Deviations: small and large

Often, our goal is to understand how collections of a large number of independent random variables behave. This is what the laws of large numbers reveal. In general, the idea is that the average of a large number of i.i.d. random variables will approach the expectation. Sometimes that is enough, but usually, we also need to understand how fast does the average converge to the expectation. We saw this already in the context of polling — we needed to understand how many people to poll in order to get a trustworthy high-precision estimate of what the population is like.

So far, our main tools have been the Markov and Chebyshev inequalities:

$$\text{(Markov)} \quad \text{If } X \geq 0, \text{ then} \quad \Pr(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$$

$$\text{(Chebyshev)} \quad \text{If } X_i \text{ are i.i.d., then} \quad \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}[X_1]\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$$

Taking the limit as n goes to infinity, Chebyshev's inequality implies that if the X_i are i.i.d, then

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}[X_1]\right| \geq \varepsilon\right) \to_{n\to\infty} 0$$

so the average of the X_i converges to the expectation $\mathbf{E}[X_1]$. This is called the *weak law of large numbers*.

The role of precision is played by the tolerance ε above. The measure of trustworthiness is how small the probability of exceeding the tolerance gets. For the sake of simplicity, let us use the reciprocal of the probability of being outside of our desired precision as our measure of confidence. Looking at Chebyshev's inequality, we are tempted to draw two conclusions about what happens as n gets large:

- If we fix the desired confidence (pick a suitably low probability of our estimate being incorrect), then the precision ε seems to be improving like $\frac{\sqrt{\mathrm{Var}(X)}}{\sqrt{n}}$.
- If we fix the precision ε, then the confidence improves linearly in n with a slope that depends on the precision.

It turns out that the first of these is essentially correct, while the second is often overly pessimistic. The first is covered by the central limit theorem and is often referred to as the study of "small deviations." Later in this lecture, we will see that actually, the confidence improves exponentially in n. That is explored using what are called Chernoff Bounds and the general area is referred to as the study of "large deviations." The difference between "small" and "large" is that the small deviations are shrinking with n.

The Central Limit Theorem: studying "small" deviations

The Central Limit Theorem has a long and illustrious history. At an intuitive level, it says that the *appropriately-scaled*[1] sum of a bunch of independent random variables behaves like a Gaussian (AKA Normal) random variable.

Theorem 19.1: (Central Limit Theorem) Let X_i be i.i.d. random variables with $\mathbf{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, and define

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n \cdot \sigma^2}}.$$

Then for for all z we have

$$\text{(CLT)} \qquad \lim_{n\to\infty} \Pr(Z_n \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x.$$

We say that $Z_n \to N(0,1)$: Z_n converges in probability towards the standard Gaussian with mean 0 and variance 1.

Notice that the convergence above is in exactly the same sense (the CDF of Z_n converges) as the convergence of the appropriately scaled geometric random variables to a continuous-valued exponential random variable. The Berry-Esséen inequality gives a more precise quantification of the speed of this convergence:

$$\text{(Berry-Esséen)} \qquad \left| \Pr(Z_n \leq z) - \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x \right| \leq \frac{0.77\,\mathbf{E}[|X_1 - \mathbf{E}[X_1]|^3]}{(\text{Var}(X_1))^{3/2}\sqrt{n}}.$$

The convergence can be much faster than that[2], but if all we know is that there exists a third moment $\mathbf{E}[|X_1 - \mathbf{E}[X_1]|^3]$, then that's pretty much the best that we can have. As a rule of thumb, you should be wary about using the CLT in the context of probabilities that are smaller than $O(1/\sqrt{n})$.

The CLT construction of Z_n is done specifically so that its mean is always zero and variance is always 1. The fact that the CLT works validates the $\sqrt{\frac{\text{Var}(X)}{n}}$ scaling of precision that even Chebyshev's coarser inequality predicted.

Example 1. (From Bertsekas and Tsitsiklis) We have 100 bags, with weight distributed uniformly in $[5,50]$, so that the mean weight is 27.5lbs and the variance is 168.75lbs. Can we upper bound the probability that the total weight is larger than 3000lbs?

For comparison, let's see what happens if we use Chebyshev's inequality,

$$\Pr\left(\left| \frac{1}{100} \sum_{i=1}^{100} X_i - \frac{2750}{100} \right| \geq 2.50 \right) \leq \frac{168.75}{100 \cdot (2.5)^2} \approx 0.27.$$

That is a strict inequality, but intuitively it is overestimating the probability by a factor of two since it is also including the case of the average being significantly smaller than the mean as well.

[1] Another way of interpreting the issue of appropriately scaling is to say that the sum of a bunch of appropriately small independent random variables behaves like a Gaussian. This is used to justify Gaussian distributions for thermal noise that results from the combined random motion of many molecules or electrons.

[2] You are strongly encouraged to use a computer to plot for yourself how fast the CDF converges for the following example: X_i uniform. The case of Bernoulli random variables is illustrated at the end of this note. Both of these converge quite rapidly.

If we use the CLT instead,

$$\Pr\left(\sum_{i=1}^{100} X_i \geq 3000\right) = 1 - \Pr\left(\sum_{i=1}^{100} X_i < 3000\right)$$

$$= 1 - \Pr\left(\frac{\sum_{i=1}^{100} X_i - 2750}{\sqrt{100 \cdot 168.75}} < \frac{250}{\sqrt{100 \cdot 168.75}}\right)$$

$$\approx 1 - \Pr\left(Z < \frac{250}{\sqrt{100 \cdot 168.75}}\right) \approx 0.027.$$

where Z is a random variable distributed as $N(0,1)$. So the CLT predicts that the probability of having a total weight of over 3000lbs is about 3%, which is much lower than the 27% bound (or 13% if we divide by two) that we got from Chebyshev's inequality!

It turns out that the Normal approximation here is very good even though the CLT is not a bound and 3% is not an appropriately low number according to the Berry-Esséen inequality. This is because the CLT is very good for the sums of bounded random variables like the uniform.

Remark. In order to use the CLT to get easily calculated bounds, the following approximations will often prove useful: for any $z > 0$,

$$\left(1 - \frac{1}{z^2}\right) \frac{e^{-z^2/2}}{z\sqrt{2\pi}} \leq \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x \leq \frac{e^{-z^2/2}}{z\sqrt{2\pi}}.$$

This way, you can approximate the tail of a Gaussian even if you don't have a calculator capable of doing numeric integration handy. It also reveals how the tail scales in a "closed form" way. We will use this later.

Example 2. Imagine a polling situation, in which we assume that people have a probability p of supporting Obama. Suppose we poll n people, and we want the result to be within $\pm 1\%$ of the true value p with probability at least 95%. Let $M_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the average vote of the polled persons. We want to choose n big enough so that $\Pr(|M_n - p| \geq 0.01) \leq 0.05$.

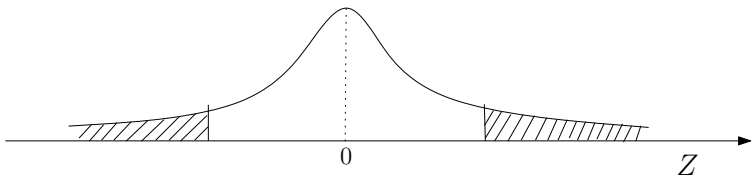

Figure 1: The distribution $M_n - p$ is symmetric, and we are looking for *tail bounds*.

Since the distribution of M_n is symmetric around its mean, we can use this symmetry to write $\Pr(|M_n - p| \geq 0.01) \approx 2\Pr(M_n - p \geq 0.01)$. Because the X_i are Bernoulli(p), we can bound their variance by $1/4$, so $\mathrm{Var}(M_n) \leq \frac{1}{4n}$. Dividing both sides by the square root of the variance to put the probability into the CLT form, we get

$$\Pr(|M_n - p| \geq 0.01) \approx 2\Pr\left(\frac{M_n - p}{\sqrt{1/4n}} \geq \frac{0.01}{\sqrt{1/4n}}\right) \approx 2\Pr(Z \geq 0.01 \cdot 2 \cdot \sqrt{n}) = 0.05.$$

This was valid because we know that $\frac{M_n - p}{\sqrt{1/4n}}$ has a variance that is something less than 1. So its tail probability is approximately no worse than a standard Gaussian. This gives us the equation $\Pr(Z \geq 0.01 \cdot 2 \cdot \sqrt{n}) = 0.025$, where the unknown is n. Equivalently $\int_{0.01 \cdot 2\sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.025$, which has a solution $n \approx 9604$. You can find this using a computer.

In this calculation we used the approximation given by the CLT, but how good is it? If we work out the exact computation, using the fact that M_n is a scaled binomial and the worse-case parameter $p = 1/2$, then we get that $n = 9604$ will give us a 95.1% probability of being within .01 of the true value p: in this case, the CLT's approximation is very good[3]!

So if the CLT is so much better, why use Chebyshev's inequality at all? The reason is that Chebyshev is a bound that works for all ε and probabilities, is always a valid upper bound but is very conservative. The CLT is often much sharper, however it only works for ε of appropriate scale and does not give a rock-solid bound as it is an approximation and not a bound.

Is it possible to strengthen Chebyshev's inequality to get a bound nearly as sharp as the CLT's, under appropriate assumptions? This is what we are going to see in the next section.

Chernoff Bounds: large deviations

Our goal is to better bound the quantity $\Pr(X \geq a)$, i.e. we want to show a *tail bound*. If we were to plot the true probability of the average of many i.i.d. random variables being far from their mean, we would find that it would make approximately a straight line on log-linear paper with the probability measured on a logarithmic scale and n on a linear one. (See Figures 3 and 4.)

That empirical observation drives us to see if we can find an exponential bound for such deviations. The CLT turns out to have an exponential dependence[4] on n, but it is not a proper bound and cannot be trusted for such small probabilities. (See Figures 3 and 4 for a demonstration of how the CLT can go wrong.)

This forces us to start from scratch to get a bound. So we return to Markov's Inequality, but instead of using a quadratic function like we did to derive Chebyshev's Inequality, we try using an exponential.

Let's choose a parameter $s > 0$ and write

$$\begin{aligned} \Pr(X \geq a) &= \Pr(e^{sX} \geq e^{sa}) \\ &\leq \mathbf{E}[e^{sX}]e^{-sa}. \qquad \text{(by Markov's inequality)} \end{aligned}$$

We could do this because exponentials are monotonically increasing as long as the base is larger than 1. The freedom to choose $s > 0$ is just our freedom to choose the base of the exponential we are using. Since this is true for every s, we have

$$\begin{aligned} \Pr(X \geq a) &\leq \min_{s>0} \left(\mathbf{E}[e^{sX}]e^{-sa}\right) \\ &= e^{-\Phi_X(a)} \end{aligned}$$

where $\Phi_X(a) = \max_{s \geq 0} \left(sa - \ln \mathbf{E}[e^{sX}]\right)$ is obtained by taking logarithms and realizing that the minus sign flips the minimization into a maximization. Notice that there is no harm in also allowing $s = 0$ into the

[3] Instead of the CLT, use Chebyshev's inequality to find the number of people n that you need to poll to get the same confidence interval with the same probability. How does this number compare to the one we got using the CLT? (Hint: you should get $n \approx 50000$.)

[4] The x^2 in the exponential cancels the $\sqrt{n}$ to give rise to a single exponential. Work it out for yourself using the bounds we gave above for the tail probability of a Gaussian.

maximization since it gives a valid bound because probabilities are always no bigger than 1. The above minimization can be solved using standard calculus techniques as we will illustrate shortly by means of an example.

Now imagine that we had started with X being an average $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right) = \Pr\left(\sum_{i=1}^{n} X_i \geq na\right) \leq e^{-\Phi_{\sum_{i=1}^{n} X_i}(na)}$, where

$$\Phi_{\sum_{i=1}^{n} X_i}(na) = \max_{s>0}\left(sna - \ln \mathbf{E}[e^{s\sum_{i=1}^{n} X_i}]\right).$$

But $e^{s\sum_{i=1}^{n} X_i} = \prod_{i=1}^{n} e^{sX_i}$, so

$$\ln \mathbf{E}[e^{s\sum_{i=1}^{n} X_i}] = \ln \prod_{i=1}^{n} \mathbf{E}[e^{sX_i}]$$

where we used the independence of the X_i to write that the expectation of the product was equal to the product of the expectations. Since the X_i are i.i.d, $\mathbf{E}[e^{sX_i}]$ is the same regardless of the value for i. This, combined with fact that the log of a product is the sum of the logs, results in

$$\begin{aligned}\Phi_{\sum_{i=1}^{n} X_i}(na) &= n \max_{s>0}\left(sa - \ln \mathbf{E}[e^{sX_1}]\right) \\ &= n\,\Phi_{X_1}(a)\end{aligned}$$

which gives us our final bound:

(Chernoff) $$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right) \leq e^{-n\Phi_{X_1}(a)}.$$

This is the exponentially decreasing bound that we expected from the form of the CLT. All that is required is to verify that the $\Phi_{X_1}(a) > 0$.

Extended Example. Consider our usual "packet loss" network experiment, and define

$$X_i = \begin{cases} 1 & \text{if a packet is dropped} \\ 0 & \text{if the packet is not dropped} \end{cases}$$

Assume that drops are i.i.d. Bernoulli(p). We want to get a bound on the probability that the proportion of drops exceeds the safety margin of an error-correcting code. Assume that our code can recover upto a fraction a of drops. Even Chebyshev's inequality tells us that as long as $a > p$ and our code-length n is large enough, we do not have to fear losing our message. But how large does n have to be to give us a very low probability of error? Or alternatively, if the n we have is known, how low does p have to be?

Suppose that the code was designed for $a = 0.15$. For probabilities like 10^{-6}, Chebyshev's inequality is hopelessly conservative. Having n in the millions would be practically infeasible. Suppose n was something more reasonable like 1024. How high of an underlying loss probability p can we tolerate while still maintaining our 10^{-6} probability guarantee for the loss of the entire message. Chebyshev's inequality would ask us to solve the equation $p(1-p)/(1024*(p-0.15)^2) = 10^{-6}$ for p. This is a quadratic that is easily solved, and gives the pitifully low value of $p = 0.000023$ for the acceptable underlying probability of a lost packet. Could this be true? That a code of length 1024 that can tolerate upto 15% packet drops can only safely be run on an essentially drop-free network? That would be bad news indeed.

This is where Chernoff bounds shine. They reveal that we can easily tolerate an underlying network that randomly drops with a probability of around 9.78%. In reality, we would be safe as long as[5] that drop probability is below[6] 10.23%, and so we can see how close the Chernoff bound is getting us. Meanwhile, the CLT would be slightly aggressive in this case and suggests[7] that we could tolerate up to a 10.45% drop probability.

To calculate the Chernoff bound, we need to evaluate

$$\begin{aligned}\mathbf{E}[e^{sX_1}] &= \sum_x \Pr(X_1 = x) e^{sX_1} \\ &= p\,e^s + (1-p)\cdot 1.\end{aligned}$$

To compute Φ_{X_1} we need to maximize the expression $sa - \ln(pe^s + (1-p))$ over all $s \geq 0$. Taking the derivative with respect to s and setting to 0, we get

$$a - \frac{pe^s}{pe^s + (1-p)} = 0$$

Solving for e^s, we get $e^s = \frac{1-p}{1-a} \cdot \frac{a}{p}$. Since we require $s \geq 0$, we should check that $e^s \geq 1$, which is the case if and only if $a \geq p$ (this makes sense, and it is the interesting case anyways). So we get that

$$\begin{aligned}\Phi_{X_1}(a) &= a \ln \frac{a}{p} + a \ln \frac{1-p}{1-a} - \ln\left((1-p)\frac{a}{1-a} + (1-p)\right) \\ &= a \ln \frac{a}{p} + (a-1) \ln \frac{1-p}{1-a} \\ &= a \ln \frac{a}{p} + (1-a) \ln \frac{1-a}{1-p}\end{aligned}$$

This last expression is called the Kullback-Liebler Divergence, usually denoted by $D(a||p) \geq 0$, and it is 0 only if $a = p$. So in this example we get the bound

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right) \leq e^{-nD(a||p)}.$$

Notice that this expression is exactly what Stirling's approximation predicted back when we first applied it to unfair coin tosses!

Anyway, if we plot this bound against the real value on a log-linear scale, it turns out that the slope given by $D(a||p)$ is asymptotically correct. Experimentation will further reveal that the Chernoff bound tends to overestimate[8] the probability by a factor of roughly $\sqrt{n}$. The deep reasons for this are a bit subtle[9], but are explored in graduate courses in the Statistics department. This turns out to have some implications for machine learning algorithms and artificial intelligence approaches to classification and search.

[5] This exact calculation involves calculating the exact probability that the Binomial $(1024, p)$ random variable is greater or equal to 154 drops. This can be done numerically. You should practice doing this for your own confidence.

[6] Some students find even the 10% number troubling. After all, the code protects against 15% drops so why can't we use it on a network that drops 15% on average? This is actually an important engineering issue and the reason is that our desired reliability is very high. We want the code to work 99.9999% of the time. This requires an extra margin of overdesign.

[7] Can you do this calculation on your own? If you knew $p < 0.15$, the CLT would say that the probability of error is the same as a standard Normal random variable exceeding $\frac{(0.15-p)1024}{1024p(1-p)}$. This can be solved numerically for the critical p that gets us just under 10^{-6}.

[8] Dividing the Chernoff Bound by $n+1$ gives a valid lower bound to the probability. This is proved in EECS 229A.

[9] You would be correct in suspecting that this has some connection to the CLT since the CLT reveals that $\frac{1}{\sqrt{n}}$ governs the essential precision that can be resolved with n samples.

Remark. We got a bound on the upper tail; if we're interested in the lower tail a similar derivation starting from $\Pr(\frac{1}{n}\sum_{i=1}^n X_i \le a) = \Pr(\frac{1}{n}\sum_{i=1}^n (-X_i) \ge -a)$ would result in a calculation of $\Phi_{-X_1}(-a)$ that would in tern yield a similar bound valid for $a \le p$:

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \le a\right) \le e^{-nD(a||p)}.$$

By luck, the expression turns out to look exactly the same. You should do the detailed derivation for yourself to make sure you understand how this sort of argument works.

The homework asks you to evaluate Chernoff-like bounds using other techniques. Those are often useful when you want to generalize to other cases.

Taylor Expansions and connecting the Chernoff Bound to the CLT We have not proved that the Central Limit Theorem holds and so the CLT might seem rather mysterious. You can get some insight into the CLT by exploring how the value of $D(a||p)$ scales when a is in the vicinity of p. To study this, we'll do a Taylor expansion of the KL-divergence. Expand $D(a||p)$ as

$$D(a||p) = a\ln a - a\ln p + (1-a)\ln(1-a) - (1-a)\ln(1-p).$$

Taking the derivative with respect to a gives

$$\begin{aligned}\frac{\partial}{\partial a}D(a||p) &= \ln a + 1 - \ln p - \ln(1-a) - 1 + \ln(1-p)\\ &= \ln\frac{a}{p} + \ln\frac{1-a}{1-p}\end{aligned}$$

So the derivative is 0 at $a = p$. This means that it is flat there where the value of the function is zero. Taking the second derivative gives us

$$\begin{aligned}\frac{\partial^2}{\partial^2 a}D(a||p) &= \frac{1}{a} + \frac{1}{1-a}\\ &= \frac{1}{a(1-a)}.\end{aligned}$$

This is always positive, so $D(a||p)$ is convex in a. Immediately, we can see that $D(a||p) \ge 0$. Furthermore, doing a Taylor expansion[10] in the vicinity of p, this means that $D(p+\varepsilon||p)$ behaves quadratically in ε for small ε. Therefore for small deviations ε we get the following approximate bound that can be massaged into a familiar form.

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \ge p+\varepsilon\right) \le\approx \exp(-n\left(\frac{\varepsilon^2}{2p(1-p)}\right))$$

$$\Pr\left(\frac{(\sum_{i=1}^n X_i) - np}{\sqrt{np(1-p)}} \ge \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \le\approx \exp(-\frac{1}{2}\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right)^2)$$

This shows that the CLT scales correctly in that the exponents agree in the vicinity of the mean. The core-reason is that the relevant second derivative of the divergence is the reciprocal of the variance.

[10] Remember, a Taylor expansion is $f(x+\varepsilon) \approx f(x) + f'(x)\varepsilon + \frac{1}{2}f''(x)\varepsilon^2 + \frac{1}{3!}f'''(x)\varepsilon^3 + \cdots$.

It turns out that one of the most popular proofs for the CLT (the one given in EECS 126, for example) is essentially based on the idea of doing a second-order Taylor expansion and seeing what happens. Here, we've just done it in the specific context of the Binomial random variable, but the same idea holds more generally.

Not only does this Taylor expansion idea reveal where the CLT is coming from, but it also gives justification to the warning that the CLT is an approximation that is only to be used for small deviations, not large ones. You remember from your calculus classes that a Taylor expansion is not a bound: it will hug the desired function for a while, but after that, it can be either higher or lower than the desired function with no guarantees as to which way it goes.

This is worth seeing in some plots. Figure 2 shows a plot of the Divergence from the Chernoff Bound as compared with the quadratic expansion that forms the heart of the CLT. Figures 3 and 4 show how the two bounds compare to the true probabilities for the simple example of the average of a bunch of iid coin tosses.

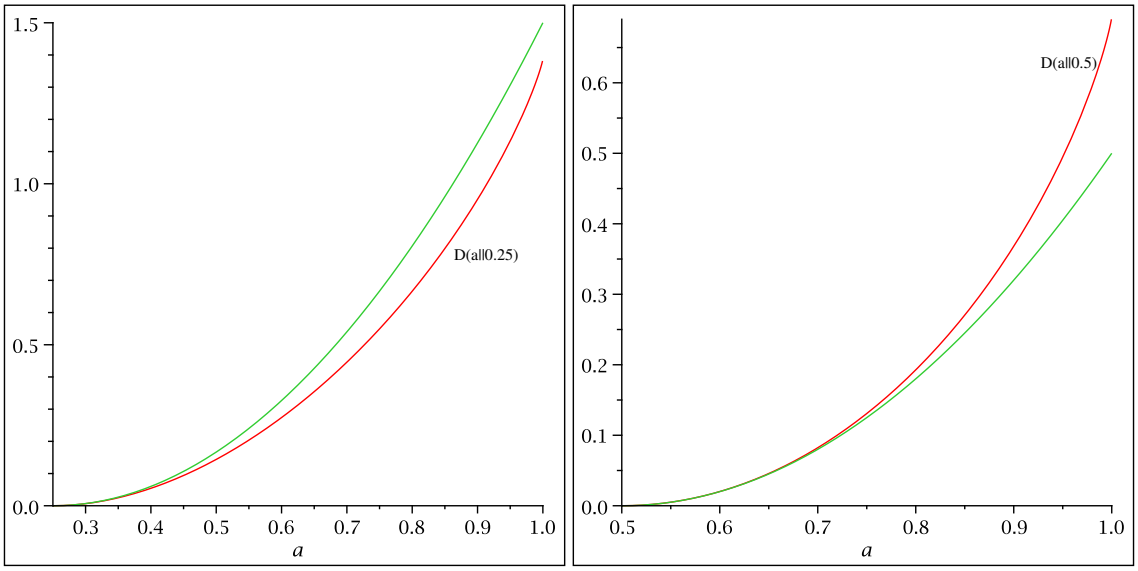

Figure 2: This pair of plots shows how the second-order Taylor expansion (corresponding to the CLT) hugs the divergence, but can be either above or below depending on the value for p. This kind of variation can occur in general — it is not a phenomenon that is limited to the simple Binomial case illustrated here.

As you have done in the virtual labs, you are encouraged to duplicate these plots on your own so you can see what happens as you go to larger n's and so on. Remember, even though this class has a big emphasis on doing proofs and reasoning correctly, never forget that the "experimental" component to learning is important. Play around with this stuff, make your own plots and try to do the same thing for other distributions.

You will never internalize this material until you have played with it enough on your own terms. Lecture, discussion, and reading can only take you so far. The homeworks and exams will push you a bit, but even these are no replacement for the free-form exploratory play that you need to do on your own. The virtual labs were meant to give you the tools you needed to play on your own.

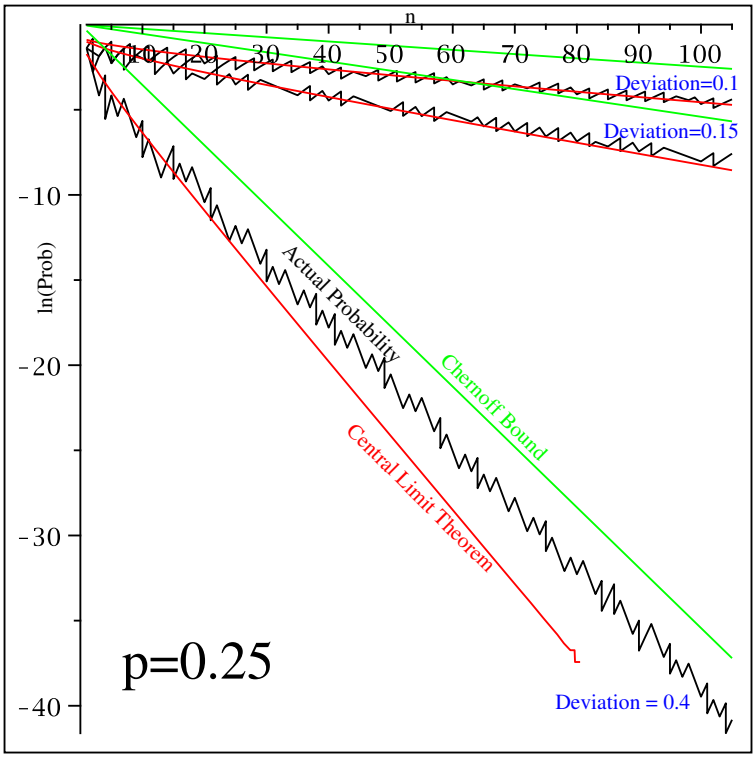

Figure 3: This plots shows how rare a deviation from the mean is as n varies. Notice how the true probabilities are jagged: this is due to an integer effect – it doesn't matter if $n = 21$ or $n = 22$, a deviation by more than 10% still means at least two extra heads in the n coin tosses. The Chernoff bound matches the average slope (on a log scale) of this jagged line perfectly, but is offset: it always gives a probability that is above the true probability. The Central Limit Theorem does not give a true bound, although it interpolates through the true curves quite impressively when the deviation is small. This matches what we would expect looking at how closely the second-order Taylor expansion hugs the Divergence curve, but it also shows that the CLT is getting the sub-exponential terms correct in a way that the Chernoff bound does not. However, the CLT is a misleading estimate when the deviation is large. If you were to use it in place of a bound, it would be overly optimistic and predict much smaller probabilities than what actually happens.

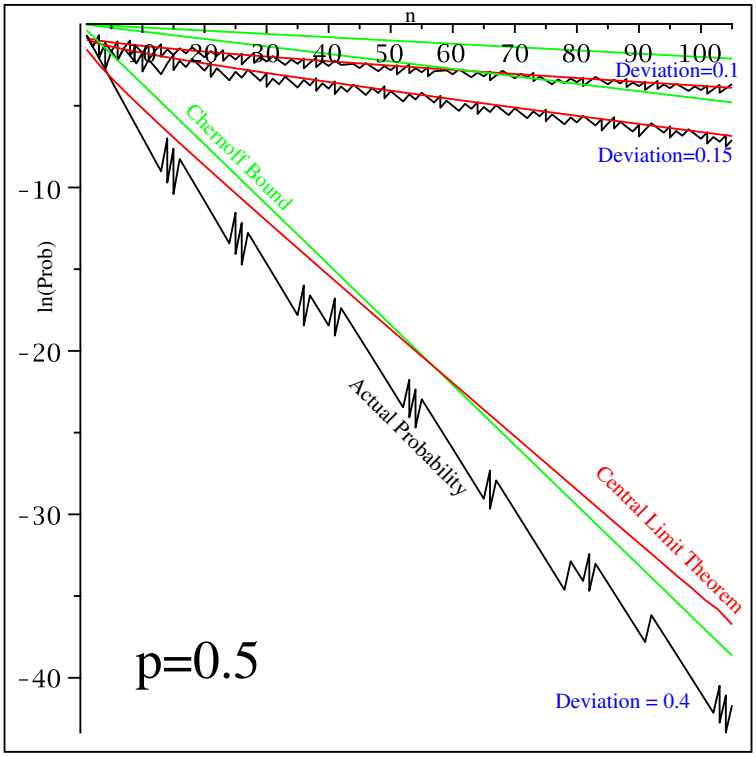

Figure 4: This plots shows how rare a deviation from the mean is as n varies. $p = 0.5$ on this plot and once again, the Chernoff bound matches the average slope (on a log scale) of this jagged line perfectly, but with an offset that keeps it safely above with a little room to spare. Once again, the CLT performs admirably for small deviations, however in this case it is too pessimistic at large deviations predicting larger probabilities than actually occur. The Chernoff bound always gets the slope right.