

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:

$$\begin{array}{ll} \text{(Markov)} & \text{If } X \geq 0, \text{ then} \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \\ \text{(Chebyshev)} & \text{If } X_i \text{ are i.i.d., then} \quad \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right) \leq \frac{\text{Var}(X_1)}{n\varepsilon^2} \end{array}$$

Taking the limit as n goes to infinity, Chebyshev's inequality implies that if the X_i are i.i.d. (independent and identically distributed), then

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right) = 0$$

so the average of the X_i converges to the expectation $\mathbb{E}[X_1]$. 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 $\frac{\sqrt{\text{Var}(X)}}{\sqrt{n}}$.
- 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 usually 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 something that is called a Gaussian (AKA Normal) random variable.

Theorem. (*Central Limit Theorem*) Let X_i be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ 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 all z we have

$$(CLT) \quad \lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

This theorem is sometimes stated as $Z_n \rightarrow N(0, 1)$: Z_n converges in probability towards something called the standard Gaussian with mean 0 and variance 1. To understand this more precisely, we would need to more carefully define something called continuous random variables² — random variables that can take on any real-valued output. For these things, the convergence is the sense that the CDF of Z_n converges³ to the desired continuous function that defines the CDF of what is called an $N(0, 1)$ continuous random variable.

Fortunately, you don’t actually need to understand precisely what a continuous random variable is to be able to use the theorem above. This is because the convergence and limits are used primarily as a justification for the use of the CLT to do *approximations* for finite n . When dealing with an appropriate sum of random variables, we use the CLT to approximate the relevant probabilities. In this way, the CLT plays a role in practice analogous to Stirling’s approximation — it turns something unwieldy into something that you can handle.

How good are these approximations? Pretty good (more on this later) and something called the Berry-Esséen inequality gives a more precise quantification of the speed of this convergence:

$$(\text{Berry-Esséen}) \quad \left| \mathbb{P}(Z_n \leq z) - \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right| \leq \frac{0.77 \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^3]}{(\text{Var}(X_1))^{3/2} \sqrt{n}}.$$

The proof of this inequality is beyond the scope of this course — you might see proofs of this in graduate courses on the subject.

The convergence can be much faster than that⁴, but if all we know is that there exists a third moment

¹Another 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 what is called a Gaussian random variable. This is used to justify Gaussian distributions for thermal noise that results from the combined random motion of many molecules or electrons.

²You’ll learn more about continuous random variables elsewhere, but one of the consequences of the uncountability of the real numbers is that for a continuous random variable Z , it is not meaningful to ask about $\mathbb{P}(Z = z)$ for any specific real number z . Such probabilities are always zero and basically meaningless. There are just too many real numbers. Instead, we look at something else called the Cumulative Distribution Function (CDF) which asks instead $\mathbb{P}(Z \leq z)$. This probability must tend to zero for $z \rightarrow -\infty$ and tend to one for $z \rightarrow +\infty$ and is a monotonically increasing continuous function in between. This function is a discontinuous piece-wise step function for a discrete random variable. Mixtures of discrete and continuous random variables also exist in-between where the function has some jump discontinuities and other places where it grows continuously.

³Elsewhere, you’ll see that this is exactly the same kind of convergence used to define other standard continuous random variables. For example, appropriately scaled geometric random variables converge to something called a continuous-valued exponential random variable. And appropriately labeled dice with increasing number of sides converge to something called a continuous-valued uniform random variable.

⁴You are strongly encouraged to use a computer to plot for yourself how fast the CDF converges for the following example: X_i is a fair roll of a six-sided die. The case of Bernoulli random variables is illustrated at the end of this note. Both of these converge quite rapidly.

$\mathbb{E}[|X_1 - \mathbb{E}[X_1]|^3]$, 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 much 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 weight variance is 168.75lbs. What is the probability that the total weight is larger than 3000lbs?

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

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

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.

If we use the CLT instead,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{100}X_i \geq 3000\right) &= 1 - \mathbb{P}\left(\sum_{i=1}^{100}X_i < 3000\right) \\ &= 1 - \mathbb{P}\left(\frac{\sum_{i=1}^{100}X_i - 2750}{\sqrt{100 \cdot 168.75}} < \frac{250}{\sqrt{100 \cdot 168.75}}\right) \\ &\approx 1 - \mathbb{P}\left(Z < \frac{250}{\sqrt{100 \cdot 168.75}}\right) \approx 1 - \mathbb{P}(Z < 1.92) \approx 0.028.\end{aligned}$$

where Z is a random variable distributed as $N(0, 1)$.

The advantage of doing this is that people have published tables of these probabilities and we can just look them up. For example, in Figure 2.

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 (i.e. CLT-based) 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 further approximations will often prove useful: for any $z > 0$,

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

This way, you can approximate the tail probabilities 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 a particular candidate. 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 $\mathbb{P}(|M_n - p| \geq 0.01) \leq 0.05$.

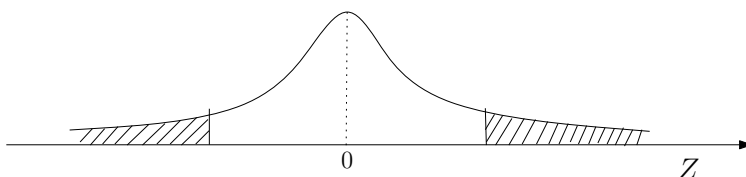


Figure 1: The distribution $M_n - p$ is symmetric, and we are looking for *tail bounds*.

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

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

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 $\mathbb{P}(Z \geq 0.01 \cdot 2 \cdot \sqrt{n}) = 0.025$, where the unknown is n . Equivalently $\int_{0.01 \cdot 2 \cdot \sqrt{n}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.025$, 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 worst-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 good⁵!


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 $\mathbb{P}(X \geq a)$, 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 4 and 5.)

⁵Instead 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 $n \approx 50000$.)



Probability Content from $-\infty$ to Z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure 2: A standard table of the values for the integral $\int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ for values of z that start at 0 and get bigger from there. The rows correspond to the most significant digits of z and the columns for the least significant digits. The symmetry of the function inside the integral can be used to understand negative values for z as well. Essentially, if $z < 0$, you look up $-z$ and then subtract that number from 1.

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 dependence⁶ on n , but it is not a proper bound and cannot be trusted for such small probabilities. (See Figures 4 and 5 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

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

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

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

where $\Phi_X(a) = \max_{s \geq 0} (sa - \ln \mathbb{E}[e^{sX}])$ 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 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 $\mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i \geq a) = \mathbb{P}(\sum_{i=1}^n X_i \geq na) \leq e^{-\Phi_{\sum_{i=1}^n X_i}(na)}$, where

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

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

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

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., $\mathbb{E}[e^{sX_i}]$ 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

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

which gives us our final bound:

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

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

⁶The 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.

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 desired 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 as⁷ that drop probability is below⁸ 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 suggests⁹ that we could tolerate up to a 10.45% drop probability.

To calculate the Chernoff bound, we need to evaluate

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

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

⁷This 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.

⁸Some 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.

⁹Can 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} .

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-Leibler 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

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

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 overestimate¹⁰ the probability by a factor of roughly \sqrt{n} . The deep reasons for this are a bit subtle¹¹, 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.

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

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

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.

Hoeffding's Inequality: a looser but Chernoff-like bound It turns out that the Chernoff-bounding technique described above tends to give exponentially-tight (i.e. the bound gets the exponential rate of decay with n correct) bounds not just in the case of iid Bernoulli random variables, but most other cases as well. The cost of this tightness is complexity and algebraic difficulty. For example, the divergence $D(a||p)$ term is not easy to solve for either a or p analytically — numeric techniques have to be used. Consequently, there is great demand for easier-to-use bounds where the probability of a tail event is bounded below something going to zero exponentially.

The basic setting¹² here is for independent random variables X_i that are all known to be bounded within an interval $[a, b]$ and all have an identical mean $\mathbb{E}[X_i] = \mu$. Let $A_n = \frac{1}{n} \sum_{i=1}^n X_i$ be the average of n such

¹⁰Dividing the Chernoff Bound by $n+1$ gives a valid lower bound to the probability. This is proved in EECS 229A.

¹¹You 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.

¹²Look in Wikipedia or any more advanced reference to see how this inequality can be generalized to cases where the means are different or the bounds are different.

independent random variables. Hoeffding's Inequality gives a pair of bounds on the deviation from the mean for this average. For all $\varepsilon > 0$,

$$\mathbb{P}(A_n \geq \mu + \varepsilon) \leq e^{-n(\frac{2\varepsilon^2}{(b-a)^2})}, \quad (1)$$

$$\mathbb{P}(A_n \leq \mu - \varepsilon) \leq e^{-n(\frac{2\varepsilon^2}{(b-a)^2})}. \quad (2)$$

The only thing that needs to be checked is independence and boundedness. The boundedness assumption here is key. So it can work with random variables like Bernoullis but not those like Geometrics. Using Hoeffding's inequality doesn't even need us to calculate a variance, much less do any calculus!

A decent proof of Hoeffding's inequality is found in Wikipedia, and it is derived from the Chernoff bound. The key steps are using the convexity of the e^{sx} function to get an upper-bound on the $\mathbb{E}[e^{sX}]$ term, and then essentially maximizing that upper-bound over all possible distributions for X_i . This gives a worst-case bound that is comparable to how when talking about variances of Bernoulli random variables, we say that $\frac{1}{4}$ is the worst one. The trick to do this is to use the elementary properties of Taylor expansions.

To see just how easy to use Hoeffding's inequality is, we will apply it to the iid Bernoulli case for our usual packet loss network experiment. If the code was designed for 0.15 fraction of drops, we know that $\mu + \varepsilon = 0.15$. If $n = 1024$ and we want to hit a probability of error of 10^{-6} we just solve $10^{-6} = e^{-1024 \frac{2\varepsilon^2}{1^2}}$ so $\mu = 0.15 - \sqrt{\frac{6 \ln 10}{2 * 1024}} = 0.15 - 0.0821 = 0.0679$. This gives a very reasonable value of a drop probability of around 7%. Not as good as the Chernoff Bound, but much better than Chebyshev's inequality.

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 $D(a||p) \geq 0$. Furthermore, doing a Taylor expansion¹³ in the vicinity of p , this means that $D(p + \varepsilon||p)$ behaves quadratically in ε for

¹³Remember, a Taylor expansion is $f(x + \varepsilon) \approx f(x) + f'(x)\varepsilon + \frac{1}{2}f''(x)\varepsilon^2 + \frac{1}{3!}f'''(x)\varepsilon^3 + \dots$.

small ε . Therefore for small deviations ε we get the following approximate bound that can be massaged into a familiar form.

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq p + \varepsilon\right) \leq \approx \exp\left(-n\left(\frac{\varepsilon^2}{2p(1-p)}\right)\right)$$

$$\mathbb{P}\left(\frac{(\sum_{i=1}^n X_i) - np}{\sqrt{np(1-p)}} \geq \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \leq \approx \exp\left(-\frac{1}{2}\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right)^2\right)$$

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.

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 3 shows a plot of the Divergence from the Chernoff Bound as compared with the quadratic expansion that forms the heart of the CLT. Figures 4 and 5 show how the two bounds compare to the true probabilities for the simple example of the average of a bunch of iid coin tosses.

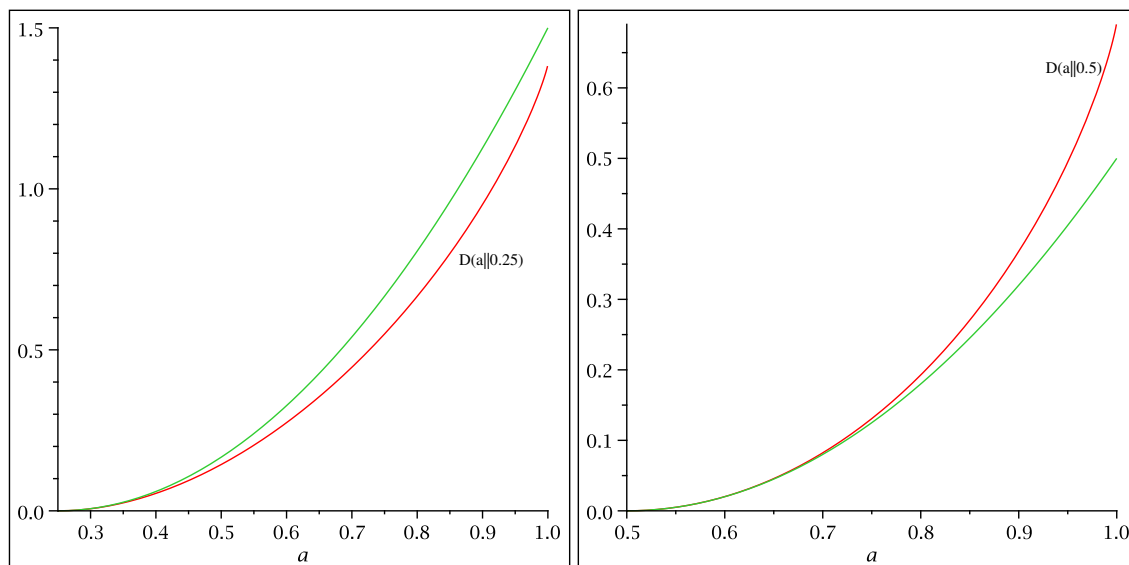


Figure 3: This pair of plots shows how the second-order Taylor expansion (corresponding to the CLT) hugs the divergence, but can be either above or below depending on the value for p . This kind of variation can occur in general — it is not a phenomenon that is limited to the simple Binomial case illustrated here.

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.

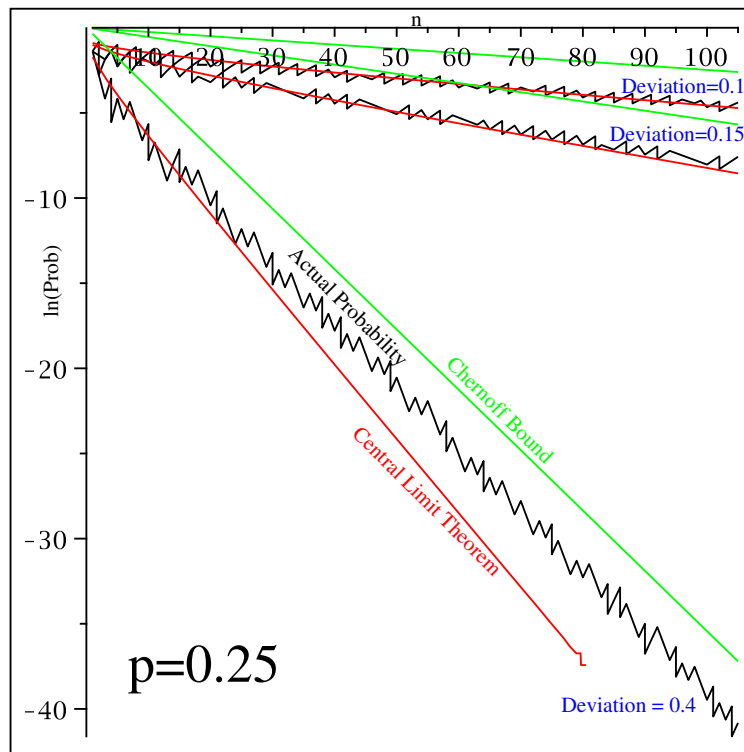


Figure 4: This plots shows how rare a deviation from the mean is as n varies. Notice how the true probabilities are jagged: this is due to an integer effect – it doesn't matter if $n = 21$ or $n = 22$, a deviation by more than 10% still means at least two extra heads in the n coin tosses. The Chernoff 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.

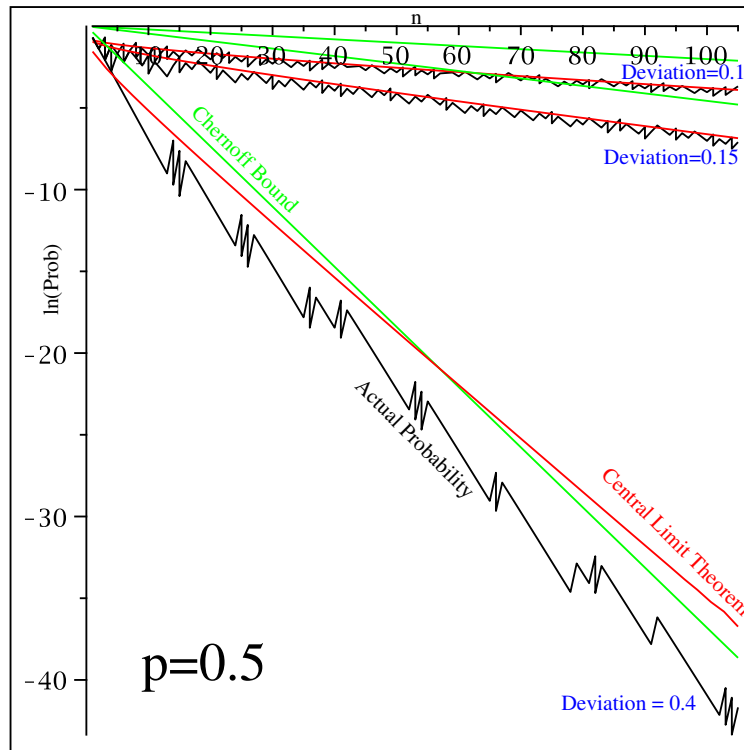


Figure 5: This plots shows how rare a deviation from the mean is as n varies. $p = 0.5$ on this plot and once again, the Chernoff bound matches the average slope (on a log scale) of this jagged line perfectly, but with an offset that keeps it safely above with a little room to spare. Once again, the 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.

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.