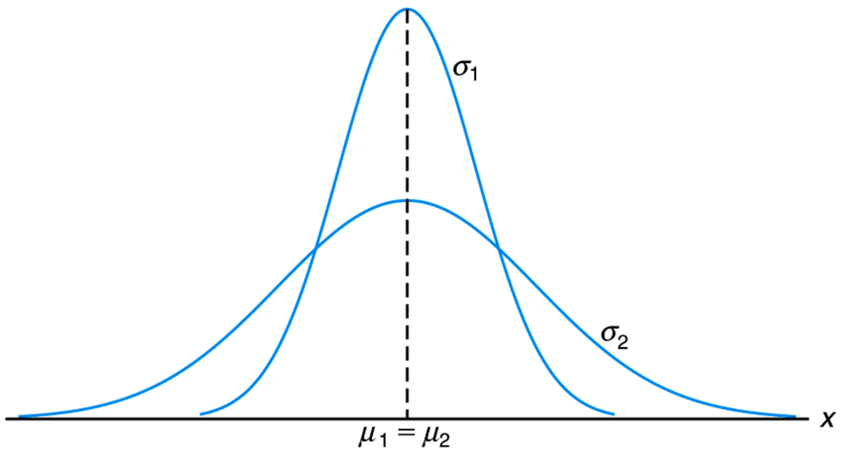
Chapter 5. Normal Distribution

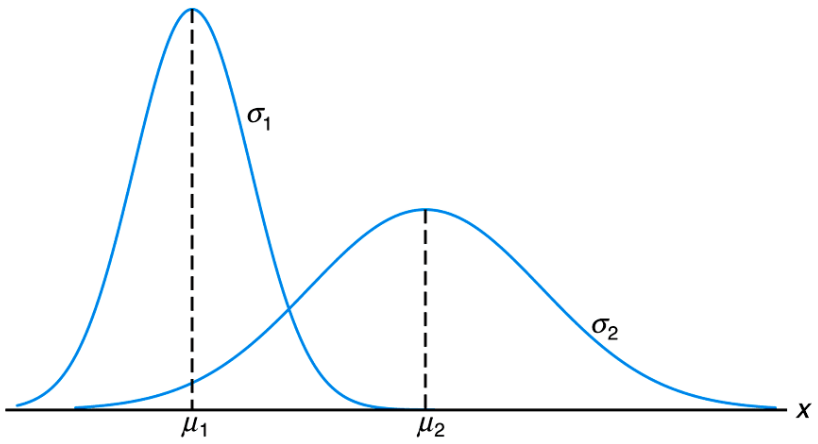
5.1 Probability Calculation Using the Normal Distribution

5.1.1 Definition of the Normal Distribution

- Normal Distribution, $N(\mu, \sigma)$

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





- The Normal distribution is also called Gaussian distribution in honor of Johann Carl F. Gauss (1777-1855).

Theorem 5.a

The mean and variance of $n(x;\mu,\sigma)$ are μ and σ^2, respectively. Hence, the standard deviation is σ.

- $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is symmetric about μ.
- $f(x) = f(2\mu - x)$
- $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\mu} xf(x)dx + \int_{\mu}^{\infty} xf(x)dx$

$$\Downarrow \ \int_{-\infty}^{\mu} xf(x)dx = \int_{-\infty}^{\mu} xf(2\mu - x)dx = \int_{\mu}^{\infty} (2\mu - y)f(y)dy$$

$$= \mu$$

- $E\left(\frac{(X-\mu)^2}{\sigma^2}\right) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma^2} f(x)dx = \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 1$

$\Uparrow \;\; y = \frac{x-\mu}{\sigma} \qquad\qquad \Uparrow$ integration by parts

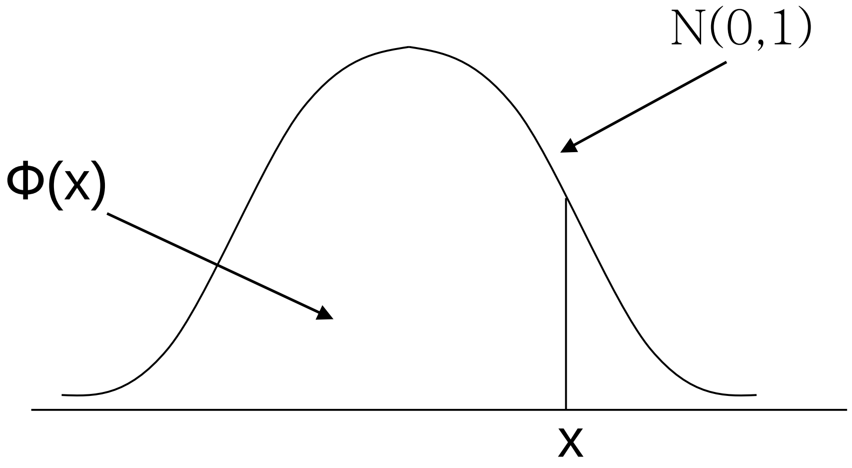
So, $E\left((X-\mu)^2\right) = \sigma^2$

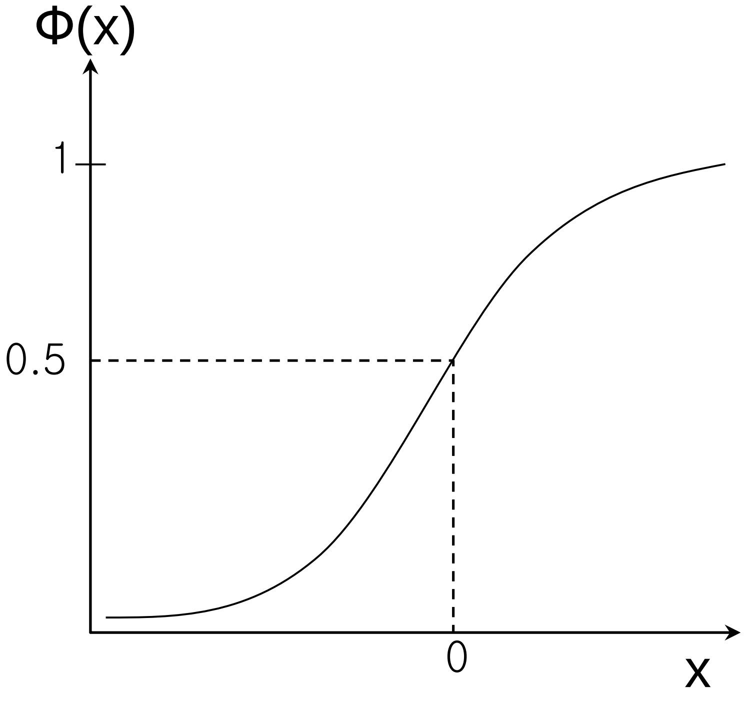
5.1.2 The Standard Normal Distribution

- Standard Normal Distribution, $N(0,1)$

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

Its cdf: $\Phi(\mathrm{x})$





5.1.3 Probability Calculation for General Normal Distributions

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

$$P(\mu - c\sigma \le X \le \mu + c\sigma) = P(-c \le Z \le c)$$

$$\Rightarrow \left\{\begin{array}{l} P(\mu-\sigma \le X \le \mu+\sigma) = P(|Z| \le 1) \simeq 0.68 \\ P(\mu-2\sigma \le X \le \mu+2\sigma) = P(|Z| \le 2) \simeq 0.95 \\ P(\mu-3\sigma \le X \le \mu+3\sigma) = P(|Z| \le 3) \simeq 0.997 \end{array}\right\}$$

$$P(X \le \mu + \sigma z_\alpha) = P(Z \le z_\alpha) = 1 - \alpha$$

5.1.4 Examples of the Normal Distributions

Example 18: Tomato Plant Heights

heights of tomato plants : mean=29.4cm, standard deviation=2.1cm

(1) Under the normal assumption, the interval of X with 1-α probability:

$$[\mu - \sigma z_{\alpha/2},\ \mu + \sigma z_{\alpha/2}] = [29.4 - 2.1 z_{\alpha/2}, 29.4 + 2.1 z_{\alpha/2}]$$

Therefore, the interval of X with 90% coverage is [25.95, 32.85] using $z_{\alpha/2}=z_{0.05}=1.645$

(2) Probability that a height is between 29cm and 30cm is

$$P(29.0 \le X \le 30.0) = \Phi\left(\frac{30.0-29.4}{2.1}\right) - \Phi\left(\frac{29.0-29.4}{2.1}\right)$$
$$= \Phi(0.29) - \Phi(-0.19) = 0.19$$

5.2 Linear Combinations of Normal Random Variables

5.2.1 The Distribution of Linear Combinations of Normal Random Variables

Linear Functions of a Normal Random Variable

$$X \sim N(\mu, \sigma^2)$$

$$\Rightarrow Y = aX + b \sim N(a\mu + b, a^2\sigma^2) \; \textit{for constant } a, b$$

$$X_1 \sim N(\mu_1, \sigma_1^{\,2}) \;\; \textit{and} \;\; X_2 \sim N(\mu_2, \sigma_2^{\,2}) \;\; \textit{are independent}$$

$$\Rightarrow Y = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^{\,2} + \sigma_2^{\,2})$$

Proof of $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$Y = X_1 + X_2$.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1)dx_1 = \int_{-\infty}^{\infty} f_1(x_1)f_2(y - x_1)\, dx_1$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}) \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-\frac{(y-x_1-\mu_2)^2}{2\sigma_2^2})\, dx_1$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}) \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-\frac{(y-x_1-\mu_2)^2}{2\sigma_2^2})\, dx_1$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp(-\frac{\left(y-(\mu_1+\mu_2)\right)^2}{2\left(\sigma_1^2+\sigma_2^2\right)})$$

Properties of independent Normal Random Variables

$$X_i \sim N(\mu_i, \sigma_i^2),\ 1 \le i \le n \text{ are independent}$$

$$a_i,\ 1 \le i \le n, \text{ and } b \text{ are constants}$$

$$\Rightarrow Y = a_1 X_1 + \cdots + a_n X_n + b \sim N(\mu, \sigma^2)$$

$$\text{where } \mu = a_1\mu_1 + \cdots + a_n\mu_n + b,\ \sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$$

$$X_i \sim N(\mu, \sigma^2),\ 1 \le i \le n \text{ are independent}$$

$$\Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

[redacted]

Example 23: Piston Head Construction

$X_1 \sim N(30.00, 0.05^2)$,: radius of piston

$X_2 \sim N(30.25, 0.06^2)$: radius of cylinder

(1) Distribution of $Y = X_2 - X_1$

$Y \sim N(30.25-30.00, 0.05^2 + 0.06^2) = N(0.25, 0.0061)$

(2) Probability that a piston head will not fit within a cylinder

$$P(Y < 0) = P\left(\frac{Y - 0.25}{\sqrt{0.0061}} < -\frac{0.25}{\sqrt{0.0061}}\right) = \Phi(-3.2) = 0.0007.$$

(3) Probability that Y is between 0.10mm and 0.35mm

$$P(0.1 \leq Y \leq 0.35) = \Phi(1.28) - \Phi(-1.92) = 0.8723.$$

Example 37: Concrete Block Weights

(1) $X_1, \cdots, X_{24}$ iid with $\mathrm{N}(11, 0.3^2)$.

$$\mathrm{Y} = X_1 + \cdots + X_{24} \sim N(24 \times 11, 24 \times 0.3^2) = N(264, 2.16)$$

(2) Find an interval of Y with 99.7% coverage:

$$[\mu - \sigma z_{\alpha/2}, \quad \mu + \sigma z_{\alpha/2}] = 264 \pm 1.47 \times 3$$

$$= [259.59, \ 268.41]$$

Example 38: Chemical Concentration Level C is measured in two methods.

- $X_A \sim N(C, 2.97)$: Method A
  $X_B \sim N(C, 1.62)$: Method B
- 99.7% coverage intervals:

  [C-5.17, C+5.17] : Method A

  [C-3.82, C+3.82] : Method B

- Combine the two measurements, $\mathrm{X_A}$ and $\mathrm{X_B}$, to

$$Y = pX_A + (1 - p)X_B$$

  so that it can minimize the variability.

$Y = pX_A + (1-p)X_B$

$Y \sim N(\mu_Y, \sigma_Y^2)$ where

$\mu_Y = C$ and $\sigma_Y^2 = p^2\sigma_A^2 + (1-p)^2\sigma_B^2$.

The minimum of σ_Y^2 is attained

when $\mathrm{p} = \frac{\sigma_B^2}{\sigma_A^2+\sigma_B^2} = 0.35$ with the minimum $\sigma_Y^2 = 1.05$.

The interval of Y with 99.7% coverage is

[C-3.07, C+3.07]

5.3 Approximating Distributions with the Normal Distribution

5.3.1 The Normal Approximation to the Binomial Distribution

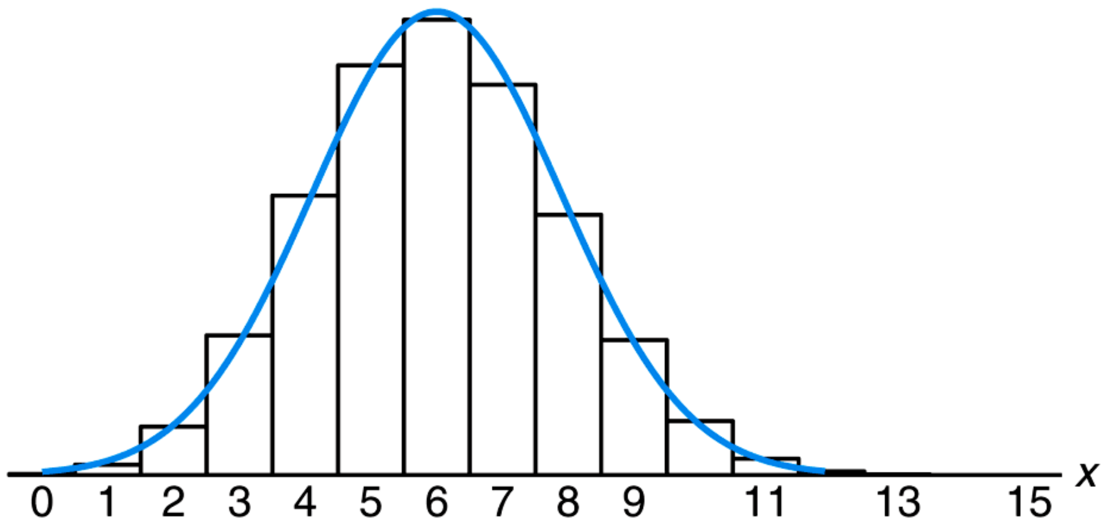
Theorem 5.3.a

If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{npq}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Figure 5.3.a Normal approximation of *B*(15,0.4)

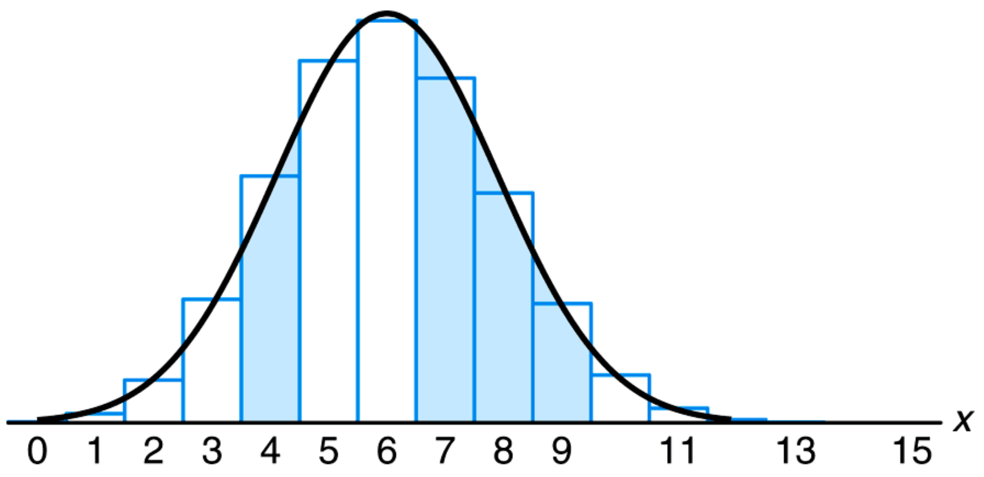


- Continuity correction in the Normal approximation

$X \sim B(n, p)$. $Z \sim N(0,1)$.

$$P(X \le x) \approx P\left(Z \le \frac{x+0.5-np}{\sqrt{npq}}\right).$$

$$P(X \ge x) \approx P\left(Z \ge \frac{x-0.5-np}{\sqrt{npq}}\right)$$



Python codes

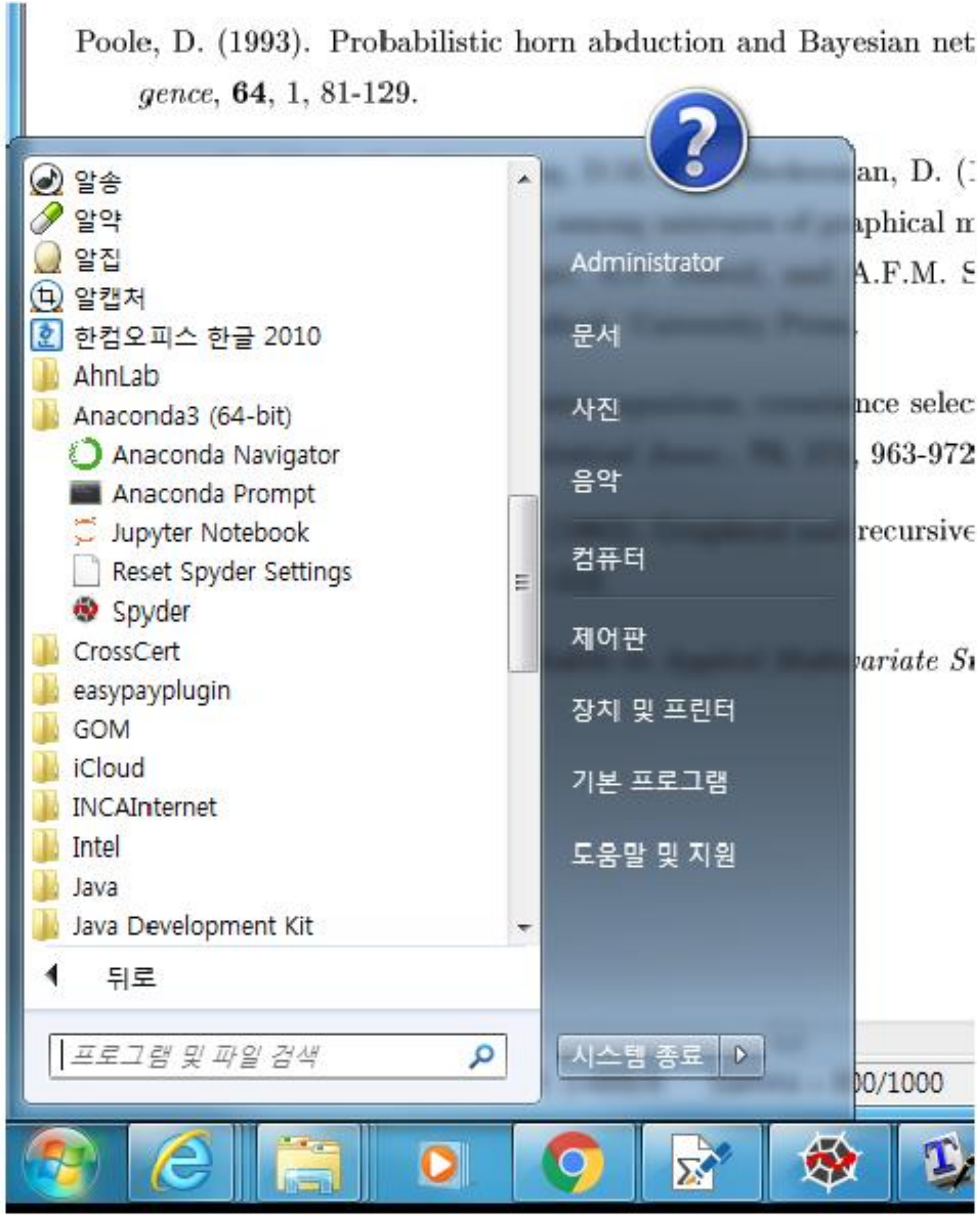
- Installing

How to install Python(Anaconda distribution):

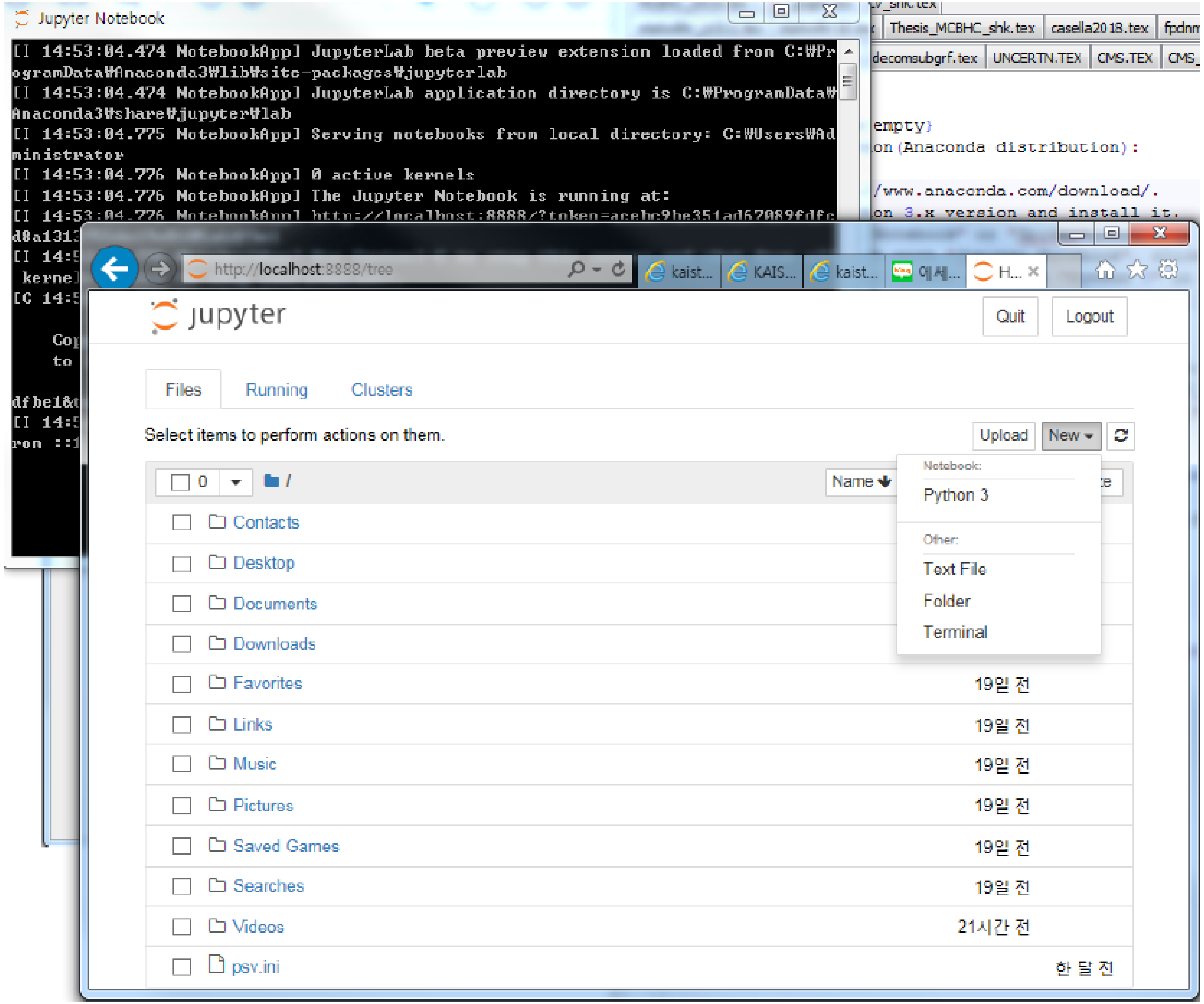
1. Enter https://www.anaconda.com/download/.
2. Download Python 3.x version and install it.
3. Run "Jupyter Notebook" or "Spyder".
4. To begin a job with "Jupyter Notebook", click "New" on the menu at the top-right corner and then "Python 3" in working directory.

Python codes

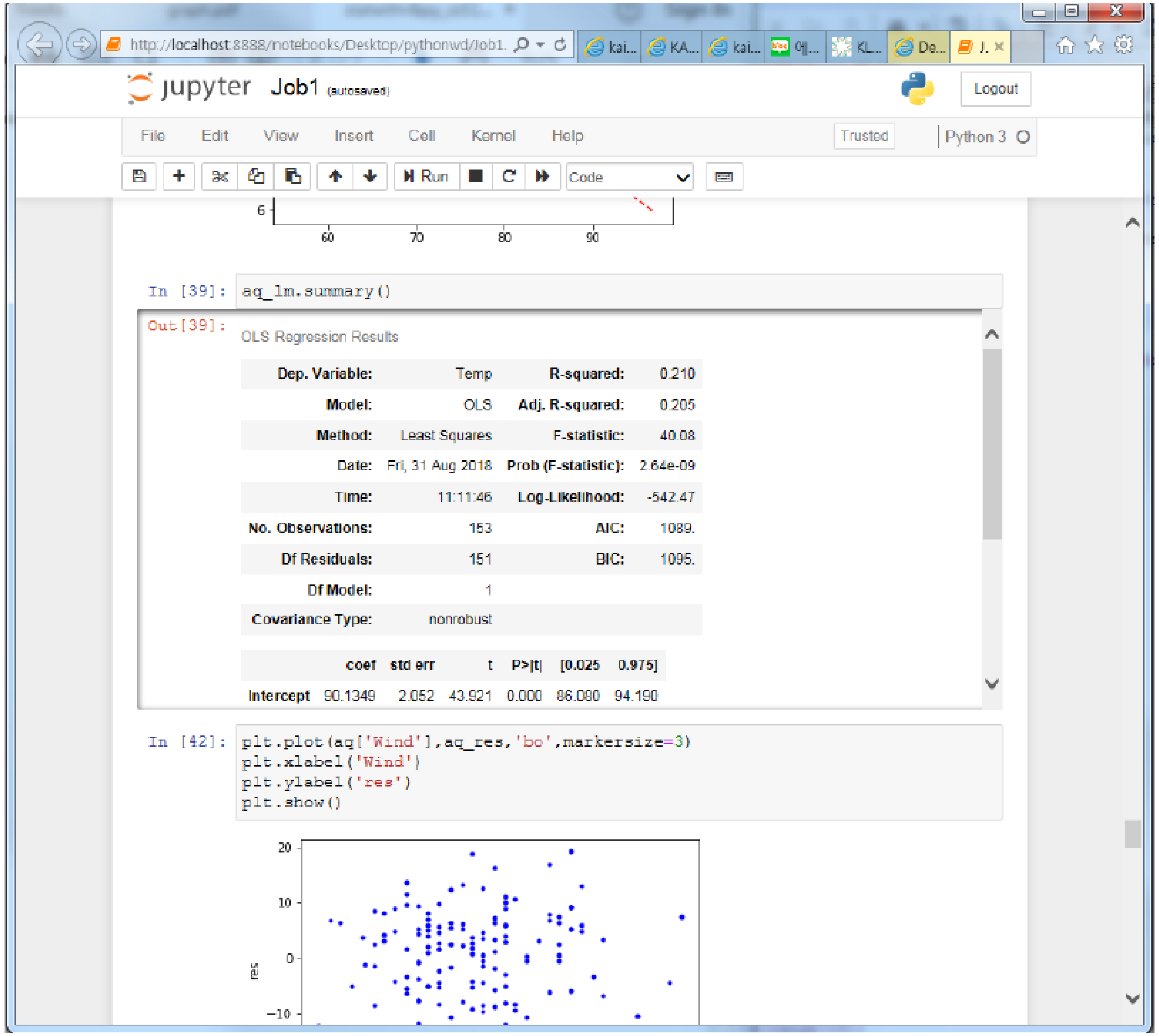
- After installing



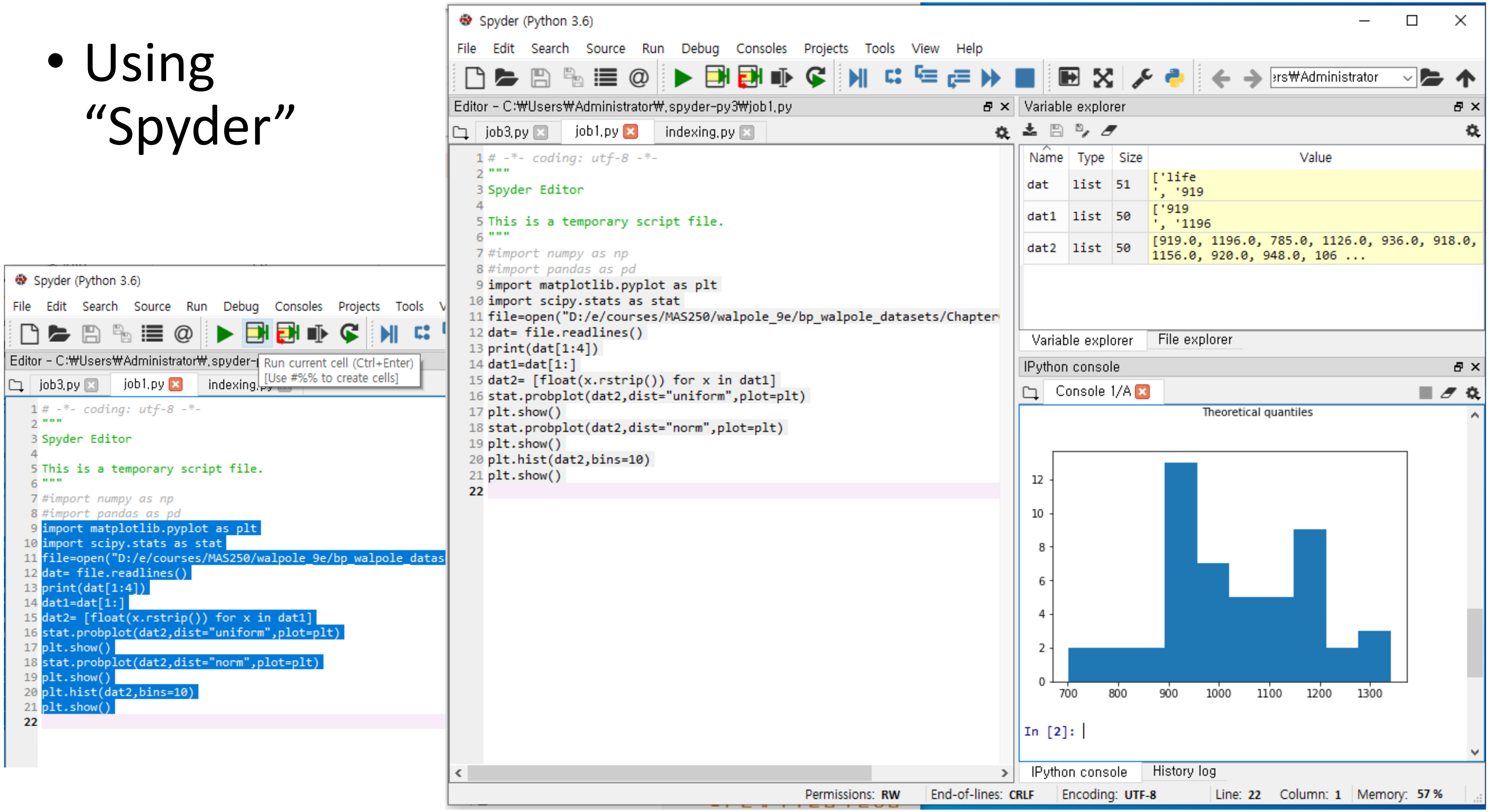
“Jupiter Notebook” clicked



- The working sheet of “Jupiter Notebook”.

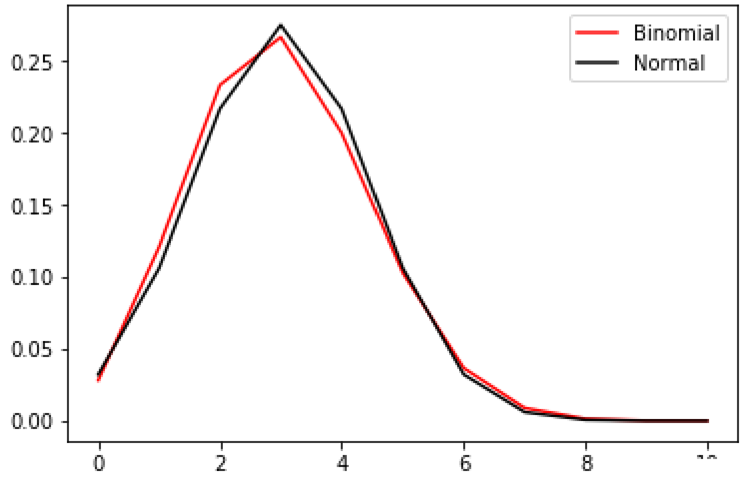


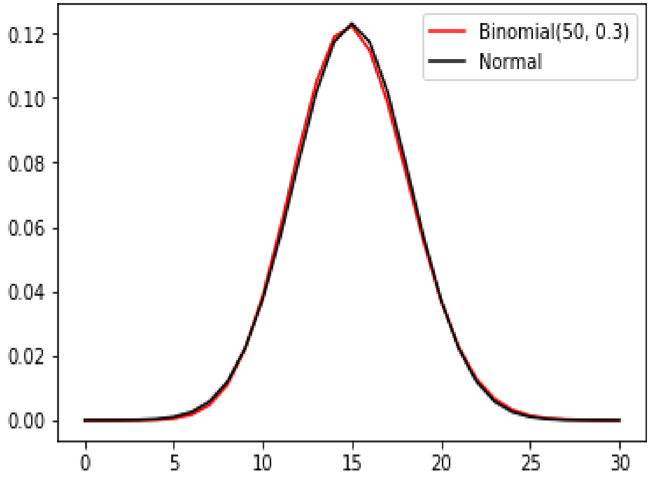
- Using "Spyder"

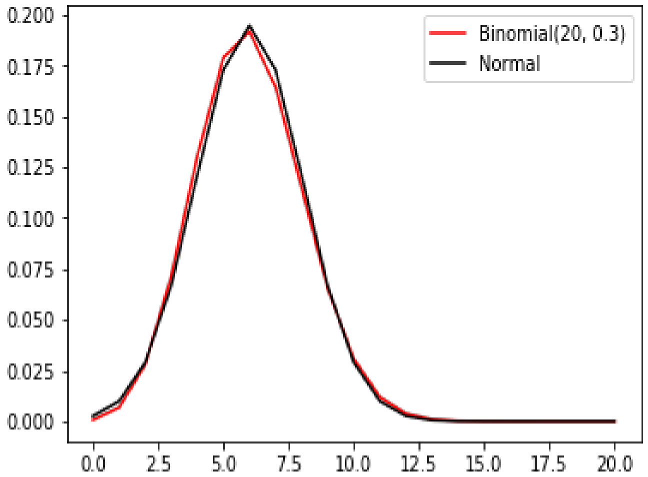


Comparison of the probability functions of Binomial and Normal distributions

$n = 10, \qquad p = 0.3$







Python codes for the distribution curves of Binomial and Normal distributions

- import numpy as np
- import matplotlib.pyplot as plt
- import scipy.stats as stat
- x=np.arange(0,31,1)
- n=50
- fig,ax=plt.subplots()
- ax.plot(x,stat.binom.pmf(x,n,0.3),'r',label='Binomial(50, 0.3)')
- ax.plot(x,stat.norm.pdf(x,n*0.3,np.sqrt(n*0.3*0.7)),'k',label='Normal')
- ax.legend()
- fig

5.3.2 The Central Limit Theorem

Let $X_1, \cdots, X_n$ be iid with a distribution with a mean μ and a variance σ^2. Then $\overline{X} = \frac{\sum_i^n X_i}{n}$ approximately follows $\mathrm{N}(\mu, \frac{\sigma^2}{n})$ for a large n.

5.3.3 Simulation Experiment 1:
An Investigation of the Central Limit Theorem

- The central limit theorem applies better if the distribution for the sample is closer the a normal distribution.
- Otherwise, the normal approximation of the distribution of the average of iid random variables will be slower.

Example 17: Milk Container Contents

(1) X~ B(20, 0.261) : the number of underweight container

Y~ N(5.22, 3.86) : approximation

$P(X \leq 3) = 0.1935$

$P(Y \leq 3.5) = 0.1922$

(2) Now suppose X~B(500, 0.261).

Then Y~ N(130.5, 96.44)

The probability that at least 150 out of 500 are underweight

$P(X \geq 150) = ?$

$P(Y \geq 149.5) = 0.0265$.

```
import numpy as np
import scipy.stats as stat
print(stat.norm.cdf(149.5,130.5,np.sqrt(96.44)))
=>0.9734895488649663
```

Example 30: Pearl Oyster Farming

The probability that an oyster produces a pearl with a diameter of at least 4mm is 0.6

How many oysters does an oyster farmer need to farm in order to be 99% confident of having at least 1000 pearls?

X : the number of pearls

X ~ B(n, 0.6) ⇒ Y ~ N(0.6n, 0.24n)

$$P(X \geq 1000) \approx P(Y \geq 999.5) = 1 - \Phi\left(\frac{999.5 - 0.6n}{\sqrt{0.24n}}\right) = 0.99.$$

$$\frac{999.5 - 0.6n}{\sqrt{0.24n}} = -z_{0.01} = -2.33.$$

$$n = 1746.$$

```
import scipy.stats as stat
x2=stat.norm.ppf(0.99,0.0,1.0)
print(x2)
=> 2.3263478740408408
```

- In conclusion, the farmer should farm about 1750 oysters in order to be 99% confident of having at least 1000 peals.

  The expected number of pearls and its variance are

  $$E(X) = 1750 \times 0.6 = 1050.$$

  $$Var(X) = 1750 \times 0.6 \times 0.4 \approx 20.5^2$$

- Suppose the diameter of pearl has mean 5.0 and variance 8.33. If 1750 pearls are obtained, the average diameter has mean 5.0 and variance 0.00476(=8.33/1750).

  Interval of the average diameter with 99.7% coverage is

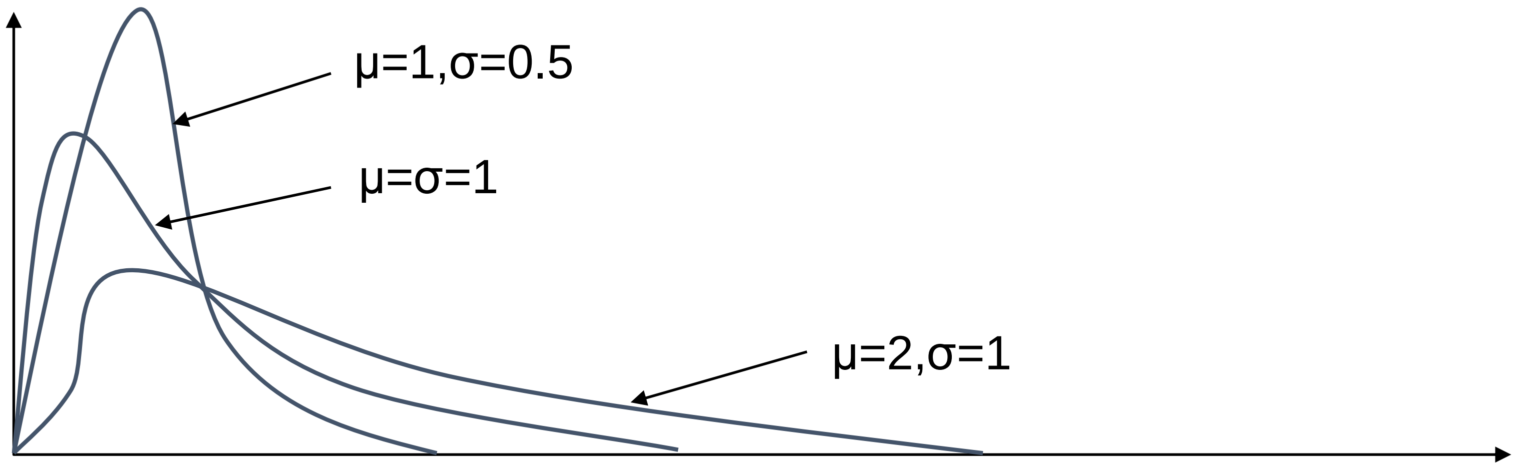
  $$5 \pm 2.97\sqrt{0.00476} = 5 \pm 2.97 \times 0.069 = 5 \pm 0.205$$

```
import scipy.stats as stat
print(stat.norm.ppf(0.9985,0.0,1.0))
=>2.9677
```

5.4 Distributions Related to the Normal Distribution

5.4.1 The Lognormal Distribution

- $Y = \ln(X) \sim \ N(\mu,\ \sigma^2)$.
- PDF: $f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp(-\frac{(\ln(x)-\mu)^2}{2\sigma^2})$ for $x > 0$.
- CDF: $F(x) = \Phi(\frac{\ln(x)-\mu}{\sigma})$.
- $E(X) = \exp(\mu + \frac{\sigma^2}{2})$ and $Var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.



5.4.2 Chi-Square Distribution

- $X_i \sim N(0,1)$. $X = \sum_{i=1}^{v} X_i^2$. X_i's are independent each other. Then

$X \sim \chi_v^2$, where v is called the degrees of freedom of the distribution.

- PDF:

$$f(x) = \frac{\frac{1}{2} e^{-x/2} \left(\frac{x}{2}\right)^{\frac{v}{2}-1}}{\Gamma\left(\frac{v}{2}\right)}$$

$$\chi_v^2 = Gam\left(\frac{v}{2}, \frac{1}{2}\right)$$

- Mean and variance:

$E(X) = v$. $\mathrm{Var}(X) = 2v$.

$X \sim N(0,1)$. Then $X^2 \sim \chi_1^2$

Proof :

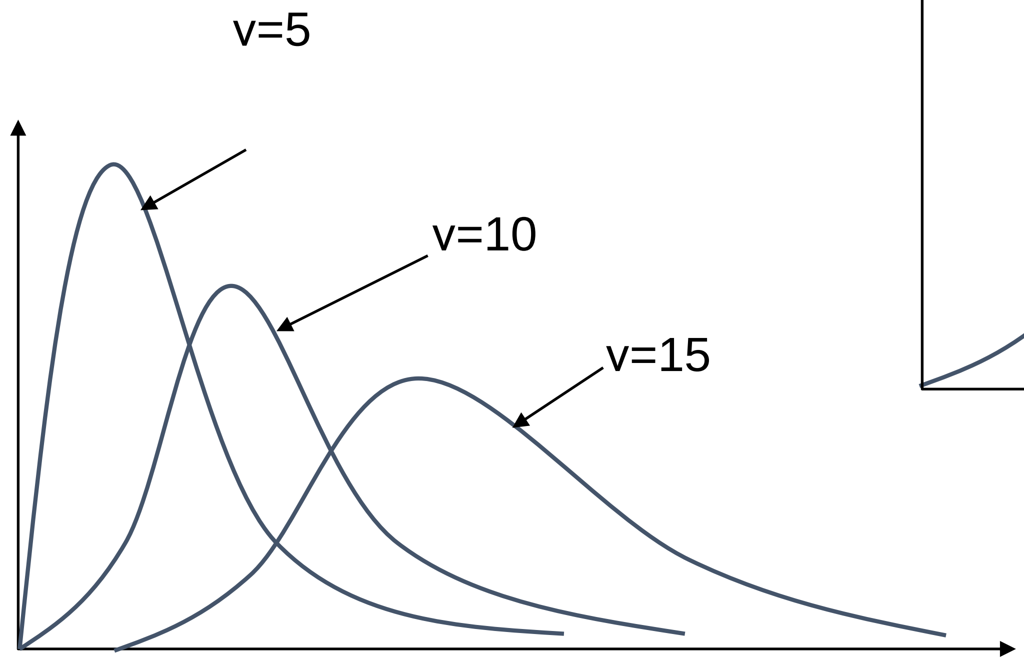
$Y = X^2$.

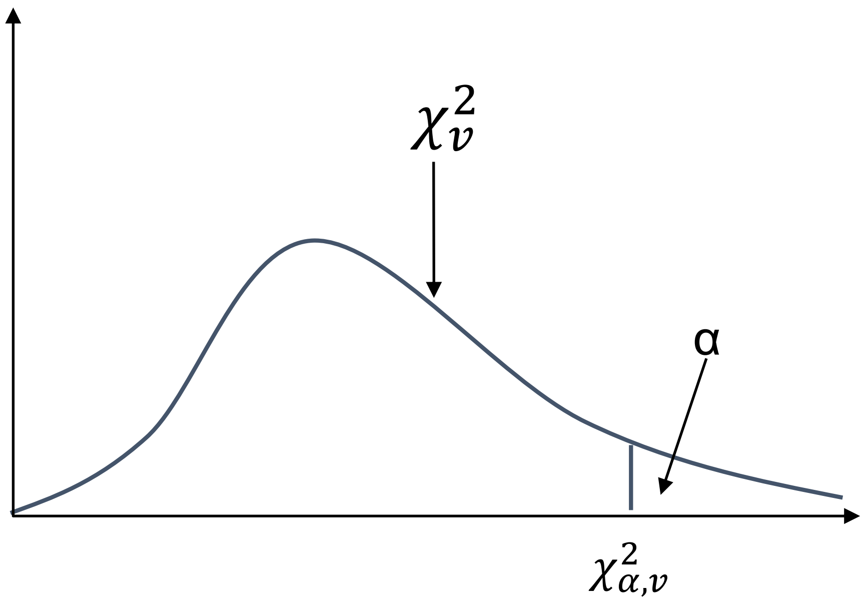
$P(Y \leq y) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) = \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)$.

$f_Y(y) = f_X\left(\sqrt{y}\right)\frac{1}{2\sqrt{y}} + f_X\left(-\sqrt{y}\right)\frac{1}{2\sqrt{y}} = \frac{y^{-1/2}}{\sqrt{2\pi}}e^{-\frac{y}{2}} = \frac{y^{-1/2}}{\Gamma\left(\frac{1}{2}\right)2^{1/2}}e^{-\frac{y}{2}}$.

$f_Y(y)$ is the pdf of Gam(½, ½) $= \chi_1^2$.

- $P\left(X \geq \chi^2_{\alpha,v}\right) = \alpha$





Example 5.4.2a:

Suppose that the coordinate errors are independent normal random variables with mean 0 and standard deviation 2.

Find the probability that the distance error between the points chosen and the target exceeds 3.

(Sol) Let $D^2 = X^2 + Y^2$, where $X,\ Y$ are independent coordinate errors.

Then $\frac{D^2}{4} \sim \ \chi_2^2$.

So the desired probability is given by

$$P(D^2 > 3^2) = P\left(\frac{D^2}{4} > \frac{3^2}{4}\right) = e^{-\left(\frac{1}{2}\right)\left(\frac{9}{4}\right)} = 0.3247.$$

```
import scipy.stats as stat
print(stat.chi2.cdf(9/4,2))
=> 0.6753475326416503
```

5.4.3 The t-Distribution

- $Z \sim N(0,1)$. $W \sim \chi^2_v$. Z and W are independent. Then
  $\mathrm{T_v} = \frac{\mathrm{Z}}{\sqrt{W/v}} \sim t_v$, a t-distribution with v degrees of freedom.

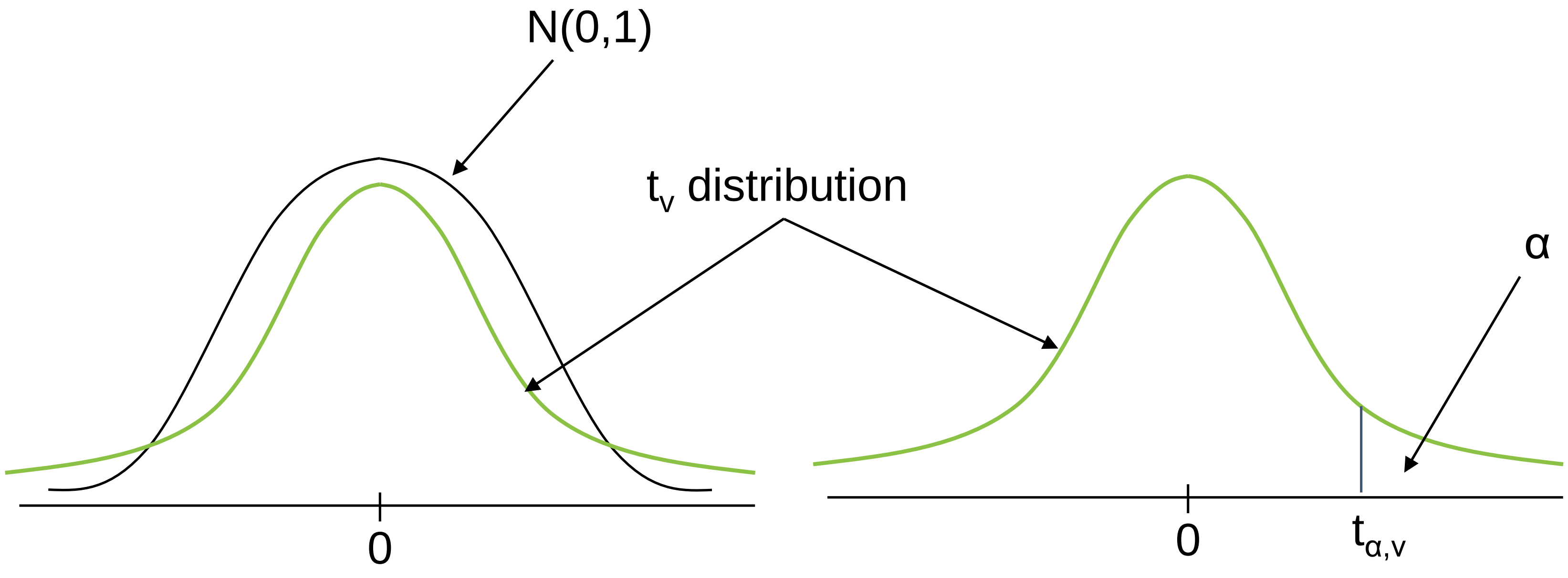
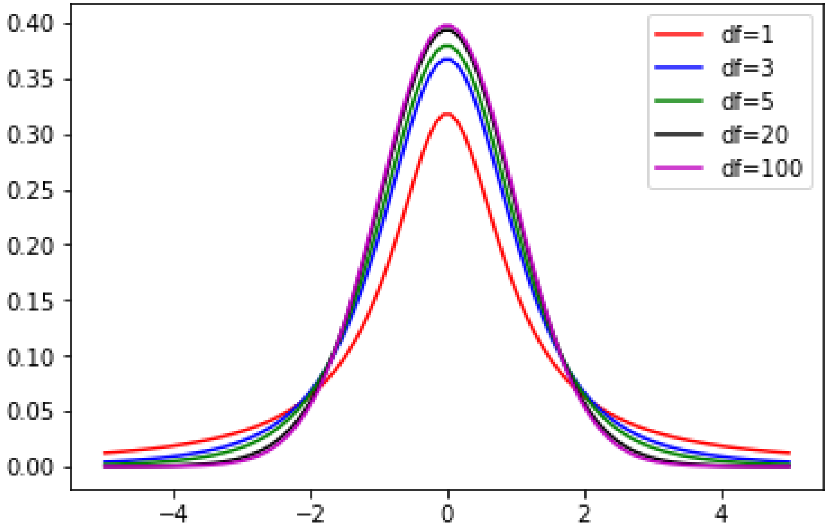


Figure 5.4a The *t*-distribution curves with Python

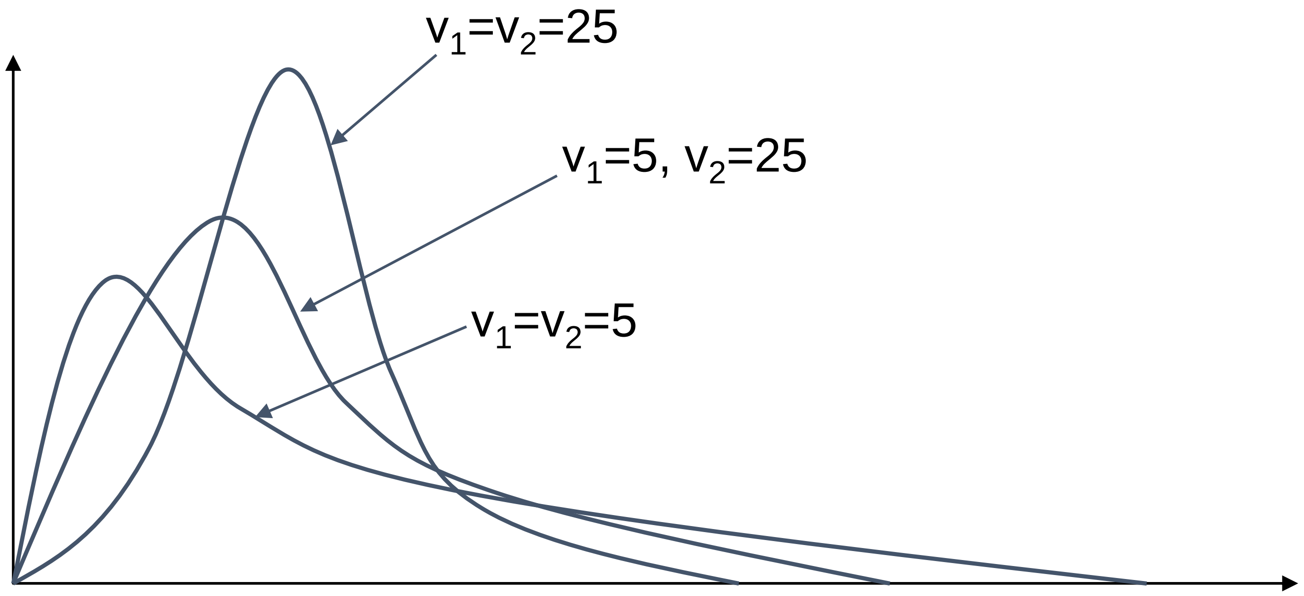


Python code

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import scipy.stats as stat
- x=np.linspace(-5,5,100)
- plt.plot(x,stat.t.pdf(x,1),'r')
- plt.plot(x,stat.t.pdf(x,3),'b')
- plt.plot(x,stat.t.pdf(x,5),'g')
- plt.plot(x,stat.t.pdf(x,20),'k')
- plt.plot(x,stat.t.pdf(x,100),'m')
- plt.legend(['df=1','df=3','df=5','df=20','df =100'],loc='upper right')
- plt.show()

5.4.4 The F-Distribution

- $W_i \sim \ \chi^2_{v_i}$ for $i = 1,2,$ and they are independent. Then
- $\frac{W_1}{v_1} / \frac{W_2}{v_2} \sim F_{v_1, v\_2}$, an F-distribution with degrees of freedom, $v_1,\ v_2$.



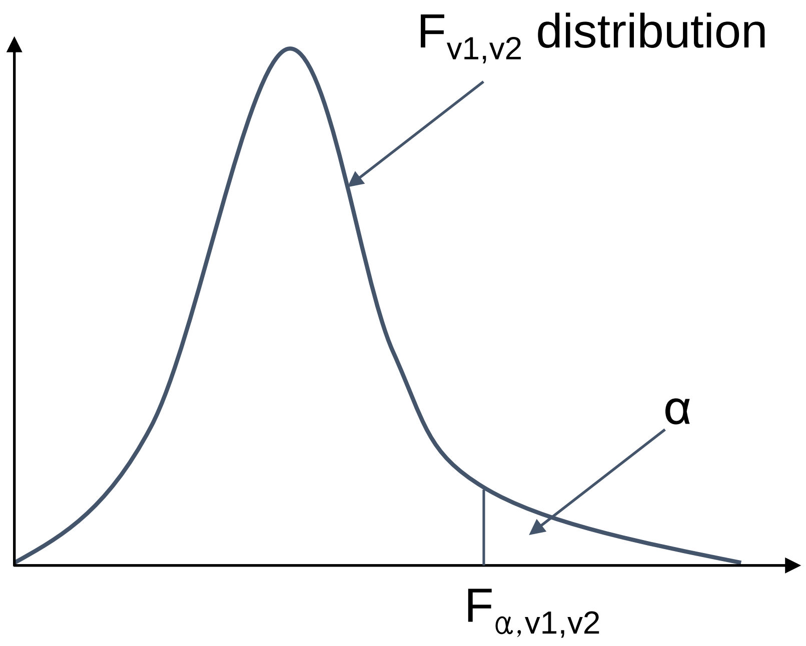
- $F_{1-\alpha,v_1,v_2} = \dfrac{1}{F_{\alpha,v_2,v_1}}.$

Proof:

$W \sim F_{v_1,v_2}$. Then $\frac{1}{W} \sim F_{v_2,v_1}$.

$P\left(W \le F_{1-\alpha,v_1,v_2}\right) = P\left(\dfrac{1}{W} \ge \dfrac{1}{F_{1-\alpha,v_1,v_2}}\right)$

$= \alpha$



5.4.5 The Multivariate Normal Distribution

- Bivariate normal distribution for (X,Y) with parameters, $\mu_1,\ \mu_2, \sigma_1^2, \sigma_2^2, \rho,$ where $\mu_1 = E(X)$,

$$\mu_2 = E(Y),\ \ \sigma_1^2 = Var(X),\ \ \sigma_2^2 = Var(Y),\ \rho = Corr(X,Y).$$

- Joint PDF of (X, Y):

$$f(x,y)$$

$$= \frac{1}{2\pi\sigma_1\,\sigma_2\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\,\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]),$$

for $\infty < x, y < \infty$.

- In particular, when $\mu_1 = \mu_2 = 0,\ \sigma_1 = \sigma_2 = 1$:

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[x^2 + y^2 - 2\rho xy\right]\right).$$

- Furthermore, when $\mu_1 = \mu_2 = 0,\ \sigma_1 = \sigma_2 = 1,\ \rho = 0$:

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left[x^2 + y^2\right]\right).$$

The last formula indicates independence between X and Y.

Chapter Summary

5.1 Probability Calculation Using the Normal Distribution

5.2 Linear Combinations of Normal Random Variables

5.3 Approximating Distributions with the Normal Distribution

5.4 Distributions Related to the Normal Distribution

Log-normal distribution, Chi-square distribution, t-distribution, F-distribution, Bivariate normal distribution.