* Announcement:
① HW8 due Tuesday (10/27)
② HW9 ~ Monday (11/02)
③ Quiz 7 : Ch 7

STAT 88: Lecture 26

Contents

Last time

Central Limit Theorem (CLT):

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $E(X_1) = \mu$ and $\text{SD}(X_1) = \sigma$. If $S_n = X_1 + \cdots + X_n$ is the sample sum, the for n large, the distribution of S_n is approximately normal (bell-shaped curve), *regardless* of the distribution of the X_i's.

Standard normal curve:

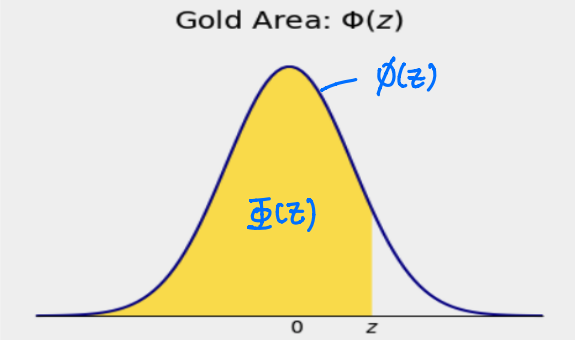
The standard normal curve is defined by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty,$$

and the standard normal CDF is

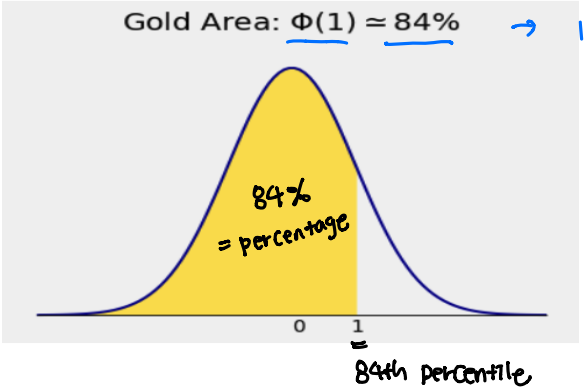
$$\Phi(z) = \int_{-\infty}^{z} \phi(x)dx.$$

Φ gives all the area under the curve ϕ to the left of z:



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

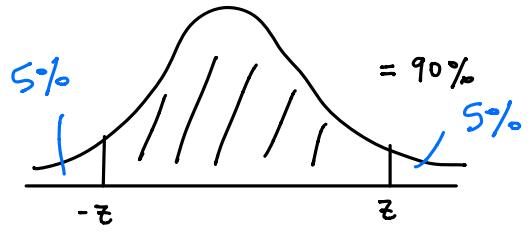
The percentage and percentile of the standard normal curve:



$\rightarrow 1 = \Phi^{-1}(0.84)$

Warm up: The standard normal curve is sketched below. Solve for z.

(a)



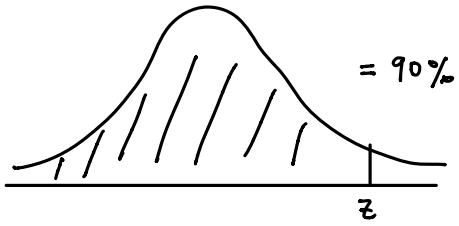
$\Phi(z) = 0.95$

$\rightarrow z = \Phi^{-1}(0.95)$

$=$ Stats.normal.ppf(0.95)

$= 1.645$

(b)

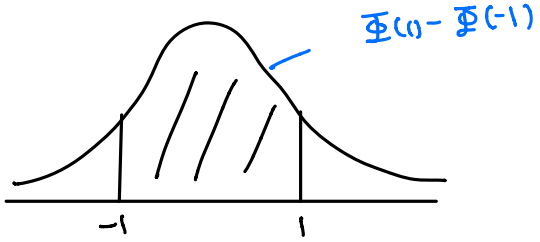


$\Phi(z) = 0.9$

$\rightarrow z = \Phi^{-1}(0.9)$

$= 1.281$

(c)



8.3. Normal Approximation

Standard Units Let X be any random variable with $E(X) = \mu$ and $\mathrm{SD}(X) = \sigma$. We define another random variable $X^\star$ such that

$$X^\star = \frac{X - \mu}{\sigma}.$$

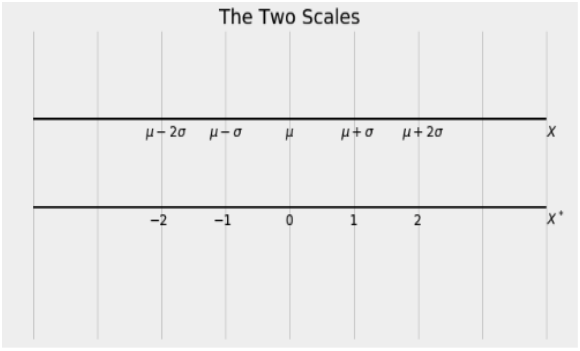
$\frac{kg}{kg}$ → $X = \sigma \cdot X^* + \mu$

Find $E(X^\star)$ and $\mathrm{SD}(X^\star)$.

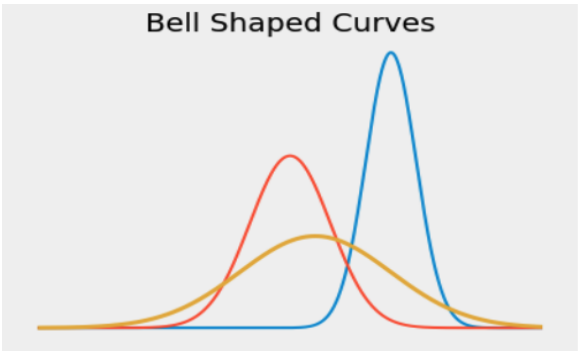
$E(X^*) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\{E(X) - \mu\} = 0$

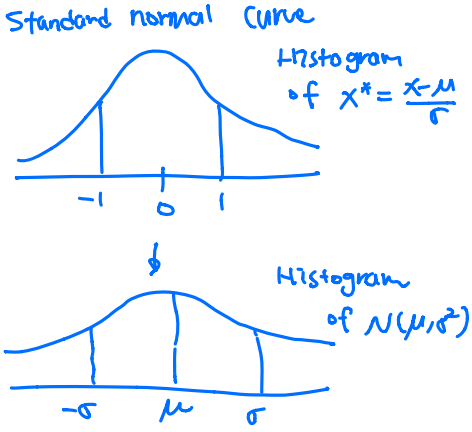
$SD(X^*) = SD\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma} SD(X) = 1$

The new random variable $X^\star$ is called X in standard units, or the standardized of X.



Normal Distributions The normal curve relates to the standard normal curve by changing the center and width of the bell, i.e. expectation and SD.





Let X be a random variable with normal distribution (bell-shaped, or Gaussian distribution) with $E(X) = \mu$ and $\mathrm{SD}(X) = \sigma$. We then write

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Applying the Central Limit Theorem We revisit the following exampe:

(Exercise 7.4.11) Each Data 8 student is asked to draw a random sample and estimate a parameter using a method that has chance 95% of resulting in a good estimate.

Suppose there are 1300 students in Data 8. Let X be the number of students who get a good estimate. Assume that all the students' samples are independent of each other.

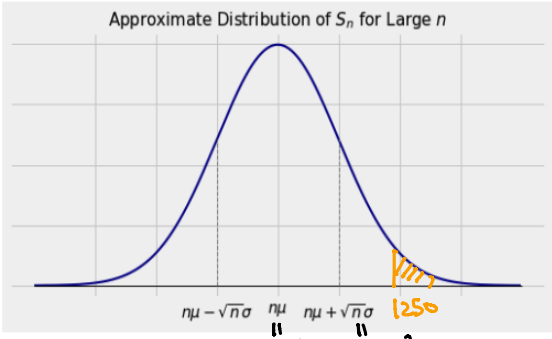
$X \sim$ Binomial (1300, 0.95) → sum of 1300 i.i.d. Ber RVs

Now we want to find $P(X > 1250)$, i.e. the chance that more than 1250 students get a good estimate.

① Apply Binomial formula
② Approximate using CLT

- The Central Limit Theorem says that the distribution of X is approximately normal with $E(X) = 1235$ and $\text{SD}(X) \approx 7.86$.

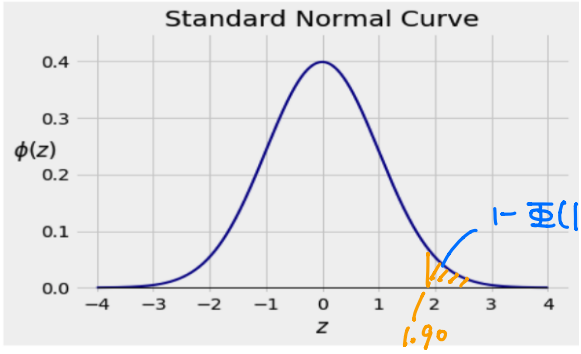
$= 1300 \cdot (0.95)$ $= \sqrt{1300 \cdot (0.95)(0.05)}$



- The standardized random variable is $Z = \frac{X - 1235}{7.86}$. To find $P(X > 1250)$, it's equivalent to finding

$$P(Z > \frac{1250 - 1235}{7.86}) = P(Z > 1.90).$$

$P\left(\frac{X-1235}{7.86} > \frac{1250-1235}{7.86}\right)$ = $P(Z > \cdot)$



$1 - \Phi(1.9) = 1 - \text{stats.norm.cdf}(1.9) \approx 0.029$

Example: (Exercise 8.5.5) Suppose the weights of sticks of butter are i.i.d. with a mean of 115 grams and an SD of 5 grams. Let X be the total weight of 600 such sticks. Find or approximate $P(X > 70000)$.

$X = X_1 + X_2 + \cdots + X_{600}$

$E(X_i) = 115$, $SD(X_i) = 5$

① Can we find $P(X > 70{,}000)$ exactly? No.

② By CLT, X is roughly normal distribution.

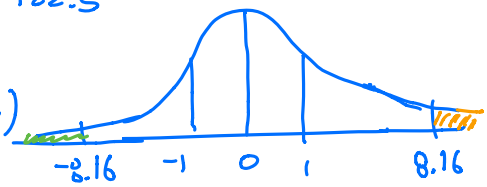
$E(X) = E(X_1 + \cdots + X_{600}) = 600 \cdot (115) = 69000$

$SD(X) = \sqrt{600} \cdot 5 = 122.5$

Now, $P(X > 70000) = P\left(\frac{X - 69000}{122.5} > \frac{70000 - 69000}{122.5}\right)$ ≈ 8.16

$= P(Z > 8.16)$

$= 1 - \Phi(8.16) = \Phi(-8.16)$



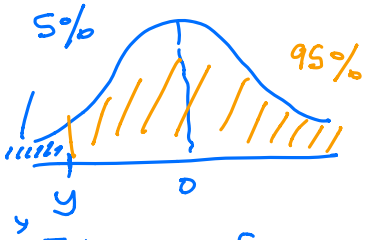
Find x such that $P(X > x)$ is approximately 95%.

$0.95 = P(X > x)$

$= P\left(\frac{X - 69000}{122.5} > \frac{x - 69000}{122.5}\right)$

$= P\left(Z > \underbrace{\frac{x - 69000}{122.5}}_{y}\right)$

$= P(Z > y)$



$\Phi(y) = 0.05$

$\Rightarrow y = \Phi^{-1}(0.05)$

$\Rightarrow \frac{x - 69000}{122.5} = \Phi^{-1}(0.05)$

$\Rightarrow x = 122.5 \cdot \Phi^{-1}(0.05) + 69000$

8.4. How Large is "Lage"?

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and SD σ, and let $S_n = X_1 + \cdots + X_n$. The Central Limit Theorem says that no matter what the distribution of X_1, after some large enough n, the distribution of S_n looks roughly normal.

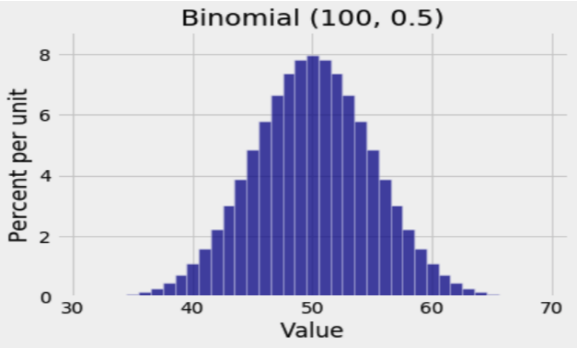
Question: How large n is "large enough"?

The answer depends on the distribution of X_1 and there is no universal way to determine the sample size.

- In general, if the distribution of X_1 is smooth and symmetric, the distribution of the sample sum can start looking normal even when the sample size n is moderate.
- If the distribution of X_1 s skewed or has gaps, then the sample size might have to get larger before the normal approximation is good.

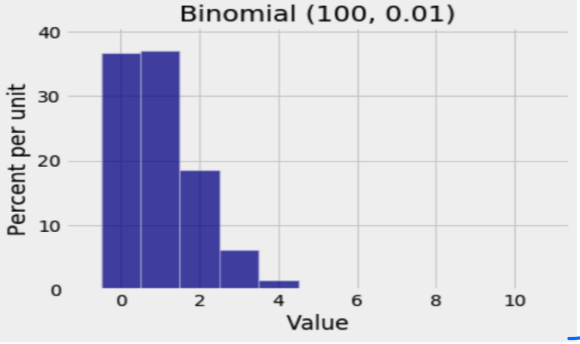
Approximations to the Binomial

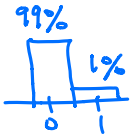
The histogram of Binomial$(100, 0.5)$:



Binomial (100, 0.5) = Sum of 100 i.i.d. Ber(0.5).
Each indicator has symmetric distribution

The histogram of Binomial$(100, 0.01)$:





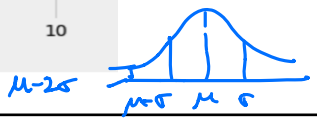
Binomial (100, 0.01) = Sum of 100 i.i.d. Ber(0.01).
Each indicator has asymmetric and skewed distribution

← This is closer to Pois(1) than $N(1,1)$

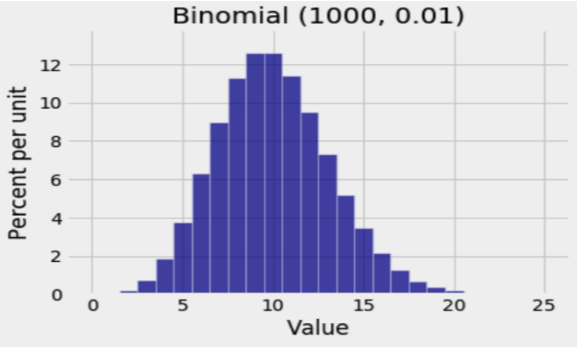
Binom $(100, 0.01) \rightarrow \mu = 100 \cdot (0.01) = 1$

$\sigma \approx 1$

$\mu - 2\sigma = -1$



When $n = 1000$, the histogram of Binomial$(1000, 0.01)$:



For a distribution to look normal, the possible values have to stretch for three or fours SDs on both sides of the mean.

Ways to Decide If n is Large Enough In the general case when the random variables being added are not indicators, there are no universal rules about how large the sample size has to be for the normal approximation to work. The textbook suggests two ways of making the decision but they are just some heuristics.

Note: The Central Limit Theorem has played an integral role in classical statistics in which there were no computers and people relied on analytical approximations to infer the distribution of statistic (e.g. sample sum or sample mean) and calculate the probabilities. Nowadays, with the recent advances in many computational tools, statistical inferences based on large sample theory are being replaced by numerical approximations such as Bootstraps or Monte Carlo simulations and the applications of CLT are being less used.