buy big enough switch to handle random variations in traffic. Each particular call may or may not be routed thru the switch, depending on its destination and source. We will treat each call as a random variable T_i:

$$T_i = \begin{cases} 1 \text{ if call } i \text{ gets routed thru switch,} \\ 0 \text{ otherwise.} \end{cases}$$

Then the total number of calls routed through the switch is given by:

$$T = \sum_{i=1}^{10^9} T_i$$

Let $P_i = \Pr(T_i = 1)$. P_i is an unknown - we don't know the probability that each of the billion calls goes thru switch. We do know the expected number of calls going thru the switch, however - it is about 1 million. So assume that $E[T] = 10^6$. Aslo, assume that the T_i are mutually independent.

We want to figure out what sort of variations T might have. In particular, we'd like to know if a 1% variation from the mean is likely to occur (on the high side). So let's compute the probability of having more than 1.01 million calls going thru the switch. We use Chernoff's bound with $c = 1.01$ and $E[T] = 10^6$, and get:

$$\begin{aligned} \Pr(T \geq 1.01 \cdot 10^6) &\leq e^{\alpha c E[T]} \\ &\leq e^{-49.8} \\ &\leq 2.3 \cdot 10^{-22} \end{aligned}$$

So the chance of being more than 1% above the mean is not really all that high!! This tells us that if we build in an extra 1% capacity, we will, for all practical purposes, never exceed the switch's capacity. Notice that, if we only expected 100 calls, the probability of getting more than 101 calls would be pretty high. Things work in our example because of the volume being so high (this means that the exponent in Chernoff's bound becomes very small).

So why do we sometimes see abnormally high loads on telephone switches in practice? The answer is that phone calls are not necessarily independent. Some events, like mother's day, earth quakes, etc. etc. change calling patterns significantly.

One last point. Suppose that we have 1000 switches in our network. What is the probability that no switch has more traffic than 1% above the mean, supposing al the switches have an expected traffic of 10^6 calls per day? Let $\mathcal{O}_s$ be the event that some switch s gets more traffic than $1.01 \cdot 10^6$ on a given day. From Boole's law we have that:

$$
\begin{aligned}
\Pr(\bigcup_s \mathcal{O}_s) &\leq 1000 \cdot \Pr(\mathcal{O}_f) \\
&\leq 2.3 \cdot 10^{-19}
\end{aligned}
$$

5 Summary

To summarize our work on deviations from the mean:

Markov's Inequality applies to any random variable with finite expectation, and gives bounds *linear* in the deviation. It applies to sums of random variables without restriction.

Chebyshev's Inequality applies to any random variable with finite variance, and gives bounds *quadratic* in the deviation. It applies to sums of random variables that are *pairwise independent.*

Chernoff Bounds apply to sums of *mutually independent* random variables and give bounds that are *exponential* in the deviation.