The central limit theorem

We just saw that when we add two independent normal random variables, we get another normal random variable. If you look back at the other examples, you might notice that when we added two independent random variables that are not normal, their sum has a distribution that looks a bit "more normal." This is not a coincidence.

The **central limit theorem** states that the sum of a large number of independent random variables (with finite variance) has a normal distribution. What makes this statement so powerful is that is does not matter what the distributions of the random variables you are adding together is, or even that they are the same. The CLT is the reason normal random variables play such an important role in probabilistic modeling. In many cases (e.g., wireless communications), you may know something like the "noise" you have to deal with is the result of many random things (e.g., the activity of other devices emitting RF energy) being aggregated. Even without knowing the individual distributions, you can make a case that the aggregate should be modeled as a Gaussian random variable.

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed (i.i.d.) random variables with finite variance. Then

$$\lim_{N\to\infty} \sum_{i=1}^{N} X_i \longrightarrow \text{normally distributed random variable.}$$

(There are also versions of the CLT with the i.i.d. assumptions relaxed, but we will start simple here.)

To be more precise, the CLT states that if the X_i have

$$\begin{aligned} \text{mean} \quad & \mathrm{E}[X_i] = \mu, \\ \text{variance} \quad & \mathrm{E}[(X_i - \mu)^2] = \sigma^2, \end{aligned}$$

then

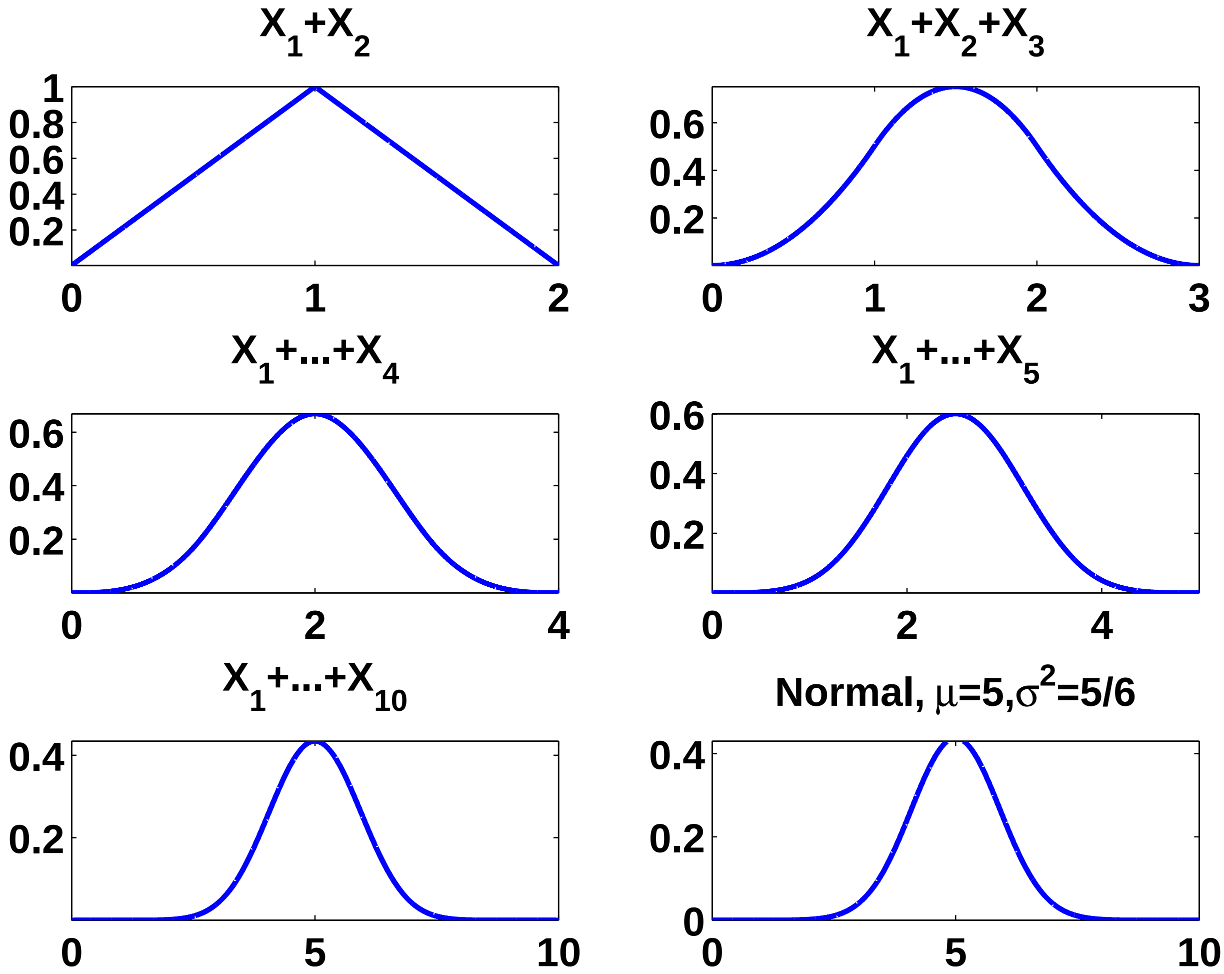
$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma} \right) \sim \text{Normal}(0,1),$$

or equivalently

$$\sum_{i=1}^{N} X_i \approx \text{Normal}(N\mu, N\sigma^2) \quad \text{for large } N.$$

We won't prove the CLT here. It's actually not that difficult, but it relies on a tool that we haven't discussed in this class. Essentially, we want to convolve a bunch of PDFs together and see what the result is. From signals and systems, we know that we can take Fourier transforms and turn this into multiplication in the frequency domain. When doing this with each PDF and writing the individual PDFs in terms of their Taylor series expansions, we can see that in the limit of all of these multiplications the only terms left in the resulting Taylor approximation are the ones corresponding to the Fourier transform of a Gaussian distribution.

Example. Suppose that $X_1, X_2, \ldots$ are uniform random variables on $[0,1]$. Then $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$. If we calculate the pdfs for the partial sums $X_1 + X_2 + \cdots + X_k$ we obtain:



Example. We have modeled *arrival times* using exponential random variables many times:

$$\begin{aligned} f_X(x) &= \lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0, \\ \mu = \mathrm{E}[X] &= \frac{1}{\lambda}, \\ \sigma^2 &= \frac{1}{\lambda^2}. \end{aligned}$$

If the difference between arrival times are exponential random variables, then the distribution of the k^{th} arrival time has what is called an *Erlang* distribution,

$$Y = X_1 + X_2 + \cdots + X_k \;\sim\; k^{\text{th}} \text{ order Erlang},$$

with pdf

$$f_Y(y) = \frac{\lambda^k y^{k-1} \mathrm{e}^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$

and mean $\mu = k/\lambda$ and variance $\sigma^2 = k/\lambda^2$.

The CLT can make it easier to do calculations with more complicated pdfs like the Erlang distribution. For example, take $\lambda = 1000$ above, and suppose we would like to know

$$\begin{aligned} &\mathrm{P}\left(\#\text{arrivals} < 900 \text{ in one second}\right) \\ &\quad = \mathrm{P}\left(900^{\text{th}} \text{ arrival comes after one second}\right) \\ &\quad = \int_1^\infty \frac{(1000)^{900} y^{899} \mathrm{e}^{-1000y}}{899!}\, \mathrm{d}y. \end{aligned}$$

This is a quantity that seems quite difficult to compute.

But the CLT tells us that

$$X_1 + \cdots + X_{900} \approx \text{Normal}\left(\frac{900}{1000}, \frac{900}{1000^2}\right)$$
$$= \text{Normal}(0.9, .0009),$$

and the probability that this exceeds 1 is:

$$1 - \Phi\left(\frac{1-0.9}{\sqrt{.0009}}\right) = 1 - \Phi(3.33) = 1 - 0.9996$$
$$= .0004$$

so about 0.04%. Much easier than trying to calculate the integral directly!

Example. You are riding on an elevator with 10 people on it (including you). The elevator is rated for 2000 pounds. You are considering whether to let an 11^{th} person on ...

Here are (what we will assume are) the facts:

1. The elevator is rated for ≤ 2000 pounds.
2. The average weight of a person is $\mu = 150$ pounds.
3. The standard deviation for a person's weight is $\sigma = 30$ pounds (so the variance is $\sigma^2 = 900$).

With the 11^{th} person on, the probability of not exceeding the limit is

$$\begin{aligned}
\mathrm{P}\left(X_1 + \cdots + X_{11} < 2000\right) &\approx \Phi\left(\frac{2000 - 1650}{\sqrt{9900}}\right) \\
&= \Phi(3.52) \\
&\approx 0.99978,
\end{aligned}$$

so the failure probability is very small.

Now let's find the largest number of people we can have in the elevator so that

$$\mathrm{P}\left(\text{Failure}\right) \;\leq\; 0.01 \qquad (1\%).$$

With k people on, their total weight is

$$X_1 + X_2 + \cdots + X_k \;\approx\; \text{Normal}(150k,\ 900k)$$

From the table, we see that

$$\Phi^{-1}(0.99) \approx 2.33,$$

This means that k must satisfy

$$\frac{2000 - 150k}{\sqrt{900k}} \geq 2.33.$$

We can solve for the critical value of k by taking $L = \sqrt{k}$, then solving

$$\begin{aligned}
150\,L^2 + 69.9\,L - 2000 = 0& \\
\Rightarrow L &= 3.4259 \\
\Rightarrow k &= L^2 = 11.7369,
\end{aligned}$$

so only let 11 people in!

Exercise:

A roulette wheel has 38 slots. 18 are red, 18 are black, 2 are green. You are betting \$20 on black repeatedly, for which you are given even odds.

Sketch the pmf for your earnings X_i on one trial.

Note that

$$\mu = 20\left(\frac{18}{38} - \frac{20}{38}\right) = -\frac{20}{19}$$

and

$$\begin{aligned}\sigma^2 &= 400\left(\frac{18}{38} + \frac{20}{38}\right) - \left(\frac{20}{19}\right)^2 \\ &= 398.892,\end{aligned}$$

so $\sigma = 19.8889$.

You play 200 times. Estimate

(a) the probability you have made any money;

(b) the probability you have lost at least \$300;

(c) the probability you have made at least \$300.

Approximation to the Binomial

Since it is a special case that we often use, let's write down the approximation for the "sum of independent Bernoullis" (i.e. the Binomial) explicitly.

Suppose the X_i are i.i.d. Bernoulli random variables

$$p_{X_i}(k) = \begin{cases} p & k = 1 \\ 1-p & k = 0. \end{cases}$$

Then

$$S_N = X_1 + X_2 + \cdots + X_N$$

is binomial with pmf

$$p_{S_N}(k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad 0 \le k \le N,$$

and

$$\begin{aligned} \mathrm{E}[S_N] &= Np \\ \mathrm{var}(S_N) &= Np(1-p). \end{aligned}$$

Using the CLT, we can make the approximation

$$\begin{aligned} \mathrm{P}\left(k \le S_N \le \ell\right) &= \mathrm{P}\left(\frac{k - Np}{\sqrt{Np(1-p)}} \le \frac{S_N - Np}{\sqrt{Np(1-p)}} \le \frac{\ell - Np}{\sqrt{Np(1-p)}}\right) \\ &\approx \Phi\left(\frac{\ell - Np}{\sqrt{Np(1-p)}}\right) - \Phi\left(\frac{k - Np}{\sqrt{Np(1-p)}}\right). \end{aligned}$$

Standard normal cdf table, $\Phi(y)$ for $y \geq 0$:

	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
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998