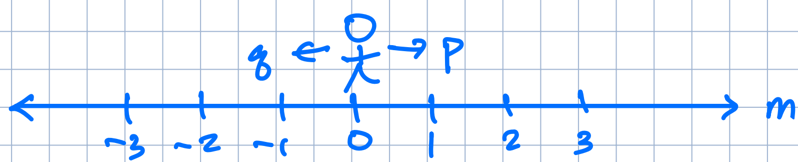
Lecture 4 - The Central Limit Theorem

A. Random Walk

As we mentioned above in our discussion on Stirling's approximation, in stat thermo, we care about many, many random variables. There are two important theorems in probability that describe behavior in the limit of many independent RVs.

Example: Random walk

consider a "drunkard's walk"



Each step is a "Bernoulli trial", n_R steps right w/ probability p and n_L steps left with probability q.

$$p(m) = \frac{N!}{n_R!\, n_L!} p^{n_R} q^{n_L}$$

After $N = n_R + n_L$ steps the man is at what position $m = n_R - n_L$?

$$p(m) = \frac{N!}{\left[\frac{1}{2}(N+m)\right]!\left[\frac{1}{2}(N-m)\right]!} p^{1/2(N+m)} q^{1/2(N-m)}$$

What happens as $N \to \infty$?

$$\lim_{N\to\infty} p(m) = \left(\frac{2}{\pi N}\right)^{1/2} \exp\left(-\frac{m^2}{2N}\right)$$

Approaches a normal distribution with mean $= 0$

and variance $= N$ ← variability decreases (std dev $\sim \sqrt{N}$)

Why? (You will prove this on your HW.)

B. Law of Large numbers

Suppose $x_1, x_2, \ldots, x_N$ are independent and identically distributed (iid) RVs of an unknown distribution with mean m and variance σ^2. The RV.

$$z = \frac{1}{N}(x_1 + x_2 + \ldots + x_N)$$

has a mean m and variance σ^2/N in the limit $N \to \infty$.

What does this mean? The mean (average) converges to what we would expect when N is very large. Also, the variance of the mean gets smaller and smaller (with rate $N^{-1/2}$).

This is the "deterministic limit."

Proof:

$$z = \frac{1}{N}(x_1 + x_2 + \ldots + x_N) = \frac{x_1}{N} + \frac{x_2}{N} + \ldots \frac{x_N}{N}$$

$$M_1(s) = E\left[e^{s x_1/N}\right] \leftarrow \text{moment generating functions}$$

$$M_2(s) = E\left[e^{s x_2/N}\right] \quad \text{of } x_i/N$$

$$\vdots$$

$$M_N(s) = E\left[e^{s x_N/N}\right]$$

By the factorization property:

$$M_z(s) = M_1(s)\, M_2(s) \cdots M_N(s) = \prod_{i=1}^{N} M_i(s)$$

Expand $M_i(s)$ using the Taylor series:

$$E\left[e^{sx_i}\right] = 1 + ms + \frac{\sigma^2 s^2}{2} + \mathcal{O}(s^3)$$

$$E\left[e^{sx_i/N}\right] = M_i(s) = 1 + \frac{ms}{N} + \frac{\sigma^2 s^2}{2N^2} + \mathcal{O}(s^3/N^3)$$

Now put both expressions together:

$$M_z(s) = \left[1 + \frac{ms}{N} + \frac{\sigma^2 s^2}{2N^2} + \mathcal{O}(s^3/N^3)\right]^N$$

Recall that

$$\lim_{N\to\infty}\left[1 + \frac{z}{N}\right]^N = \exp(z)$$

Therefore

$$\lim_{N\to\infty} M_z(s) = \lim_{N\to\infty}\left[1 + \frac{1}{N}\underbrace{\left(ms + \frac{\sigma^2 s^2}{2N} + \mathcal{O}(s^3/N^2)\right)}_{z}\right]^N$$

$$\lim_{N\to\infty} M_z(s) = \exp\left[ms + \frac{\sigma^2 s^2}{2N} + \mathcal{O}\left(\frac{s^3}{N^2}\right)\right]$$

The cgf is

$$Z(s) = ms + \frac{\sigma^2 s^2}{2N} + \ldots$$

$$\text{mean} = \left.\frac{dZ}{ds}\right|_{s=0} = m + \left.\frac{\sigma^2 s}{N}\right|_{s=0} = m$$

$$\text{var} = \left.\frac{d^2 Z}{ds^2}\right|_{s=0} = \frac{\sigma^2}{N}$$

$$\boxed{\text{mean} = m, \quad \text{var} = \sigma^2/N}$$

C. Central limit theorem (Lindeberg–Lévy theorem)

Suppose that $X_1, X_2, \ldots, X_N$ are IID RVs with mean m and variance σ^2. If Y is defined as

$$Y = \sum_{i=1}^{N} \left(\frac{X_i - m}{\sigma\sqrt{N}} \right)$$

mean $=0$
var $=1$
$\mathcal{N}(0,1)$

then the pdf of Y converges to a standard normal distribution in the limit that $N \to \infty$.

What does this mean? Sums of normalized independent variables converge to the normal distribution for large N!

Proof (very similar): scale by $\sqrt{N}$ instead

$$Y = \frac{X_1 - m}{\sigma\sqrt{N}} + \frac{X_2 - m}{\sigma\sqrt{N}} + \ldots + \frac{X_N - m}{\sigma\sqrt{N}}$$

let $W_i = \frac{X_i - m}{\sigma}$ (write less)

$$Y = \frac{W_1}{\sqrt{N}} + \frac{W_2}{\sqrt{N}} + \ldots + \frac{W_N}{\sqrt{N}}$$

$m_{W_i} = 0 \qquad \sigma^2_{W_i} = 1$

We want to find the mgf of Y:

$$
\begin{aligned}
M_Y(s) &= E\left[\exp(Ys)\right] \\
&= E\left[\exp\left(\frac{W_1 s}{\sqrt{N}} + \frac{W_2 s}{\sqrt{N}} + \ldots + \frac{W_N s}{\sqrt{N}}\right)\right] \\
&= E\left[\exp\left(\frac{W_1 s}{\sqrt{N}}\right)\exp\left(\frac{W_2 s}{\sqrt{N}}\right) \cdots \exp\left(\frac{W_N s}{\sqrt{N}}\right)\right]
\end{aligned}
$$

The Z_i's are independent (factorization)

$$= E\left[\exp\left(\frac{W_1 s}{\sqrt{N}}\right)\right] E\left[\exp\left(\frac{W_2 s}{\sqrt{N}}\right)\right] \cdots E\left[\exp\left(\frac{W_N s}{\sqrt{N}}\right)\right]$$

$$= \prod_{i=1}^{N} \boxed{E\left[\exp\left(\frac{w_i s}{\sqrt{N}}\right)\right]}$$

Expand $\square$ as a Taylor series

$$E[e^{sx}] = 1 + ms + \frac{s^2\sigma^2}{2} + O(s^3)$$

scale $s \to s/\sqrt{N}$

$m=0$, $\sigma^2=1$ for w_i

$$E\left[\exp\left(\frac{w_i s}{\sqrt{N}}\right)\right] = 1 + 0 \cdot \frac{s}{\sqrt{N}} + \frac{s^2 \cdot 1}{2N} + O\left(s^3/N^{3/2}\right)$$

$$= 1 + \frac{s^2}{2N} + O\left(s^3/N^{3/2}\right)$$

Now plug back into $M_Y(s)$:

$$M_Y(s) = \prod_{i=1}^{N}\left[1 + \frac{s^2}{2N} + O\left(s^3/N^{3/2}\right)\right] = \left[1 + \frac{s^2}{N} + O\left(\frac{s^3}{N^{3/2}}\right)\right]^N$$

Now, recall that

$$\lim_{N\to\infty}\left[1 + \frac{x}{N}\right]^N = \exp(x)$$

Scaling by $\sqrt{N}$ means only terms w/ N are $O(s^3)$

Therefore

$$\lim_{N\to\infty}\left[1 + \frac{s^2}{2N}\right]^N = \exp(s^2/2)$$

$$\boxed{\lim_{N\to\infty} M_Y(s) = \exp(s^2/2)}$$

The mgf for a normal distribution is

$$M(s) = \exp(ms + \sigma^2 s^2/2)$$

By comparison, the pdf is normal with mean $= 0$ and variance $= 1$.