Central Limit Theorem (CLT) Chapter 7.3

- Considers the <u>CDF</u> of the sample mean

- intuition: Sample mean has a CDF the approaches the Gaussian CDF as we increase the number of samples ($n \to \infty$)

$\Rightarrow$ we can use the Gaussian CDF to approximate the CDF of the sample mean when we have many samples

- This is a key reason the Gaussian distribution is so useful and important.

Let $X_1, X_2, \ldots, X_n$ be a sequence of iid RVs of <u>any</u> distribution (discrete or continuous).

Then, if $\mu = E(X)$ and $\sigma^2 = Var(X)$ are both finite, and we define

$$Z_n = \frac{nM_n - n\mu}{\sigma\sqrt{n}} = \frac{M_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{M_n - \mu}{\sigma}$$

then Z_n has zero mean and unit variance

and

$$\lim_{n \to \infty} P(Z_n \leq z) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{-x^2/2}\, dx = \Phi(z)$$

In words: The CDF of Z_n is well-approximated by the CDF of a Gaussian w/ mean 0 and variance 1.

And the approximation gets better as n increases.

How can we use this?

Recall $Z = \frac{X-\mu}{\sigma}$ is Gaussian with mean 0 variance 1, if X is Gaussian with mean μ, variance σ^2.

So CLT says that

M_n is "approximately Gaussian" (or more precisely, that the CDF of M_n is approximately a Gaussian CDF)

and $E(M_n) = \mu$ and $Var(M_n) = \frac{\sigma^2}{n}$

So

$$Z_n = \frac{M_n - E(M_n)}{\sqrt{Var(M_n)}} = \frac{M_n - \mu}{\sigma/\sqrt{n}} = Z_n$$

Z_n has a CDF that is well approximated by a Gaussian CDF w/ mean 0 variance 1

Example of applying CLT

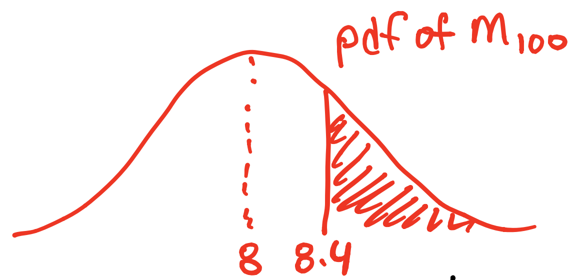
The orders at a restaurant are IID, $\mu = \$8$ and $\sigma = \$2$.

Estimate the probability the first 100 customers spend a total of more than \$840.

$$M_{100} = \frac{1}{100} \sum_{i=1}^{100} X_i$$ ≡ the RV of how much the 1st 100 customers spend

$$E(M_{100}) = \mu = 8$$

$$Var(M_{100}) = \frac{\sigma^2}{100} = 0.04$$



By the Central Limit Theorem, we know the CDF of M_{100} is well-approximated by a Gaussian CDF

so $$Z_{100} = \frac{M_{100} - \mu}{\sigma/\sqrt{100}} = \frac{M_{100} - \mu}{\sigma/10} = \frac{M_{100} - 8}{2/10}$$

So we want $(M_{100})100 > 840 \Rightarrow M_{100} > 8.4$

$$\Rightarrow Z_{100} = \frac{M_{100} - 8}{2/10} > \frac{8.4 - 8}{2/10} = \frac{4/10}{2/10} = 2$$

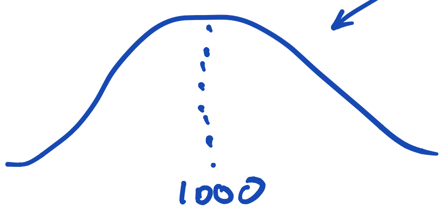
so $P(Z_{100} > 2) = 1 - \Phi(2) = 0.0228$

How many orders are necessary for us to be 90% certain that the total spent by all customers is more than \$1000.

Intuition: As n increases, the mean M_n increases. When is n large enough to be "confident" that nM_n is larger than 1000

If M_{125}

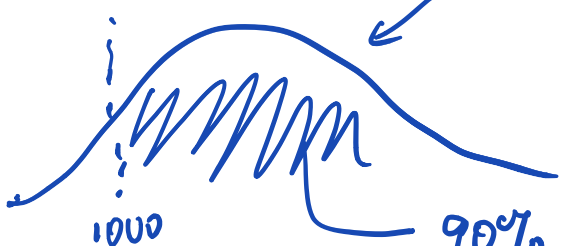
CDF of $125\,M_{125}$



$E(125\,M_{125}) = 1000$

so not yet big enough.

CDF of nM_n for desired goal



$$Z_n = \frac{M_n - \mu}{\sigma/\sqrt{n}}$$

want $P(nM_n > 1000) > 0.90$

$$P\left(M_n > \frac{1000}{n}\right) = P\left(\frac{M_n - \mu_n}{\sigma/\sqrt{n}} > \frac{\frac{1000}{n} - \mu_n}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z_n > \frac{\sqrt{n}\left(\frac{1000}{n} - \mu\right)}{\sigma^2}\right)$$

(and simplify to get)

$$= P\left(Z_n > \frac{1000 - 8n}{2\sqrt{n}}\right) = 1 - \Phi\left(\frac{1000 - 8n}{2\sqrt{n}}\right) = 0.9$$

from the table, $\dfrac{1000 - 8n}{2\sqrt{n}} = -1.28 \Rightarrow \sqrt{n} = 11.34$

$n = 128.6 \Rightarrow$ need 129 customers

<u>Example</u> Suppose the times between events are iid exponential RVs with mean μ. Find the probability that the 1000^{th} event occurs in the time interval $(1000 \pm 50)\mu$ i.e. between 950μ and 1050μ.

Let X_n be the time between events (exponential)

$S_n = \sum_{i=1}^{n} X_i = n M_n$ is the time of the n^{th} event.

$$E(X_i) = \mu \quad \text{so} \quad E(S_n) = n\mu$$

$$Var(X_i) = \mu^2 \quad \text{so} \quad Var(S_n) = n\mu^2$$

By CLT, let $Z_n = \dfrac{S_n - n\mu}{\sqrt{n}\,\mu}$ (Gaussian zero mean unit var.)

$$Z_{1000} = \frac{S_{1000} - 1000\mu}{\sqrt{1000}\,\mu}$$

$$P\left(\frac{950\mu - 1000\mu}{\sqrt{1000}\,\mu} \le Z_{1000} \le \frac{1050\mu - 1000\mu}{\sqrt{1000}\,\mu}\right)$$

$$= \Phi\left(\frac{50\mu}{\sqrt{1000}\,\mu}\right) - \Phi\left(-\frac{50\mu}{\sqrt{1000}\,\mu}\right)$$

$$= \Phi(1.58) - \Phi(-1.58) = 1 - 2\Phi(-1.58)$$

$$= 0.9418 - 0.0582 = 0.8836$$

The fact that X_i is exponential is immaterial!

The Central Limit Theorem works for discrete RVs too. Example: binomial

Binomial RV is a sum of iid Bernoullis.

Let X be binomial, mean np, variance $np(1-p)$

Let Y be Gaussian, same mean and variance

For large n, $P(X=k) \approx P\left(k-\frac{1}{2} \le Y \le k+1/2\right)$

We could compute this using the Φ function as usual.

Or we could recognize that for large n, the interval $\left[k-\frac{1}{2}, k+\frac{1}{2}\right]$ is quite narrow

Simplify this approximation further by

$$P(X=k) \approx \frac{\exp\left(-(k-np)^2/2np(1-p)\right)}{\sqrt{2\pi np(1-p)}}$$