Last time:

Sec. 6.4 Chebyshev's inequality

Sec. 7.1, 7.2 Variance of sums (Binom, Poisson, NegBinom, HyperGeom)

Sec 7.3 Law of averages / Law of large number

Today: the Central Limit Theorem "CLT"

Sec. 8.1 The distr. of an i.i.d. sum.

$X_1, X_2, X_3, \cdots, X_n$ i.i.d. with mean μ, sd σ

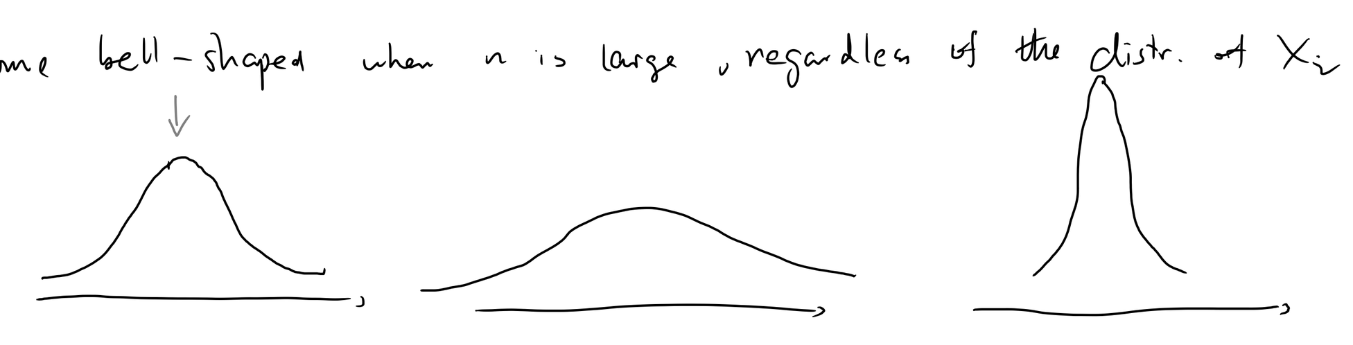
Let $S_n := \sum_{i=1}^{n} X_i$

We want to see the shape of the distr. of S_n

$\mathbb{E}S_n = n\mu$ $\quad Var(S_n) = n\,Var(X_n) = n\sigma^2$

$SD(S_n) = \sqrt{n}\,\sigma$

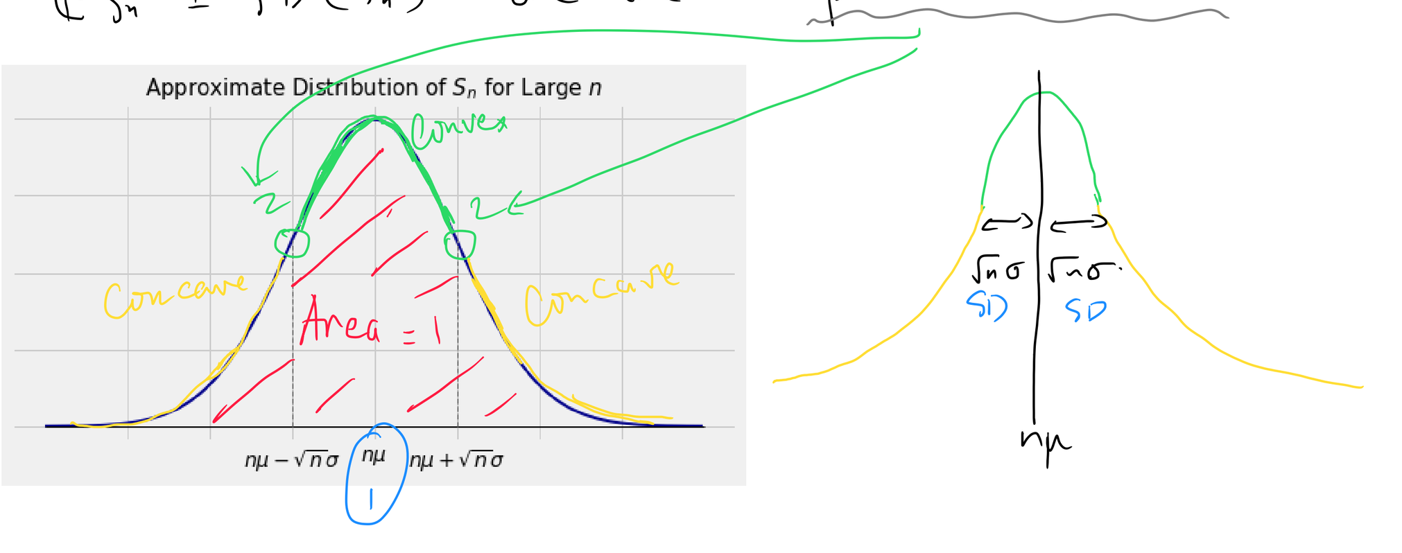
Roughly speaking, CLT tells us that S_n will eventually become bell-shaped when n is large, regardless of the distr. of X_i



To draw the graph

① the center of the bell is $\mathbb{E}S_n = n\mu$

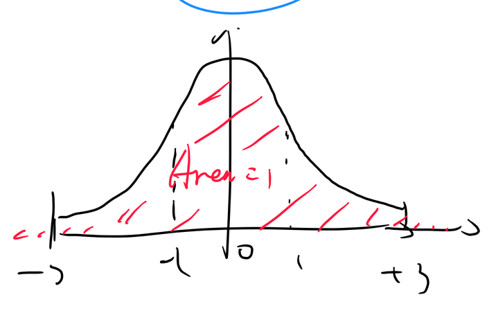
② $\mathbb{E}S_n \pm SD(S_n)$ are the two points of inflection



Sec. 8.2 the Standard Normal Curve

For different $n\mu$ & $\sqrt{n}\sigma$, you may result in different bell-shaped curves, but they are essentially the same in the sense that all curves can be derived by changing the origin and the units of measurement from the standard normal curve, where mean = 0 & sd = 1

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



Properties

① bell-shaped

② symmetric about 0 / even function

③ two points of inflection ± 1 — $\phi''(x) = 0$

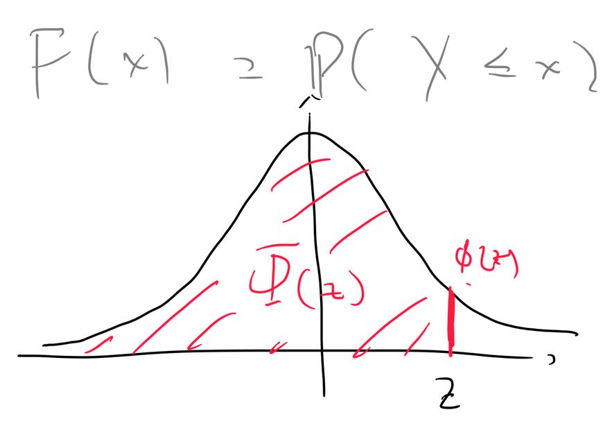
④ pretty close to 0 outside of $(-3, +3)$

⑤ a prob. distr. histogram (prob. density function) from next week.

the standard Normal CDF → $F(x) = P(X \le x)$

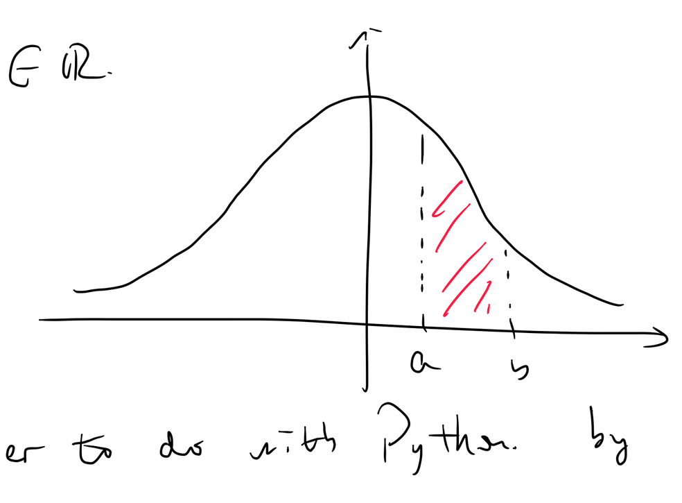
$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\,dx$$

"Phi" "phi"



There is no closed-form formula for $\Phi(z)$ for most $z \in \mathbb{R}$.

$P(\ldots) = \Phi(b) - \Phi(a)$



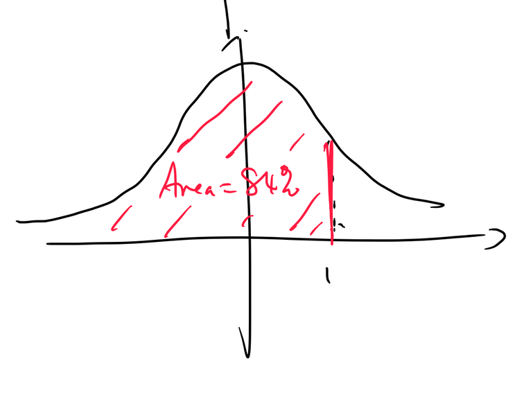
"Normal (Distr.) Table"

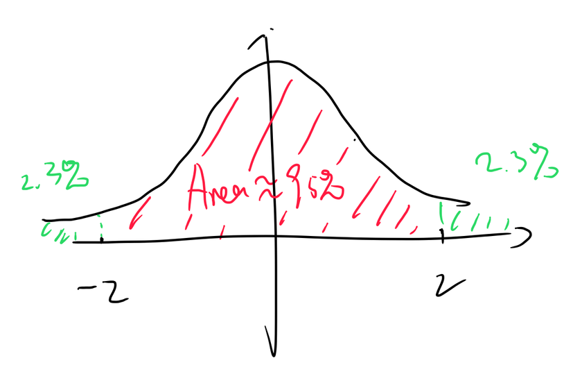
z	$\Phi(z)$
1	0.84
2	0.977

Easier to do with Python by

stats.norm.cdf(z)

$= \Phi(z)$



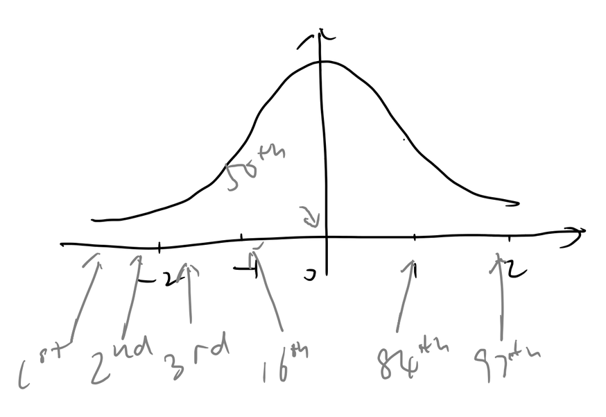


Percentiles

the 84th percentile is ≈ 1

the 98th percentile is ≈ 2

the p-th percentile is given by



$z = \Phi^{-1}(p\%)$

since $\Phi(z) = p\% \iff z = \Phi^{-1}(p\%)$

In python, use the percent point function (ppf)

stats.norm.ppf(p)

Sec 8.3 Normal Approximation

How could we use CLT and the normal curve to approx. a prob distr.?

First, we consider standard unit. Recall the restatement of Chebyshev's inequality.

For X

{ use $\mu = \mathbb{E}X$ as origin
{ use $\sigma = SD(X)$ as the unit distance

to create a new R.V. X^* — X in standard units

$$X^* = \frac{X - \mathbb{E}X}{SD(X)} = \frac{X - \mu}{\sigma} \iff X = \mu + \sigma X^*$$

$\mathbb{E}X^* = \frac{1}{\sigma}(\mathbb{E}(X) - \mathbb{E}(\mu)) = \frac{1}{\sigma}(\mu - \mu) = 0$

$SD(X^*) = \frac{1}{\sigma} SD(X - \mu) = \frac{1}{\sigma} SD(X) = \frac{1}{\sigma}\cdot\sigma = 1$

Example: X random weight in pounds (lbs)

$\mathbb{E}X = 150$

$SD(X) = 20$

$X = 170 = 150 + 20 \iff X^* = \frac{170 - 150}{20} = 1$

X is one SD to the right of mean

$X^* = -1.5 \iff X = 150 + 20 \times (-1.5) = 120$

↑ negative means smaller than expectation

Example 2.

$N = 10{,}000$ $\quad \mathbb{E}X = 3$ $\quad SD(X) = 2$

$n = 400$ $\quad P(S_{400} > 1260)$

Since n is large, use CLT

{ $\mathbb{E}S_{400} = n\,\mathbb{E}X = 1200$
{ $SD(S_{400}) = \sqrt{n}\,SD(X) = 20 \cdot 2 = 40$

$S^*_{400} = \frac{S_{400} - 1200}{40}$

Do the same on both sides

$P(S_{400} > 1260) = P\left(\frac{S_{400} - 1200}{40} > \frac{1260 - 1200}{40}\right)$

$= P(S^*_{400} > 1.5)$

$= 1 - P(S^*_{400} \le 1.5)$

$= 1 - \Phi(1.5) = \Phi(-1.5)$

by symmetry.

Example 3.

Suppose that the rent paid by a random group of 900 people $X_1, \cdots, X_{900}$ are i.i.d. each with expectation $1500 and sd $810

There is ≈ 90% chance that the avg rent of the 900 people is in the range of $1500 ± $ ______ ?

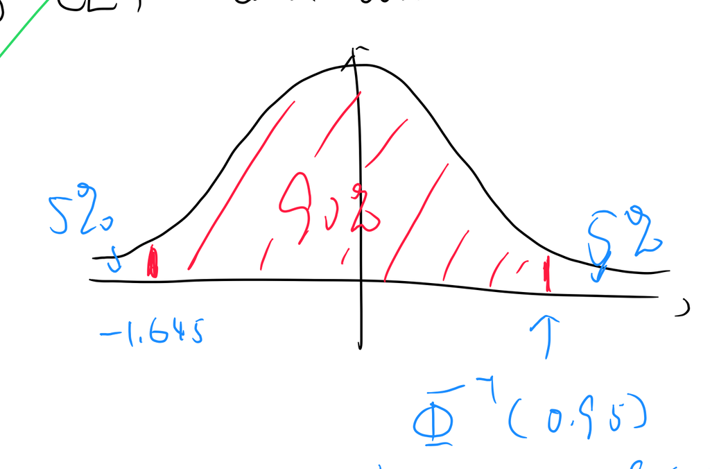
A_{900} := avg. rent (unit $)

$\mathbb{E}A_{900} = 1500$, $\quad SD(A_{900}) = \frac{1}{\sqrt{900}} \cdot SD(X) = \frac{1}{30} \cdot 810 = 27$.

$n = 90000$ $\quad SD(A_{90000}) = \frac{1}{\sqrt{90000}} SD(X) = \frac{1}{300} \cdot 810 = 2.7$

Since 900 is large, apply CLT and convert into standard unit measurement.

$$A^*_{900} = \frac{A_{900} - 1500}{27}$$



$\Phi^{-1}(0.95)$ ≈ 1.645

stats.norm.ppf(0.95)

± 1.645 SD

Answer: 1.645×27

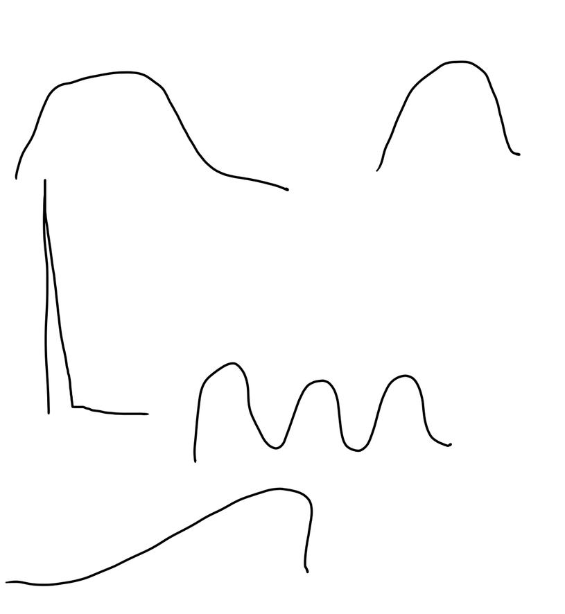
Sec. 8.4 How large is large?

- If the distr. is good { Smooth, uni-modal, balanced

⇒ not very large n needed. (30)

- If the distr. is bad { sharp, many curves, skewed

⇒ need larger n. (100 or 1000)



In this course: Consider good distr. only

calculate the mean and sd. and check if mean ± 3SD are within reasonable / possible region.

Example. $X_i \overset{i.i.d}{\sim} Bern\left(\frac{1}{100}\right)$ $\quad \mathbb{E}X_i = \frac{1}{100}$ $\quad SD(X) = \sqrt{\frac{1}{100} \cdot \frac{99}{100}} \approx \frac{1}{10}$.

$n = 100$

$S_n = X_1 + X_2 + \cdots + X_{100}$

$\mathbb{E}S_n = 1$ $\quad SD(S_n) \approx \sqrt{100} \cdot \frac{1}{10} = 1$

$1 \pm 3 = -2$ or 4

↑ -2 not reasonable ⇒ $n = 100$ not large enough.

$n = 10000$

$\mathbb{E}S_n = 100$ $\quad SD(S_n) = \sqrt{10000 \times \frac{1}{100}} = 10$

$100 \pm 3 \times 10 = 70$ or 130 ⇒ $n = 10000$ large enough.