§4.3 Normal Distribution (The most important!)

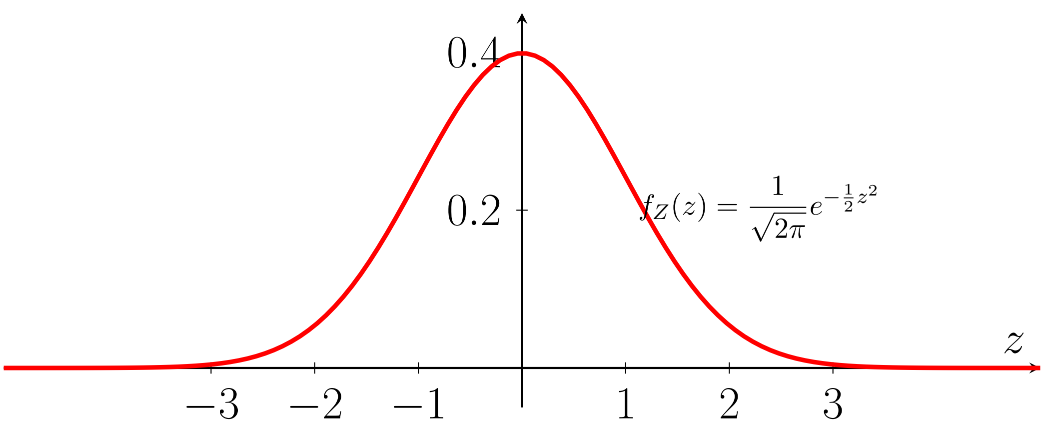
Definition.

The **standard normal distribution** is a continuous **pdf** defined by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

for $-\infty < z < \infty$.

The graph is **Gaussian** curve (bell curve).

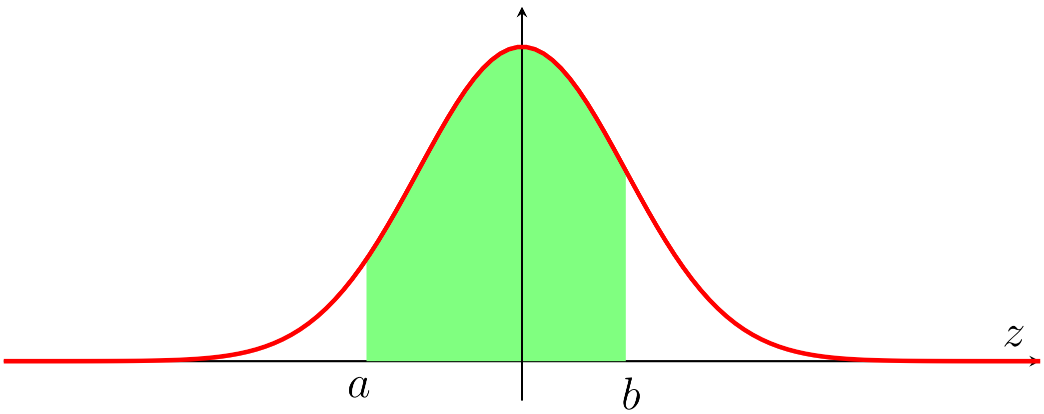


Theorem.

(1) It is a well defined **pdf**, i.e., $\int_{-\infty}^{\infty} f_Z(z) = 1$
(2) The mean is $E(Z) = \mu = 0$.
(3) The variance is $\mathrm{Var}(Z) = \sigma^2 = 1$.

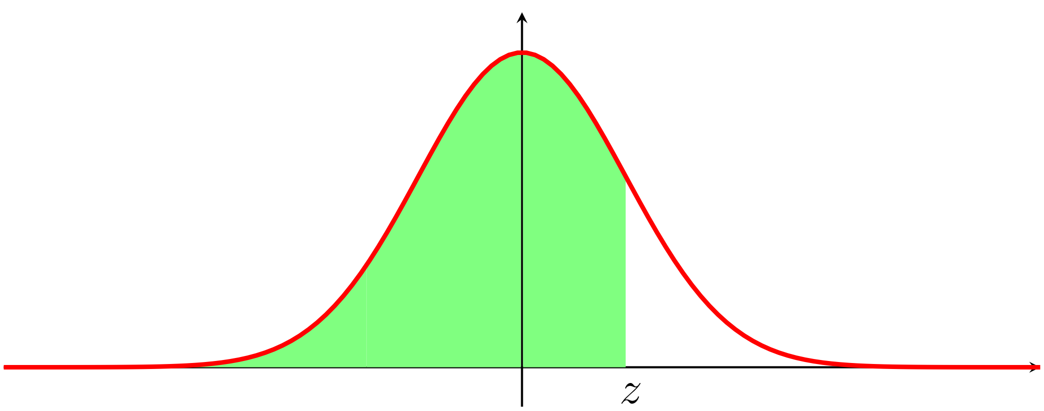
In theory, we know the **probability** of $a \le Z \le b$ is

$$P(a \le Z \le b) = \int_a^b f_Z(z)dz.$$



The **cdf** function is

$$F_Z(z) = \int_{-\infty}^{z} f_Z(u)du.$$



However, the function $f_Z(z) = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ has **no** elementary function as its antiderivative.

We can use calculator (TI-83/ TI-84 or higher version) to find the calculation of probability in exams.

(For homework, we can also use www.wolframalpha.com to check you answer. We can also the table in the Appendix **table** A.1 of the book.)

TI-83/TI-84: $\boxed{\text{2ND}} \rightarrow \boxed{\text{VARS}} \rightarrow \boxed{\text{2:normalcdf(}}$

Example 1. Let Z be the standard normal distribution.

(1) Find $P(Z \leq 1.31)$ or find $\int_{-\infty}^{1.31} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$

$P(Z \leq 1.31) =$ **normalcdf**$(-1000, 1.31, 0, 1) = 0.9049$
(Or use the table, look up cdf for $z = 1.31$ gives 0.9049)

(2) Find $P(Z \geq -0.45)$

By calculator $P(Z \geq -0.45) =$**normalcdf**(-0.45,1000,0,1)= 0.6736
(Or by table use $P(Z \geq -0.45) = 1 - P(Z \leq -0.45) =$)

(3) Find $P(-1 < Z < 1)$ or find $\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$

By calculator $P(-1 < Z < 1) =$ **normalcdf**$(-1, 1, 0, 1) = 0.6827$
(Or by table use $P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) =$)

(4) Find the 70th percentiles. (Find a such that $P(Z \leq a) = 0.7$)

Use calculator **invNorm**(0.7,0,1)= 0.5244. So, $a = 0.5244$
(Or use the table, find the closed number to 0.7, which is 0.6985 comes from z=0.52.)

(5) Find numbers a and b such that $P(a \leq Z \leq b) = 0.95$ (Hint: you can assume $a = -b$, then $P(Z \leq a) = (1 - 0.95)/2 = 0.025$)

Use calculator **invNorm**(0.025,0,1)= -1.96
So, $a = -1.96$ and $b = 1.96$.

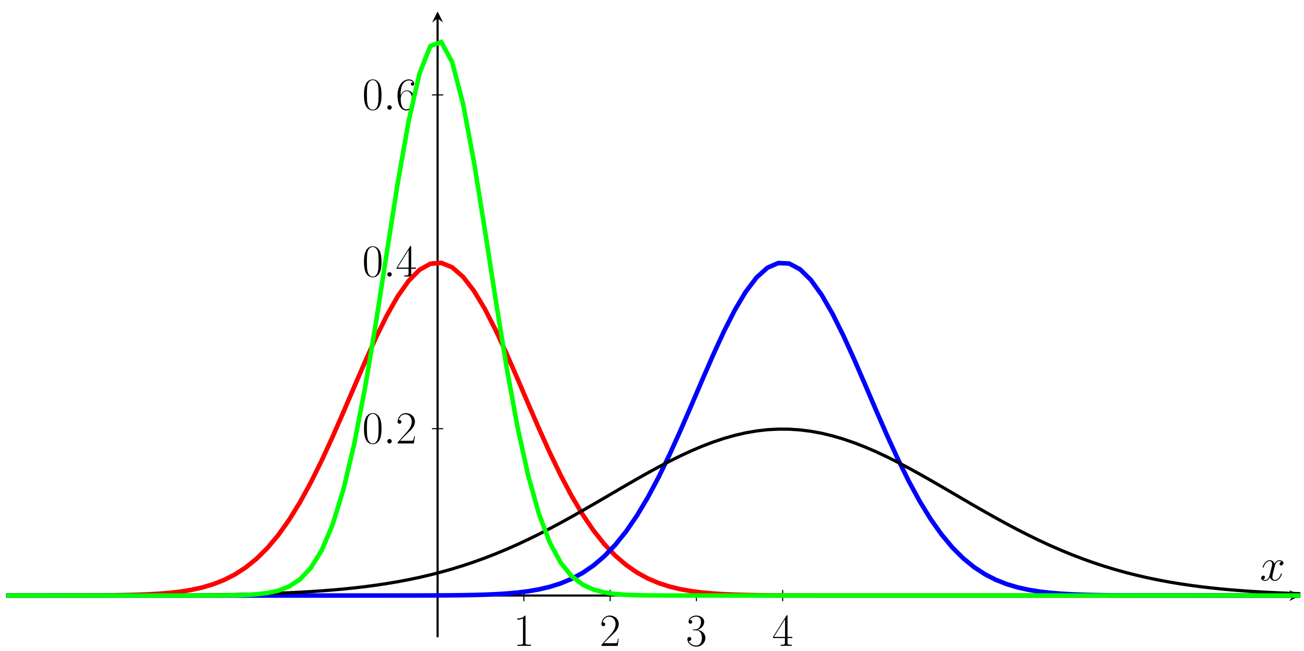
Definition. (Normal Distribution)

The **Normal Distribution** is a continuous **pdf** function defined as

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$

Theorem.

(1) It is a well defined **pdf** , i.e., $\int_{-\infty}^{\infty} f_X(x) = 1$
(2) The mean is $E(X) = \mu$.
(3) The variance is $\mathrm{Var}(X) = \sigma^2$.



Red: $\mu = \mathbf{0}, \sigma = \mathbf{1}$. **Green:** $\mu = \mathbf{0}, \sigma = \mathbf{0.6}$. **Blue:** $\mu = \mathbf{4}, \sigma = \mathbf{1}$. **Black:** $\mu = \mathbf{4}, \sigma = \mathbf{2}$.

There are two parameters in the definition. We usually denote

$$X \sim \mathrm{Normal}(\mu, \sigma^2)$$

and say that X is a random variable *normally distributed with mean μ and variance σ^2 (or standard deviation σ).*

Theorem.

The relationship between standard normal distribution $Z \sim \text{Normal}(0,1)$ and normal distribution $X \sim \text{Normal}(\mu, \sigma^2)$ is that

$$X = \mu + \sigma Z$$

or write it in another way

$$Z = \frac{X - \mu}{\sigma}$$

Example 2. Suppose the national Mathematics SAT scores is **normally** distributed with mean of 500 and a standard deviation 100. What percentage score between 400 and 600?

Solution:
Method 1. Use calculator

$P(400 \le X \le 600)$ =**normalcdf**(400,600,500,100)= 0.6827

Method 2.

$$\begin{aligned} P(400 \le X \le 600) &= P(400 \le 500 + 100Z \le 600) \\ &= P(-1 \le Z \le 1) \\ &= \textbf{normalcdf}(-1, 1, 0, 1) \\ &= 0.6827 \end{aligned}$$

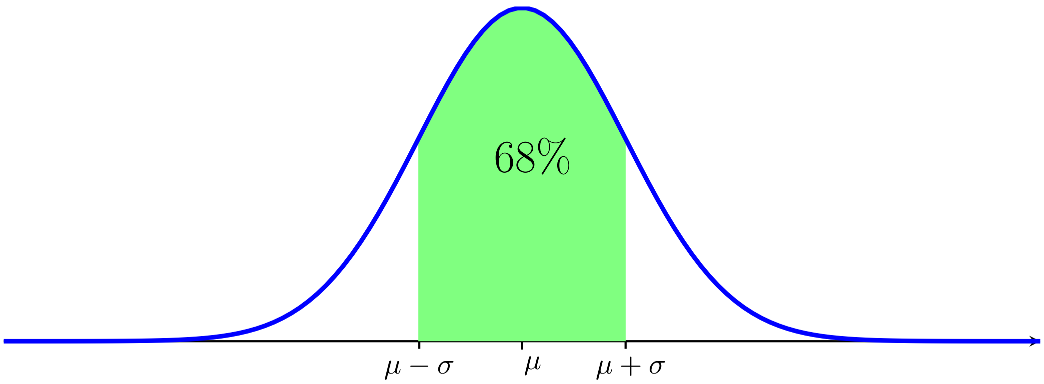
You can also try

$P(300 \le X \le 700) = 0.9545$

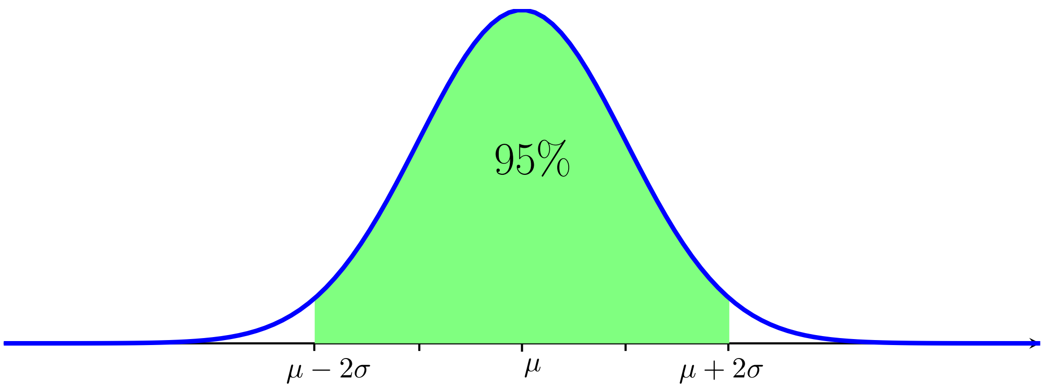
$P(200 \le X \le 800) = 0.9973$

68-95-99.7 rule in the normal distribution $X \sim \text{Normal}(\mu, \sigma^2)$.
(Do not memorize.)

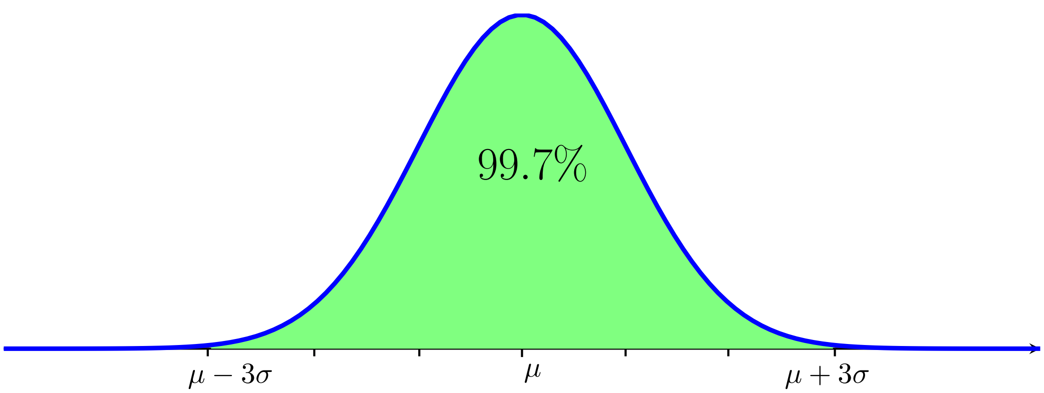
$P(\mu - \sigma \le X \le \mu + \sigma) \approx$ **68**.27%



$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx$ **95**.45%



$P(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx$ **99.7**3%



Applications: Central Limit Theorem!

Suppose random variables $X_1, X_2, \cdots, X_n$ are **independent** and **identically distributed**(**IID**) from **any** distribution,(i.e., $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.)

The **sample sum** $X = X_1 + X_2 + \cdots + X_n$.

The **sample mean** is $\overline{X} := \dfrac{X}{n} = \dfrac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

From §3.9, $E(X) = n\mu$; $\mathrm{Var}(X) = n\sigma^2$; $E(\overline{X}) = \mu$; $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

Theorem. (**Central Limit Theorem**) for sample mean

Under above assumption (IID) and when n is **large** enough,

$$\overline{X} \sim \mathrm{Normal}(\mu, \frac{\sigma^2}{n})$$

That is, $P\Big(a \le \overline{X} \le b\Big) = \textbf{normalcdf}(a, b, \mu, \frac{\sigma}{\sqrt{n}})$

Theorem. **CLT** as standard normal distribution

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{Normal}(0,1), \quad or, \quad \frac{X - n\mu}{\sqrt{n}\,\sigma} \sim \mathrm{Normal}(0,1)$$

Theorem. **CLT** for sample sum

$$X \sim \mathrm{Normal}(n\mu, n\sigma^2)$$

$$P\Big(a \le X \le b\Big) = \textbf{normalcdf}(a, b, n\mu, \sqrt{n}\sigma)$$

More precisely,

$$\lim_{n\to\infty} P\Big(a \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le b\Big) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Historically, normal distribution is used as an approximation for binomial distribution. Later, people found that it can approximate "everything"!

Proof of CLT?

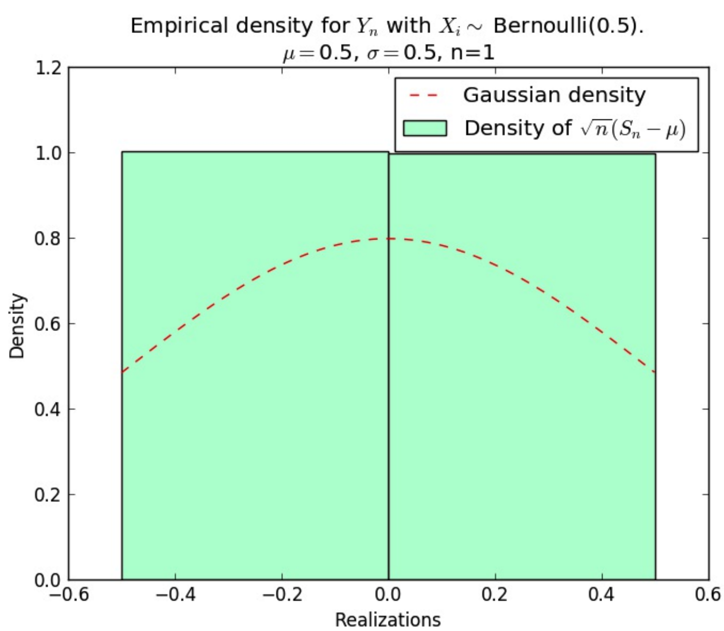
Simulations:

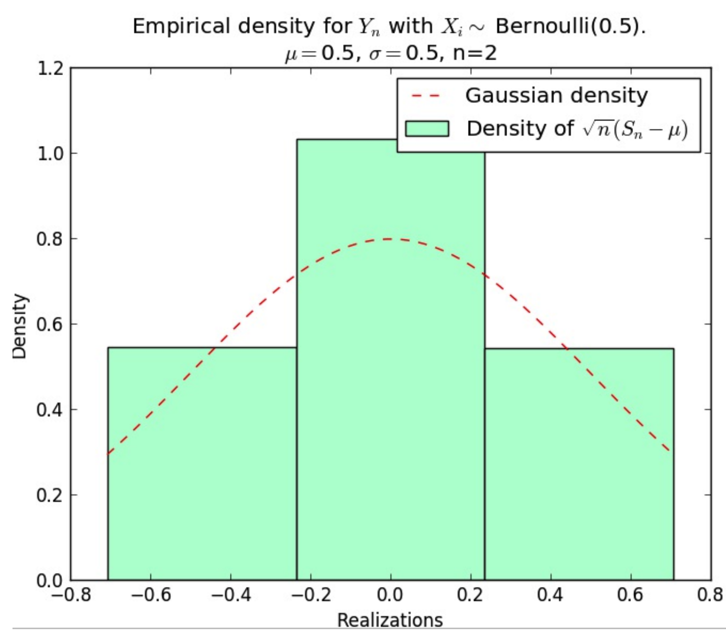
Suppose each X_i is the Bernoulli with $p = 0.5$. (Flip a fair coin)

Denote $S_n = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

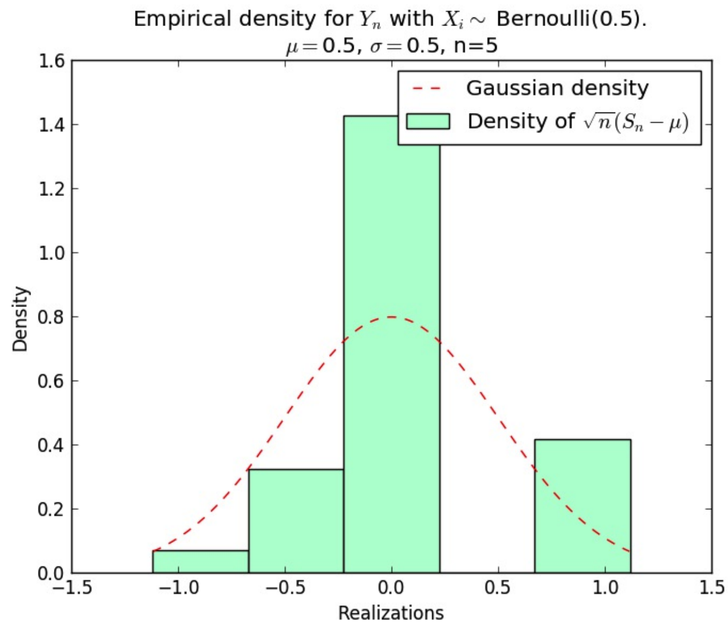
Make 1000 simulations for each case: $S_1, S_2, S_5, S_{10}, S_{50}, S_{100}$.

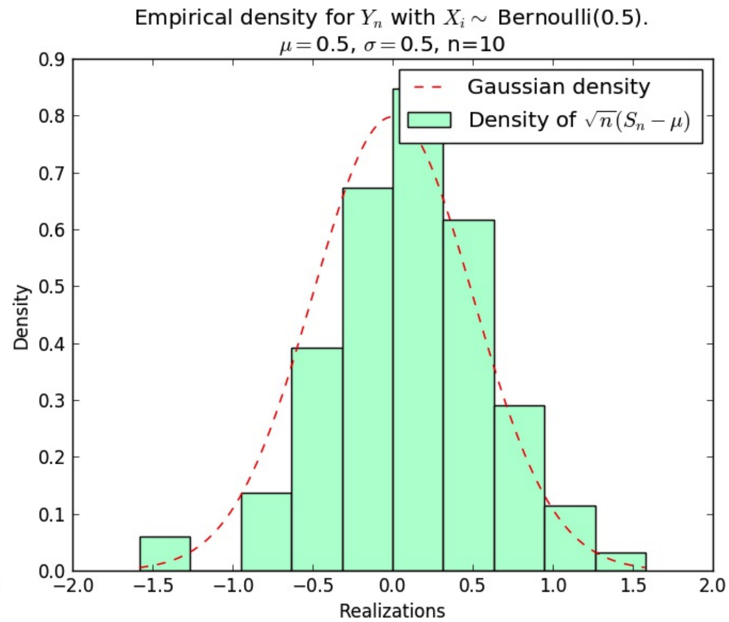
Examples: $S_1 = X_1$ and $S_2 = X_1 + X_2$



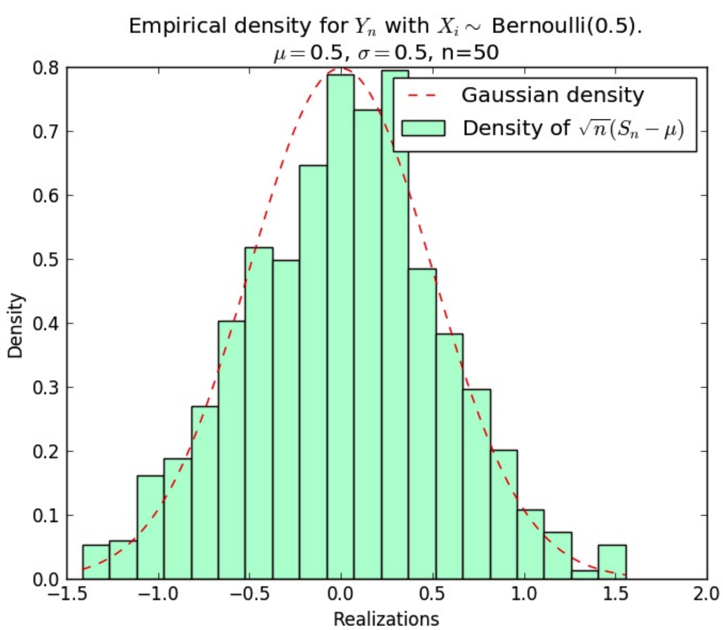


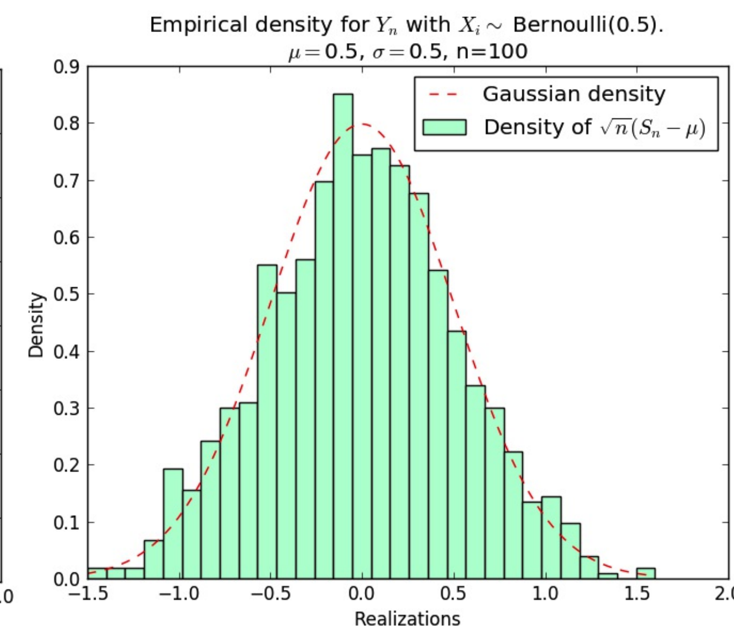
Examples: S_5 and S_{10}





Examples: S_{50} and S_{100}





http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html

http://simulations.lpsm.paris/tcl/

Example 3. Suppose we have a sample of 200 from a distribution with $\mu = 25$ and $\sigma = 16$. Find $P(\overline{X} > 26)$.

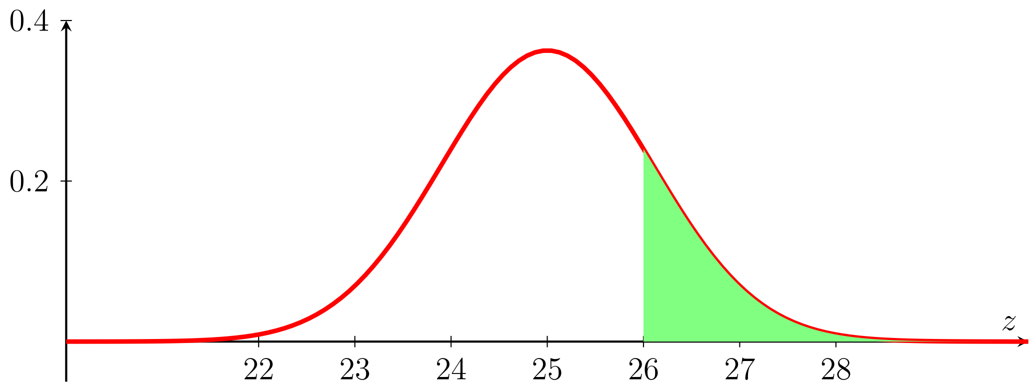
Solution: By CLT, we have

$$\overline{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n}) = \text{Normal}(25, \frac{16^2}{200})$$

Using Calculator,

$$P(\overline{X} > 26) \approx \textbf{normalcdf}(26, 1000, 25, 16/\sqrt{200}) \approx 0.188$$

Geometric meaning for the question:



Remark: In this example, $\overline{X}$ is the sample mean

$$\overline{X} = \frac{1}{100}(X_1 + X_2 + \cdots + X_{200})$$

Even we don't know what is the distribution of X_i. Only use the mean μ and standard deviation σ of X_i, we can use normal distribution to study the data.

Example 4. (Roulette Wheel) Let X be the amount won or lost in betting \$2 on **00** in roulette. The Payout is 35:1. Then $p_X(70) = 1/38$ and $p_X(-2) = 37/38$. If a gambler bets on **00** one hundred times, use the central limit theorem to estimate the probability that those wagers result in less than \$20 in losses.

Solution:
X: the amount won or lost in betting \$2 on **00** in roulette.
$E(X) = 35(2)(1/38) + (-2)(37/38) = -2/19 \approx -0.1053$
$E(X^2) = (70^2)(1/38) + (2^2)(37/38) = 2524/19 \approx 132.84$
$\sigma^2 = \mathrm{Var}(X) = E(X^2) - E(X)^2 = 132.83$
Denote $Y = X_1 + X_2 + \cdots + X_{100}$ which is the wagers result.
By CLT for sample sum, we have

$$Y \sim \text{Normal}(n\mu, n\sigma^2) = \text{Normal}(-10.53, 100(132.83))$$

Using Calculator,

$$P(Y > -20) \approx \textbf{normalcdf}(-20, 10000, -10.53, 10\sqrt{132.83}) \approx 0.5327$$

Remark: Try to find the precise result using binomial distribution. Let Z be the number of times the gambler win. We know $Z \sim \mathit{Binomial}(n = 100, p = 1/38)$. The wagers result $Y = 70Z - 2(100 - Z) = 72Z - 200$. The probability $P(Y > -20) = 1 - P(Y \leq -20) = 1 - P(Y \leq -20) = 1 - P(Z \leq 180/72) = 1 - \textbf{binomialcdf}(100, 1/38, 2) \approx 0.5084$

We use continuous random variable to estimate discrete random variable. We use a correction

$$P(Y > -20) = P(Y \geq -19) \approx \textbf{normalcdf}(-19.5, 10000, -10.53, 10\sqrt{132.83}) \approx 0.5310$$

We will see details in the following.

Estimate Binomial Distribution. Let X be a **binomial** random variable with parameters n and p. it is the sum of n independent Bernoulli variables $X_1, X_2, \ldots, X_n$,

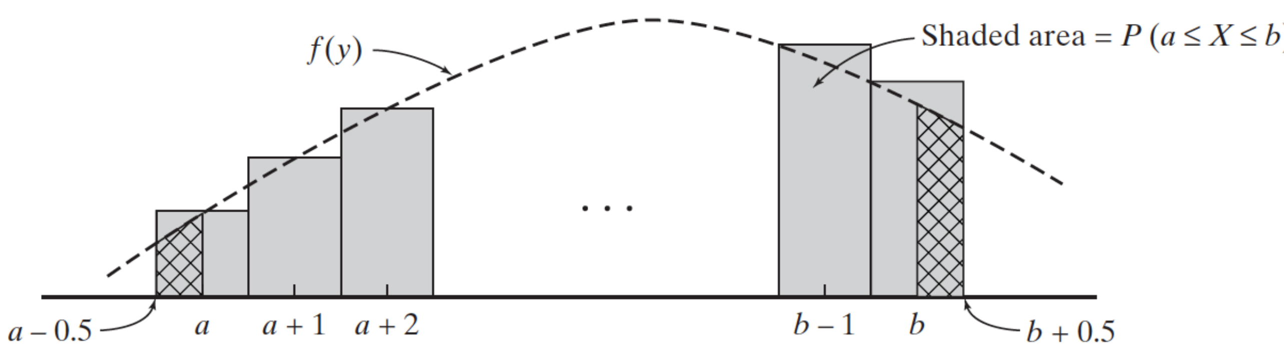
$$X = X_1 + X_2 + \cdots + X_n$$

We know that $E(X) = np$ and $\mathrm{Var}(X) = np(1-p)$.

So, by CLT for sample sum,

$$\boxed{X \sim \mathrm{Normal}(E(X), \mathrm{Var}(X)) = \mathrm{Normal}(np, np(1-p))}$$

One problem is that the binomial distribution is discrete. We need to use **The Continuity Correction** to fix the difference.



Theorem. CLT for Binomial with Continuity Correction

If X **binomial** random variable with parameters n and p.

$$P\Big(a \le X \le b\Big) \approx \textbf{normalcdf}(a - 0.5, b + 0.5, np, \sqrt{np(1-p)})$$

$$P\Big(X \le b\Big) \approx \textbf{normalcdf}(-10000, b + 0.5, np, \sqrt{np(1-p)})$$

$$P\Big(X \ge a\Big) \approx \textbf{normalcdf}(a - 0.5, 10000, np, \sqrt{np(1-p)})$$

Example 5. Roll a die 600 times, use the normal approximation with continuity correction to estimate the probability that the number of 6's is smaller than 100.

Solution:
Let X be the number of 6's.
The question is to find $P(X \leq 99)$.
By CLT for binomial with continuity correction, we have

$$P\Big(X \leq b\Big) \approx \textbf{normalcdf}(-10000, b+0.5, np, \sqrt{np(1-p)})$$

Here, $n = 600$ and $p = 1/6$.
So, $np = 100$ and $\sqrt{np(1-p)} = 10\sqrt{5/6}$
Using Calculator,

$$P(X \leq 99) \approx \textbf{normalcdf}(-10000, 99.5, 100, 10\sqrt{5/6}) \approx 0.4782$$

Example 6. About 37% people of in a country smoke(retired/ drink/ watch TV...). If you randomly choose 500 people from this country, use the normal approximation with continuity correction to estimate the probability that there are at least 200 people who smoke (retired/...).

Solution:
Let X be the number of people who are 30 or under in the sample.
The question is to find $P(X \geq 200)$.
By CLT for binomial with continuity correction, we have

$$P\Big(X \geq a\Big) \approx \textbf{normalcdf}(a-0.5, 10000, np, \sqrt{np(1-p)})$$

Here, $n = 500$ and $p = 0.37$.
So, $np = 185$ and $\sqrt{np(1-p)} = \sqrt{116.55}$
Using Calculator,

$$P(Y \geq 200) \approx \textbf{normalcdf}(199.5, 10000, 185, \sqrt{116.55}) \approx 0.0896$$

Compare with the precise answer for the above examples

$P(X \leq 99)$=**binomcdf**$(600, 1/6, 99) \approx 0.4830$

$P(Y \geq 200) = 1 - P(Y \leq 199)$=1-**binomcdf**$(500, 0.37, 199) \approx 0.0901$

More Examples:

Example 7. Standardized IQ tests are designed so that their scores have a **normal** distribution in the general population with a mean of 100 and the standard deviation 15. Randomly choose 1245 people, person A gets a score 125. How many people have a better score than this person A?

Hint: First find the probability that a person's score is higher than 125, P(score >125). Then multiply with the population 1245.

Solution:
Let X be the IQ test score of a person.
So, $X \sim \text{Normal}(\mu, \sigma^2)$ with $\mu = 100$ and $\sigma = 15$.
Using Calculator,

$$P(X > 125) \approx \textbf{normalcdf}(125, 10000, 100, 15) \approx 0.0478$$

$1245(0.0478) \approx 59.5$. So, the number of people with a better score than person A is about 59 or 60.

Example 8. (Practice) The exam score of all students (8 sections, 530 students) are recorded. Assuming the distribution of the score is normal with a mean of 83 and a standard deviation of 7.

(1) Let Y be the *average* score in section 1 (with 71 students). What is the probability that the average score will exceed 84?

Solution: By **CLT** for sample mean, $Y \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$ with $\mu = 83$, $\sigma = 7$ and **n=71**.
By calculator,

$$P(Y > 84) \approx \textbf{normalcdf}(84, 10000, 83, 7/\sqrt{71}) \approx 0.1143$$

(2) Randomly choose a student's score Y_i. What is the probability that the score will exceed 90?

Solution: From the assumption, $Y_i \sim \text{Normal}(\mu, \sigma^2)$ with $\mu = 83$ and $\sigma = 7$.
By calculator,

$$P(Y_i > 90) \approx \textbf{normalcdf}(90, 10000, 83, 7) \approx 0.1587$$

(3) What is the probability that more than 5 of the student's scores will exceed or equal 93 in section 1 (with 71 students)?

Solution:
Let X be the number students whose score is higher than 93. **Find** $P(X > 5)$.
This is a **binomial** distribution with $n = 71$ and p given by

$$p = P(Y_i \geq 93) \approx \textbf{normalcdf}(93, 10000, 83, 7) \approx 0.0766$$

Method 1: Direct computation by binomial

$P(X > 5) = 1 - P(X \leq 5) = 1-$**binomcdf**(71,0.0766,5)$\approx$ 1-0.5365=0.4635

Method 2: Poisson approximation ($\lambda = np = 71(0.0766) = 5.4386$)

$P(X > 5) = 1 - P(X \leq 5) = 1-$**poissoncdf**(5.4386)$\approx$ 1-0.5395=0.4605

Method 3: CLT with continuity correction ($np = 5.4386$, $\sqrt{np(1-p)} = \sqrt{71(0.0766)(1-0.0766)} \approx 2.241$)

$P(X > 5) = \textbf{normalcdf}(4.5, 1000, 5.4386, 2.241) \approx 0.4891$

Example 9. Suppose the random variable Y (for example, the weight, or IQ, or ... of a group of people) can be described by a normal curve with that 95 and 143 are equidistant from the average μ. For what value of σ is

$$P(95 \leq Y \leq 143) = 0.8$$

Solution:
The mean $\mu = \frac{95+143}{2} = 119$.
For standard normal variable Z satisfies $Y = \mu + \sigma Z$. So,

$$P(95 \leq \mu + \sigma Z \leq 143) = 0.8$$

Hence,

$$P(\frac{95-119}{\sigma} \leq Z \leq \frac{143-119}{\sigma}) = 0.8$$

By calculator **invNorm**$(0.1, 0, 1) = -1.2815$. So,

$$P(-1.2815 \leq Z \leq 1.2815) = 0.8$$

So, $\dfrac{143-119}{\sigma} = 1.2815$. Then, $\sigma = 18.728$

Example 10. Let $X_1, X_2, \cdots, X_{100}$ be the independent random variables with **uniform** distribution in $[0, 2]$. Let $\overline{X}$ be the sample mean of $X_1, X_2, \cdots, X_{100}$.

(1). Find the **pdf** and **cdf** of X_i and draw the graph.

See Example 7. in §3.6 for a more general calculation. $f_{X_i}(x) = 1/2$ and $F_{X_i} = x/2$ for $x \in [0, 2]$

(2). Find the mean, variance and standard deviation of X_i.

See Example 7. in §3.6.
$E(X_i) = 1$ and $\text{Var}(X_i) = 1/3$

(3). Find $P(0.5 \leq X_1 \leq 1)$.

$P(0.5 \leq X_1 \leq 1) = F_{X_1}(1) - F_{X_1}(0.5) = 1/4$

(4). Find $P(0.5 \leq \overline{X} \leq 1)$.

$\overline{X} \sim \text{Normal}(1, 1/300)$ So, $P(0.5 \leq \overline{X} \leq 1) = \textbf{normalcdf}(0.5, 1, 1, \sqrt{1/300}) \approx 0.5$.

Summary of Normal Distribution

(I) Normal Distribution

1. Standard Normal Distribution, $Z \sim \text{Normal}(0, 1)$.

2. Normal Distribution $X \sim \text{Normal}(\mu, \sigma)$ and $X = \mu + \sigma Z$.

3. Graph, mean, variance of normal distribution.

4. Calculate probability.

5. Calculate standard deviation σ from probability.

(II). **Central Limit Theorem (CLT).** Sample sum or Sample mean of **any** distribution admits Normal Distribution.

1. CLT of sample mean and sample sum.

2. CLT for binomial with continuity correction.

Application questions.