Algorithm Trading

Day2 - Statistics & Probability

James Ahn

STATISTICS

Statistics

- Statistics is the science and practice of developing human knowledge through the use of empirical data expressed in quantitative form. It is based on statistical theory which is a branch of applied mathematics. Within statistical theory, randomness and uncertainty are modelled by probability theory

(Wikipedia Encyclopedia)

데이터의 전체적인 모습을 파악하기 위해서

Why Statistics for Algo-Trading?

EDA와 검증을 위해 통계적 지식이 요구된다.

EDA
(Explanatory Data Analysis)

Alpha Signal

By collecting, summarizing, and analyzing of data

Signal or Model

Verification

By hypothesis testing, is it luck or not?

Types of Statistics

Descriptive

- How many men work at Fraser Health?
- How many hours a week do employees spend at their desks?

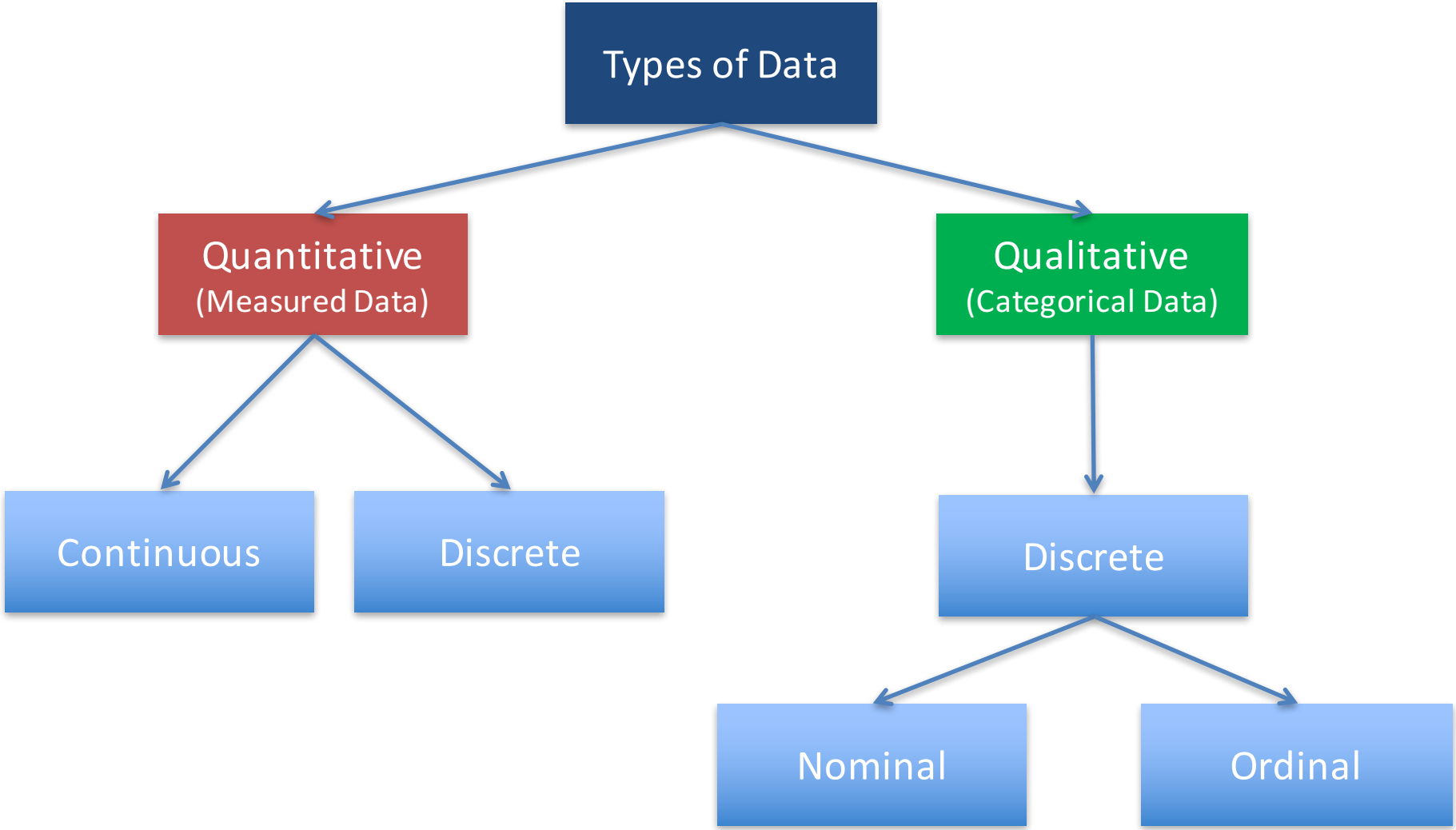
변수들간의 관계를 설명하는데 관심

Inferential

- Does having a science degree help students learn statistical concepts?
- What risk factors most predict heart disease?

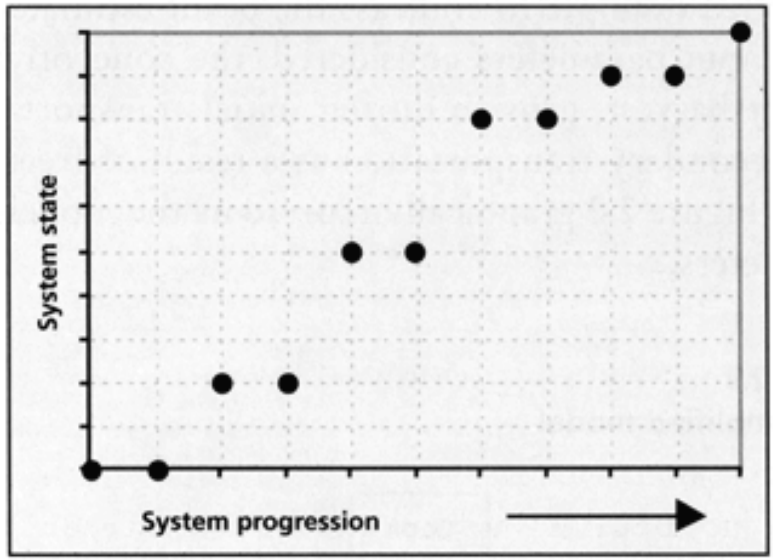
데이터의 원래 모습을 추론에 관심

Data

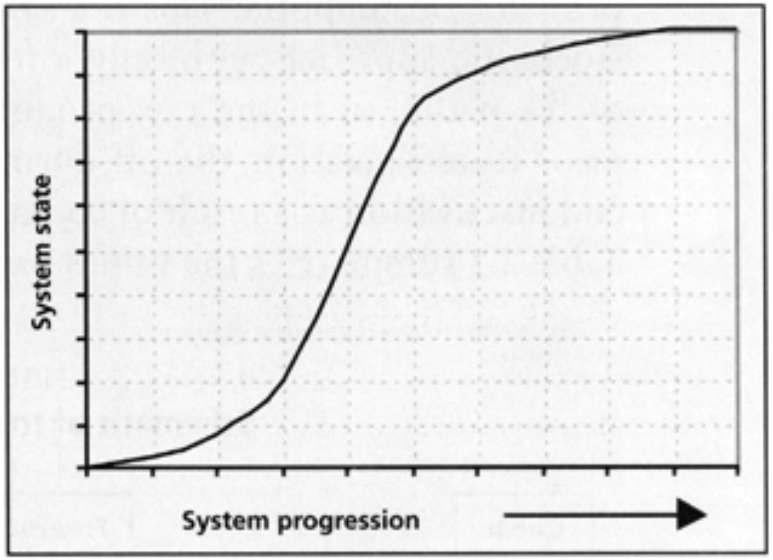

Types of Data
Quantitative
(Measured Data)
Qualitative
(Categorical Data)
Continuous
Discrete
Discrete
Nominal
Ordinal

Measured Data

Discrete Variable (정수)



Continuous Variable (실수)



Categorical Data

- Categorical (Qualitative) Data occur when we assign objects into labelled groups or categories
- 예) 성별, 정치인 선호도, 영화 평점, 친절도 평가등
- **Ordinal variables**
  - a natural ordering e.g. gold/silver/bronze medal
  - 영화평점(1 to 5), 정치인 선호도
- **Nominal variables**
  - do not have a natural ordering e.g. gender
  - Label Encoding
  - 남녀성별, 자동차회사

Nominal Variable은 크기가 없지만 Ordinal Variables은 크기가 있다.

Variable

A variable is a property of an object or event that can take on different values.

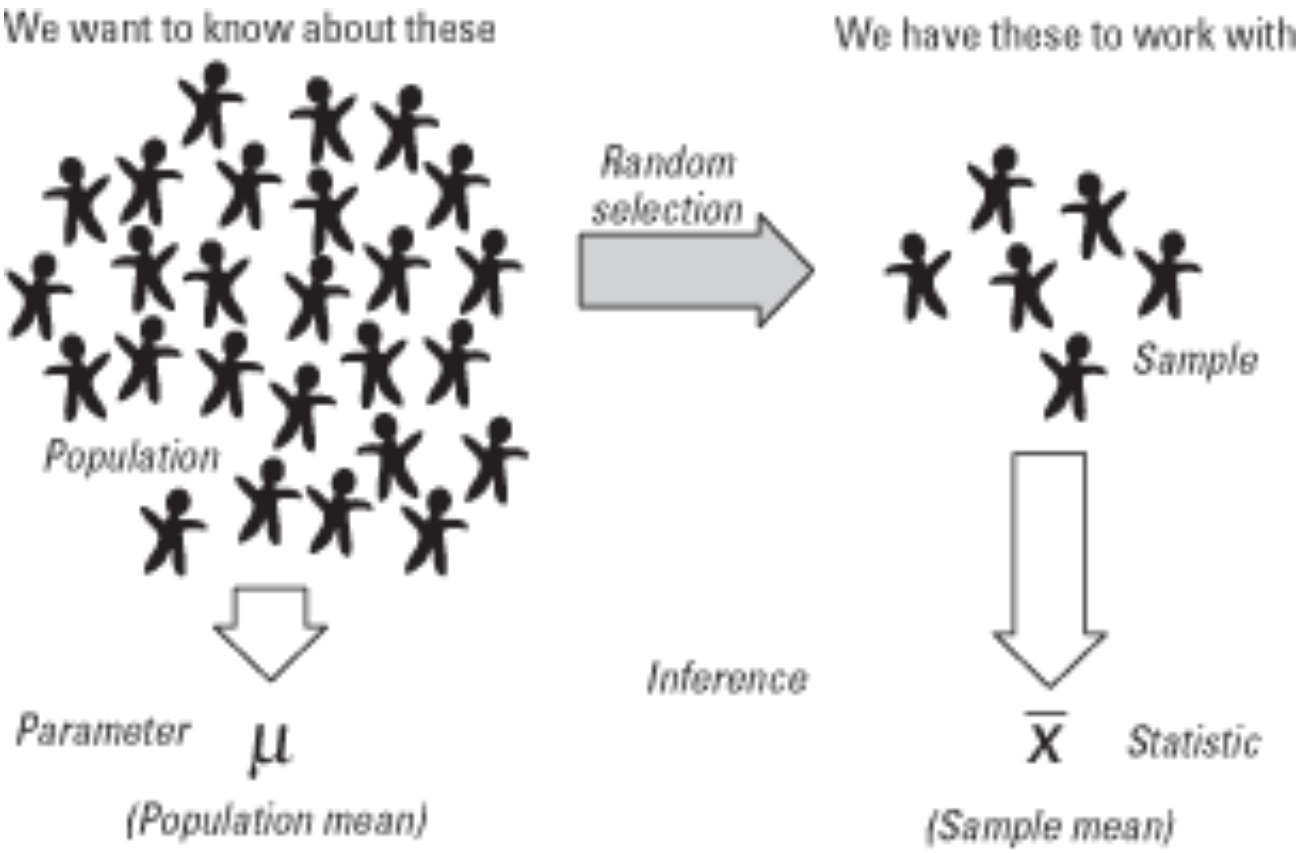
- **independent variable(독립변수)**
  - the variable that you believe will **influence your outcome measure.**
  - **Input Variable**
- **dependent variable(종속변수)**
  - the variable that is dependent on or influenced by the independent variable(s).
  - A dependent variable may also be the variable you are trying to predict.
  - Output Variable
- Ex) 주택가격에 집의 크기와 위치가 영향을 미친다고 하면 주택가격은 Dependent Variable, 크기와 위치는 Independent Variable이다.

Population and Sample

- A statistical population is the **set of all measurements** (or record of some quality trait) corresponding to each unit in the entire population of units about which information is sought.
- A sample from a statistical population is the **subset of measurements** that are actually collected in the course of an investigation.

매우 중요한 개념, 하지만 자주 잊어버리는 개념

Population and Sample Example



Mean

Height to Weight Chart for Boys:

Age (Mths)	Weight (Pds)	Length (Ins)
0	7.4	19.6
1	9.8	21.6
2	12.3	23.0
3	14.1	24.2
4	15.4	25.2
5	16.6	26.0
6	17.5	26.6
7	18.3	27.2
8	19.0	27.8
9	19.6	28.3
10	20.1	28.8
11	20.8	29.3
12	21.3	29.8
13	21.8	30.3
14	22.3	30.7
15	22.7	31.2
16	23.2	31.6
17	23.7	32.0
18	24.1	32.4
19	24.6	32.8
20	25.0	33.1
21	25.5	33.5
22	25.9	33.9
23	26.3	34.2

Age (Yrs)	Weight (Pds)	Height (Ins)
2	27.5	34.2
3	31.0	37.5
4	36.0	40.3
5	40.5	43.0
6	45.5	45.5
7	50.5	48.0
8	56.5	50.4
9	63.0	52.5
10	70.5	54.5
11	78.5	56.5
12	88.0	58.7
13	100.0	61.5
14	112.0	64.5
15	123.5	67.0
16	134.0	68.3
17	142.0	69.0
18	147.5	69.2
19	152.0	69.5
20	155.0	69.7

- The sum of all the scores divided by the number of scores.
- Often referred to as the average.
- **Good measure of central tendency.**
- Central tendency is simply the location of the middle in a distribution of scores.

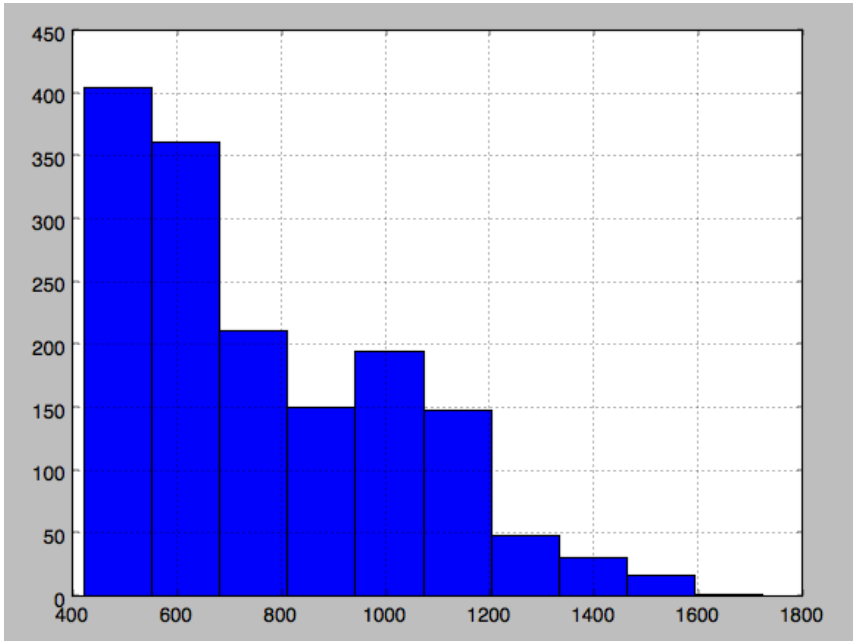
- dataframe.mean()

모든 값을 가장 설명할 수 있는 값

Min, Max and Range

- Min
  - dataframe.min()
- Max
  - dataframe.max()
- Range = 데이터의 범위
  - dataframe.max() – dataframe.min()

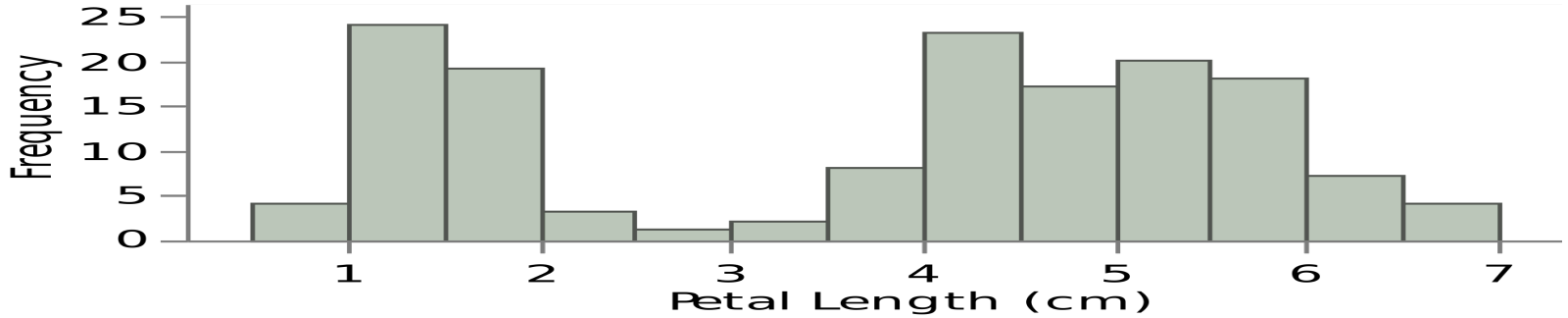
Histogram(Distribution)



- A histogram is a graphical representation of the **distribution of numerical data**.
- dataframe.hist()

- 데이터의 전체적인 모습을 시각적으로 확인할 수 있는 유용한 도구
- 통계와 확률의 중요한 기초개념

Variance & Standard Deviation



- The variance is a measure of how spread out a distribution is.
- The larger the variance, the further spread out the data
- dataframe.var()

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- Standard Deviation

$$\sigma = \sqrt{\mathrm{E}(X - \mathrm{E}(X))^2} = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2}$$

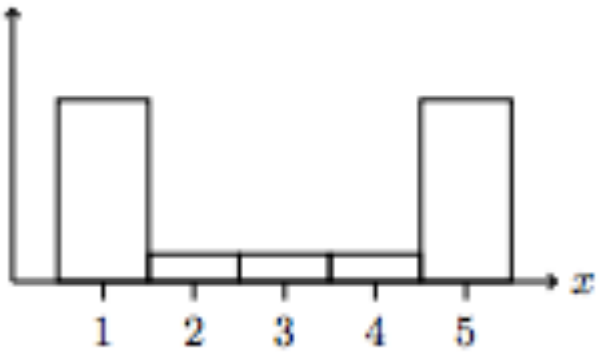
- dataframe.std()

반드시 숙지해야 하는 매우 매우 중요한 개념

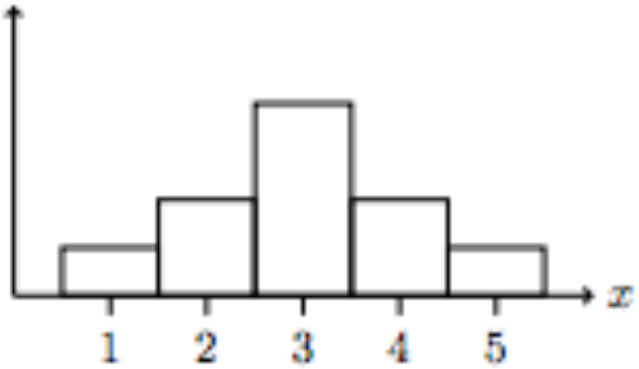
Quiz

다음 2개의 히스토그램중 Standard Deviation이 큰 것은?

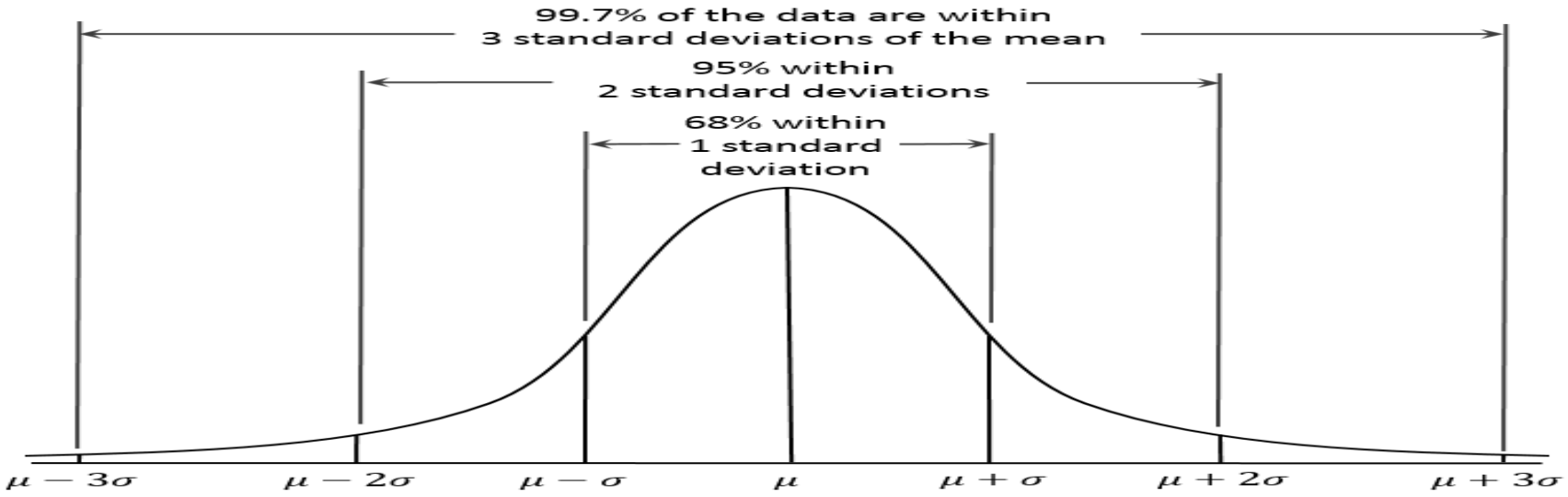
A)



B)



Normal Distribution



- **<u>평균이 0이고 표준편차가 1인 분포 N (0,1) 을 표준정규분포라고 한다.</u>**
- 확률론과 통계학에서, 정규분포(正規分布, 영어: **<u>normal distribution</u>**) 또는 가우스 분포(Gauß分布, 영어: **<u>Gaussian distribution</u>**)는 연속 확률 분포의 하나이다.
- 정규분포는 수집된 자료의 분포를 근사하는 데에 자주 사용되며, 이것은 중심극한정리에 의하여 독립적인 확률변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문이다.

STATISTICS EXERCISE

Tips

- Loading data from data file
  - Import pandas as pd
  - df = pd.read_pickle(file_name)
- Draw chart
  - import matplotlib.pyplot as plt
  - plt.show()
- Data file format
  - 일자, 시작가,최고가,최저가,종가,거래량,수정종가

Pandas Reference
http://pandas.pydata.org/pandas-docs/stable/index.html

Descriptive Statistics

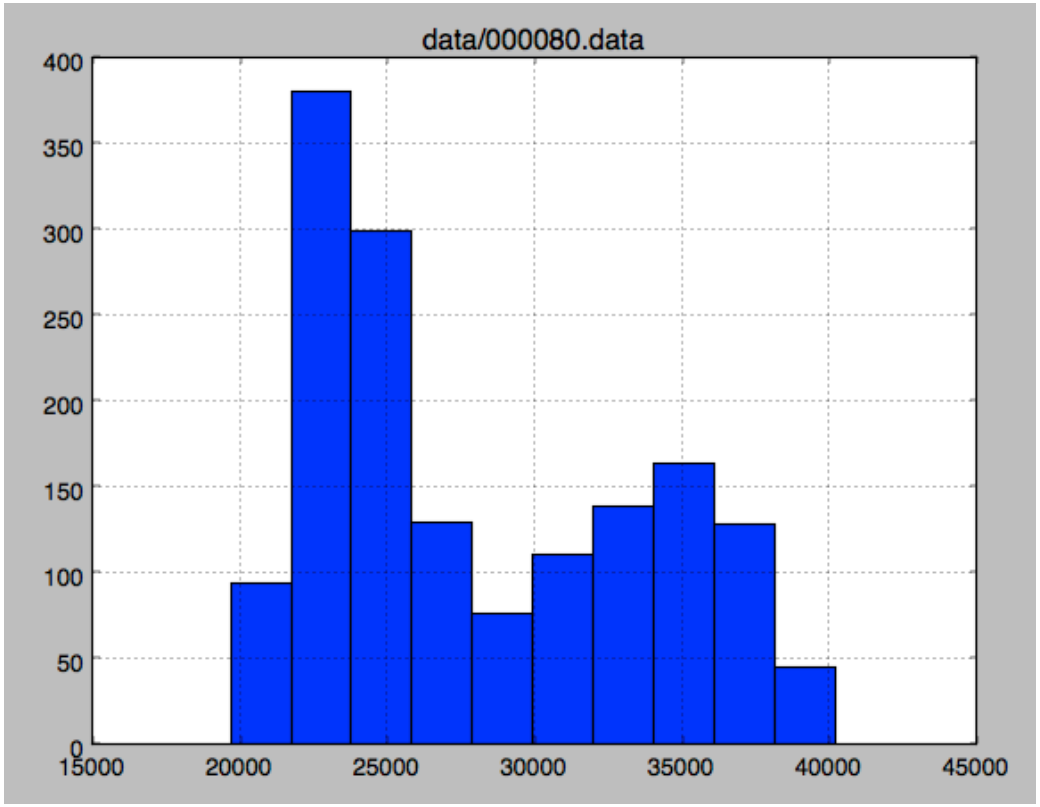
- 다음 3개 종목의 기초통계량을 계산하고 히스토그램을 그려라
  - 000040, 000215, 000480
  - min,max,mean,var,std

종목 선택

- 다음 5개 종목들에 대해 답하라.
  - 000040, 000215, 000480,001515,003547
  - 가장 수익성이 좋을 것으로 예상되는 종목은?
  - 개발한 모델의 적중율이 가장 높은 것으로 예상되는 종목은?

Histogram Analysis

- 다음은 하이트진로의 주가 히스토그램이다.
  - 데이터분산의 측면에서 무엇을 유추할 수 있는가?
  - 아래의 데이터를 그대로 Algorithm Trading에 사용하는 것은 좋은가?



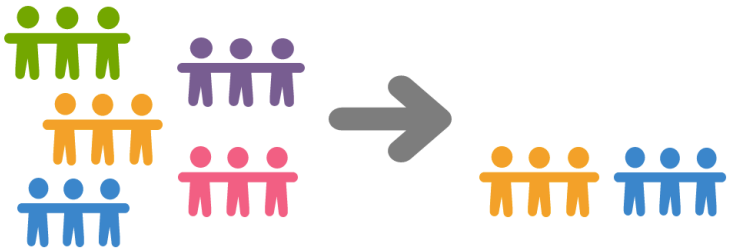
BRAINTEASER

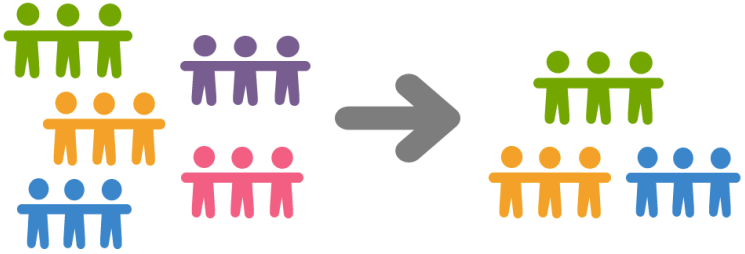
Posco 주가 데이터

- 아래의 그래프는 포스코의 2006-2010년 주가데이터이다.
  - 이 데이터를 이용해 Alpha Model을 만들었다면, 여기에는 통계적인 측면에서 어떤 잠재적인 문제가 있을까?
  - 주의해야 할 사항은 무엇인가?

Sampling Bias

In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. **It results in a biased sample**, a non-random sample[1] of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected.

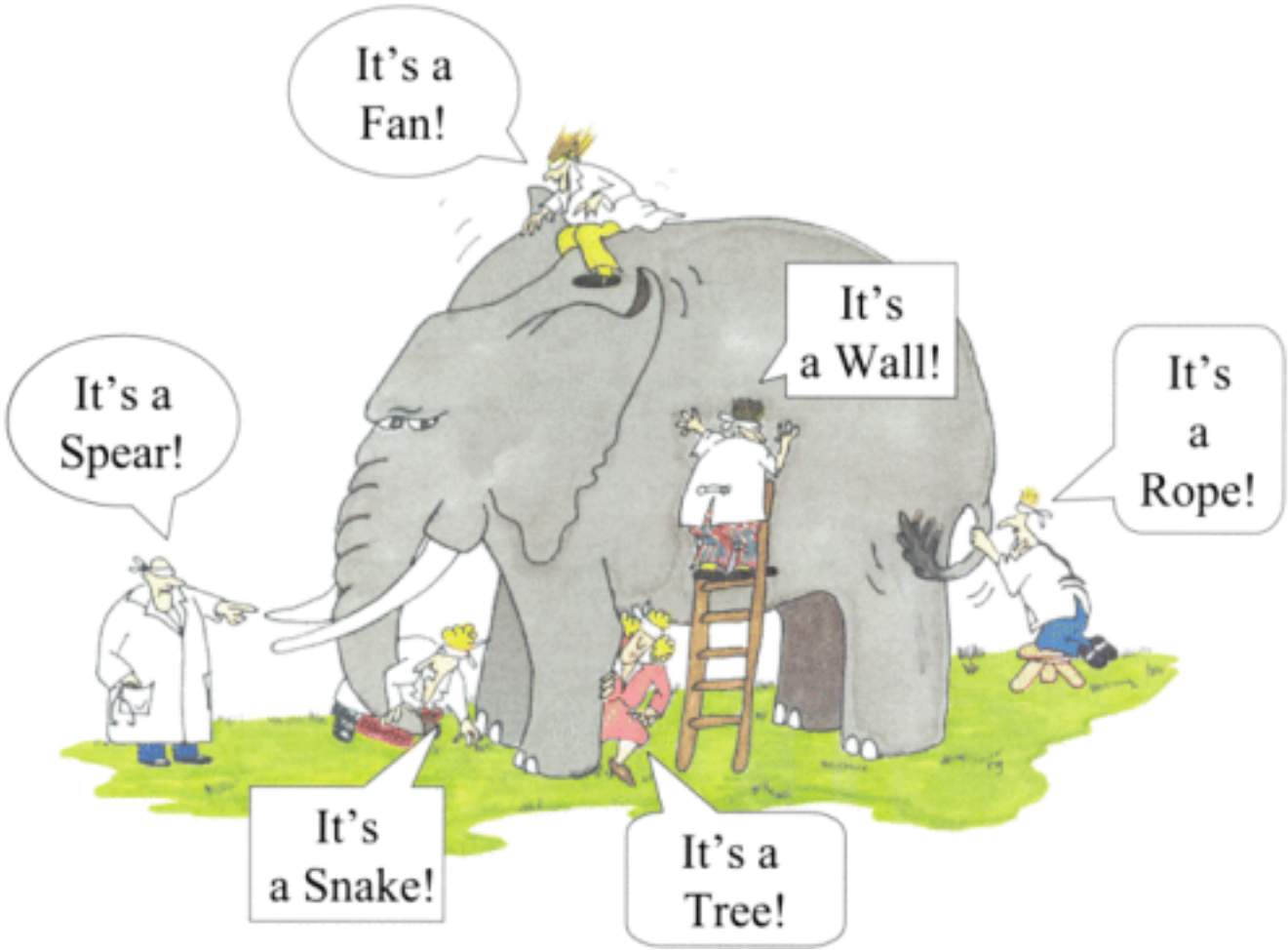




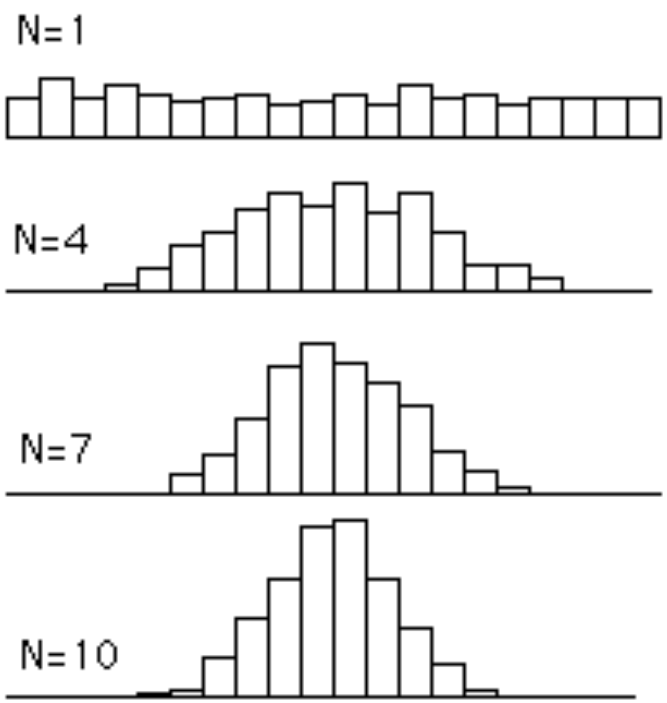
Alpha Model의 개발을 위해, 혹은 종목 분석을 위해 사용한 주가데이터는 Sampling Bias를 가지고 있다. 절대 자유로울 수 없다.!!!

Sampling Bias is a Fate!!!

주가데이터를 다루는 우리의 모습



CLT(Central Limit Theorem)



- 중심극한정리central limit theorem, 약자 CLT)는 동일한 확률분포를 가진 독립 확률변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 정리이다.
- **매우 불규칙한 분포도 충분히 많은 수를 더하면 중심극한정리에 따라 결국 정규분포로 수렴한다.**

반드시 숙지해야 하는 매우 중요한 개념

CLT And Interpretation

EDA를 이용해 y = sin(ax+b) 라는 Alpha Model을 찾았다고 가정하자.
그리고 종목 데이터를 이용해 a=2, b=1 값이 최적의 결과를 나타낸다는 것을 알았다.
최종적으로 완성된 alpha model은 y = sin(2x+1) 이다.

Q) Alpha Model 개발에 사용된 데이터가 Sampling Bias를 가지고 있다면, Parameter A와 B의 값을 어떻게 해석하는 것이 좋은가?

PROBABILITY

Random Variable

- A random variable *x* takes on **a defined set of values with different probabilities.**
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
- **Continuous Random Variable**
    - an infinite continuum of possible values
    - 혈압, 자동차의 속도
- **Discrete Random Variable**
    - a countable number of outcomes
    - 동전의 앞뒤면, 주사위눈

Definition of Probability

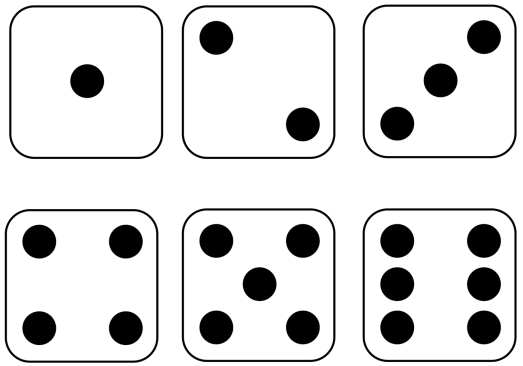
- **Probability is the ratio of the number of occurrences of an event to the total number of experiments**, in the limit of very large number of repeatable experiments.
- Can only be applied to a specific classes of events (**repeatable experiments**)

Probability

$$\text{Probability} = \frac{\text{Number of favorable cases}}{\text{Number of total cases}}$$

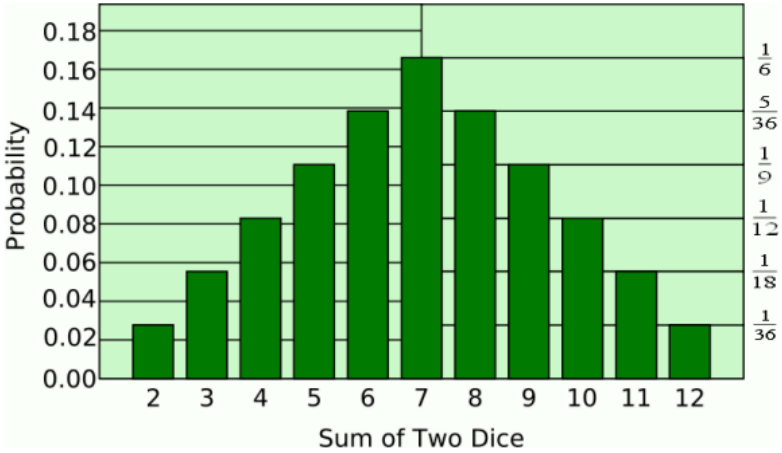
P = 1/2

P = 1/6



Sample Space

set of all possible outcomes in Events



E.g:

2 = {(1,1)}
3 = {(1,2), (2,1)}
4 = {(1,3), (2,2), (3,1)}
5 = {(1,4), (2,3), (3,2), (4,1)}
etc. …

Practice Problem

포트폴리오로 보유중인 6개의 주식중 동반상승할 확률이 다음과 같다.

X	1	2	3	4	5	6
P(x)	0.4	0.3	0.1	0.1	0.05	0.05

다음의 확률을 계산하라.

a. 4개가 상승할 확률은?

$P(x) = P(4) = 0.1$

b. 3개이상 상승할 확률은?

$P(x) = P(3) + P(4) + P(5) + P(6) = 0.1 + 0.1 + 0.05 + 0.05$

c. 2개 이하가 상승할 확률은?

$P(x) = P(1) + P(2) = 0.4 + 0.3$

Expected Value and Variance

x	*P(x≤A)*
1	*P(x≤1)*=1/6
2	*P(x≤2)*=2/6
3	*P(x≤3)*=3/6
4	*P(x≤4)*=4/6
5	*P(x≤5)*=5/6
6	*P(x≤6)*=6/6

All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).

통계에서는 평균, 확률에서는 기대값

Expected Value

- **Expected value is just the average or mean (μ) of random variable *x*.**
- It's sometimes called a "weighted average" because more frequent values of X are weighted more highly in the average.
- It's also how we expect X to behave on-average **over the long run**

Discrete case:

$$E(X) = \sum_{\text{all x}} x_i \, p(x_i)$$

Continuous case:

$$E(X) = \int_{\text{all x}} x_i \, p(x_i) dx$$

Example : Expected Value

포트폴리오로 보유중인 6개의 주식중 동반상승할 확률이 다음과 같다.

X	1	2	3	4	5	6
P(x)	0.4	0.3	0.1	0.1	0.05	0.05

$$E(X) = \sum_{\text{all } x} x_i \, p(x_i)$$

E(x) = 1*0.4 + 2*0.3 + 3*0.1 + 4*0.1 + 5*0.05 + 6*0.05 = 2.25

평균적으로 2.25개의 주식이 동반 상승할 것으로 기대된다.

Variance

The expected (or average) squared distance (or deviation) from the mean"

Discrete case:

$$Var(X) = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Continuous case:

$$Var(X) = \int_{\text{all x}} (x_i - \mu)^2 \, p(x_i) dx$$

Example : Variance

포트폴리오로 보유중인 6개의 주식중 동반상승할 확률이 다음과 같다.

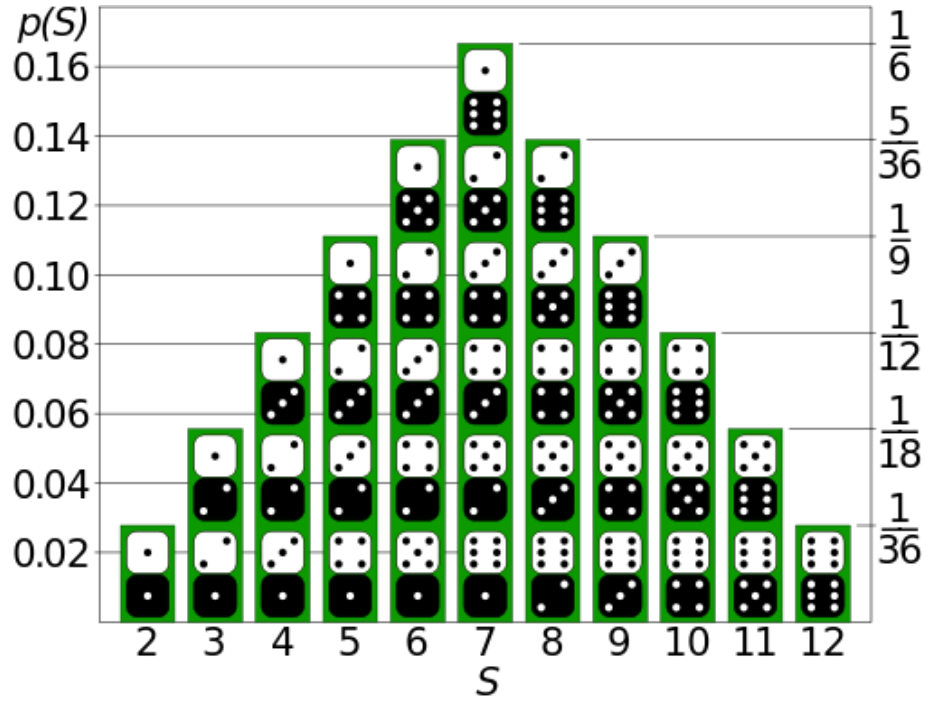
X	1	2	3	4	5	6
P(x)	0.4	0.3	0.1	0.1	0.05	0.05

$$E(x) = 1*0.4 + 2*0.3 + 3*0.1 + 4*0.1 + 5*0.05 + 6*0.05 = 2.25$$

$$\begin{aligned} Var(x) &= (1-2.25)^2 * 0.4 + (2-2.25)^**0.3 + (3-2.25)^2*0.1 + \\ &(4-2.25)^2*0.1 + (5-2.25)^2*0.05 + (6-2.25)^2*0.05 \\ &= 1.5625*0.4 + 0.0625*0.3 + 0.5625*0.1 + \\ &3.0625*0.1 + 7.5625*0.05 + 14.0625*0.05 \\ &= 2.0875 \end{aligned}$$

Probability Functions

확률값



2개 주사위눈의 확률분포

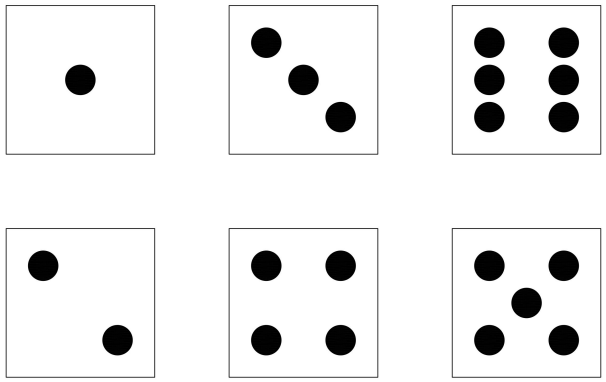
- **A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)***
- *P(X) = [0,1]*

$$\sum_{\text{all } x} P(x) = 1$$

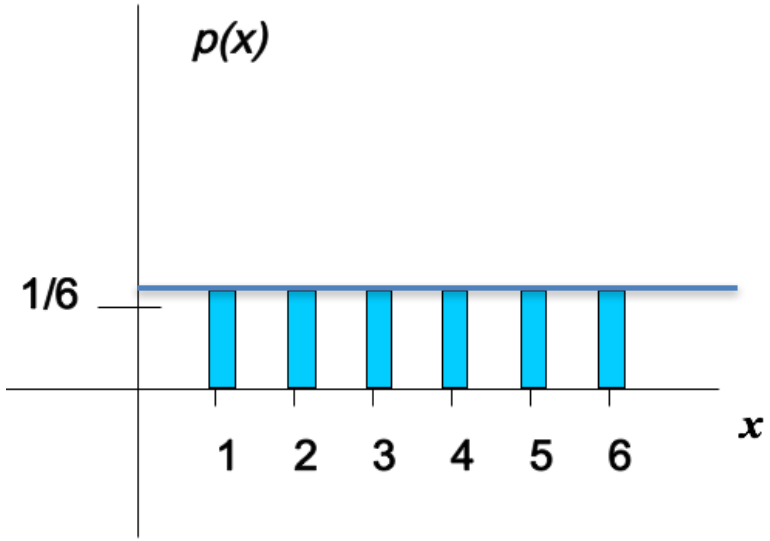
Probability Function을 구할 수 있다면 X에 대한 확률값을 알 수 있다.

Example of Probability Function

Q) 주사위에 대한 Probability Function은?



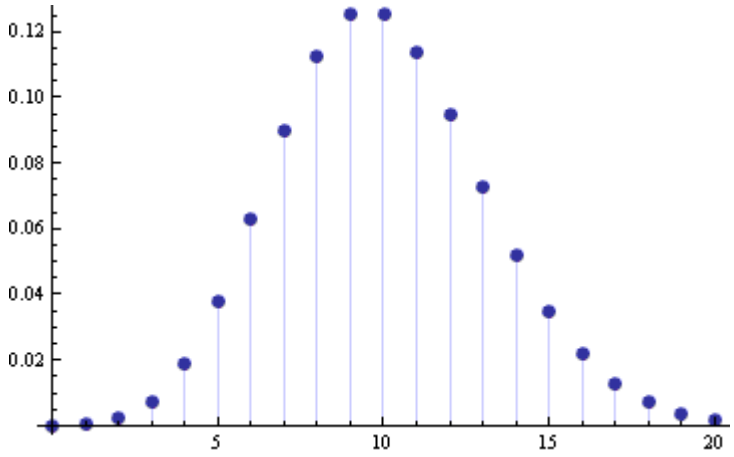
A) Probability Function = 1/6



Two Probability Function Types

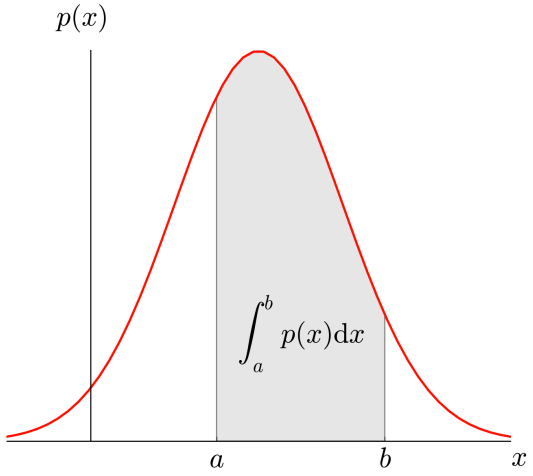
Discrete case:

Probability Mass Function
PMF



Continuous case:

Probability Density Function
PDF



Continuous Probability Distributions

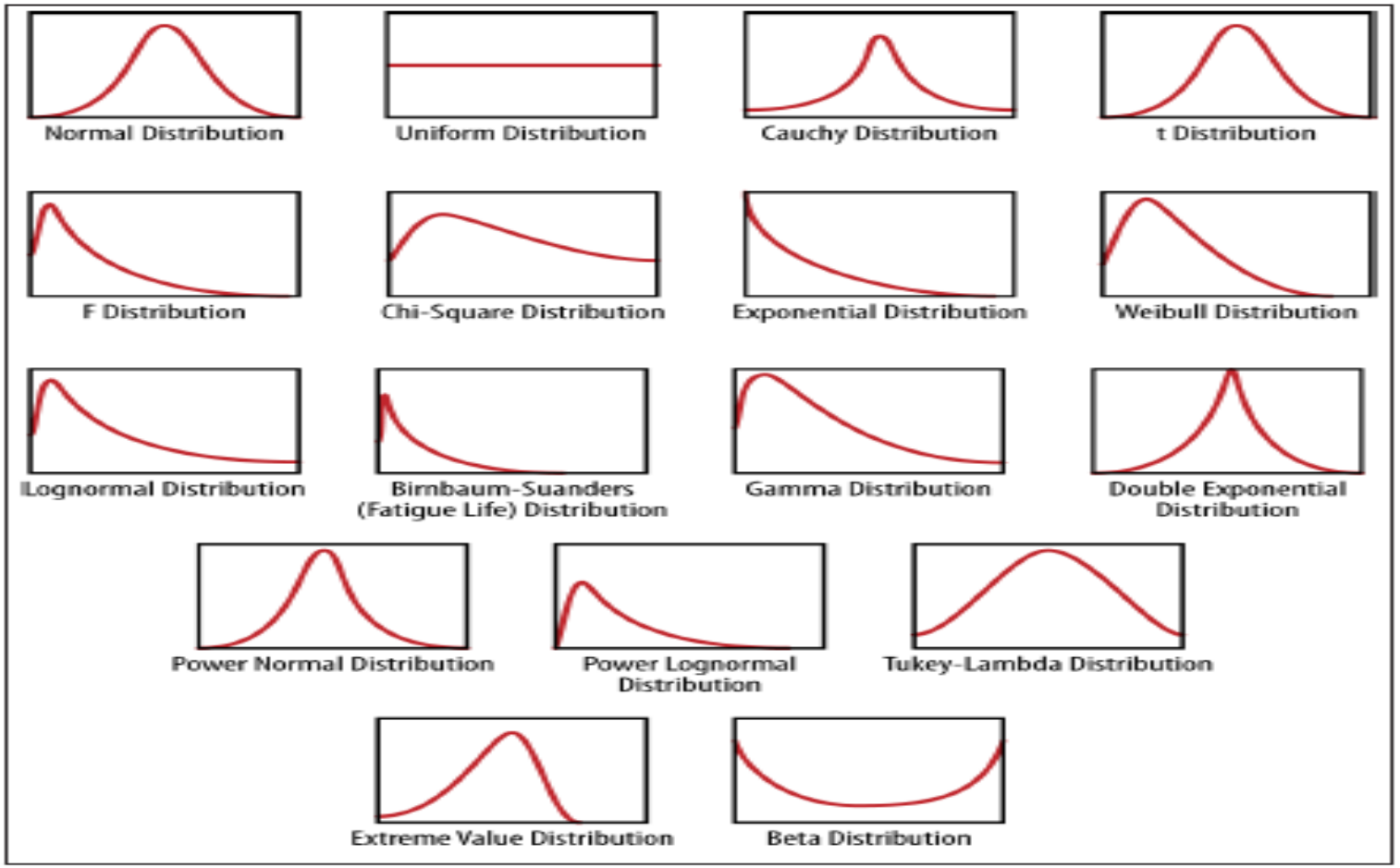
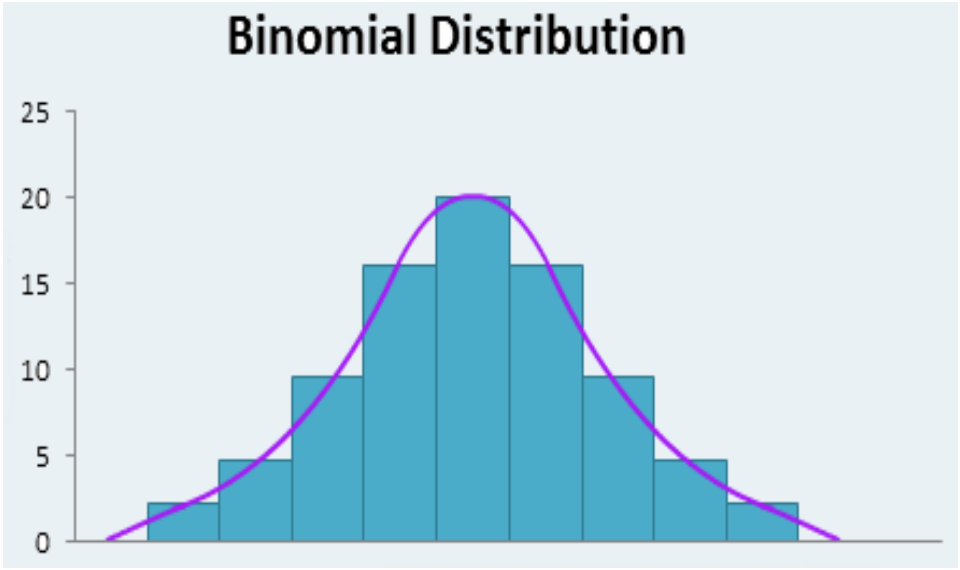
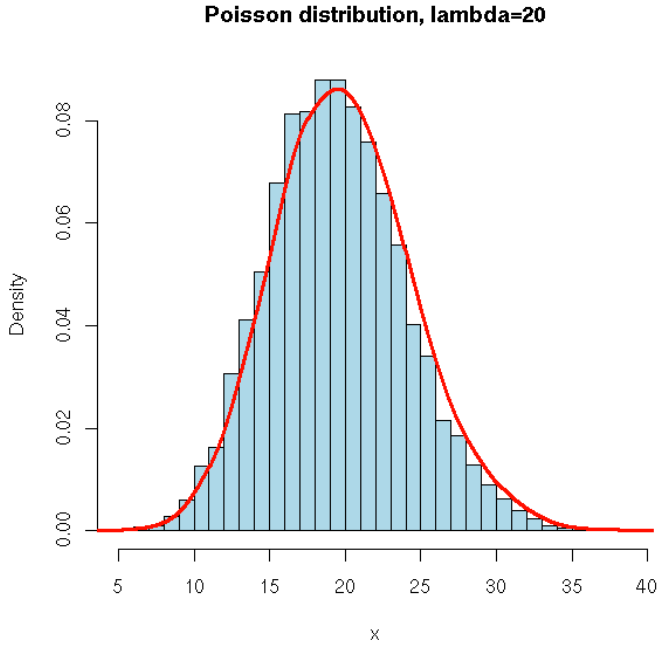


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Discrete Probability Distributions



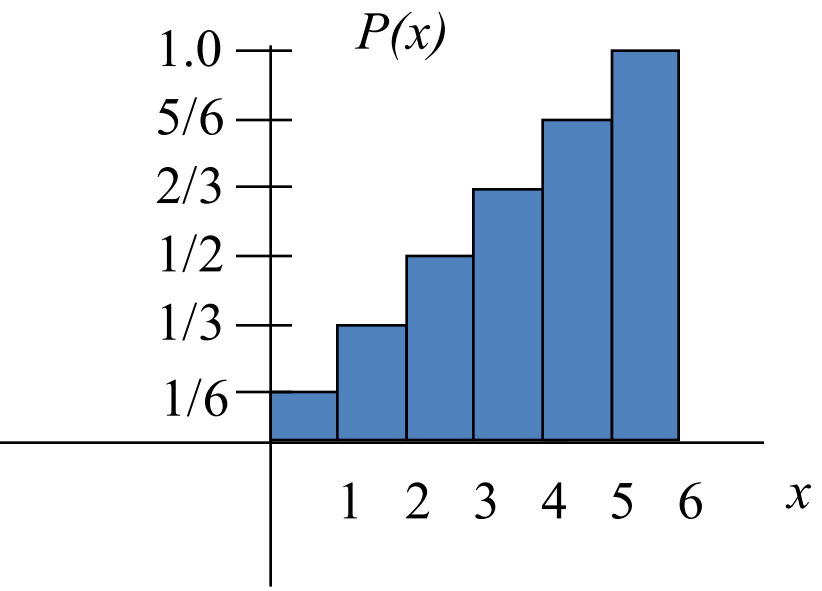


Python Scipy has 81 continuous probability distributions and 10 discrete distributions

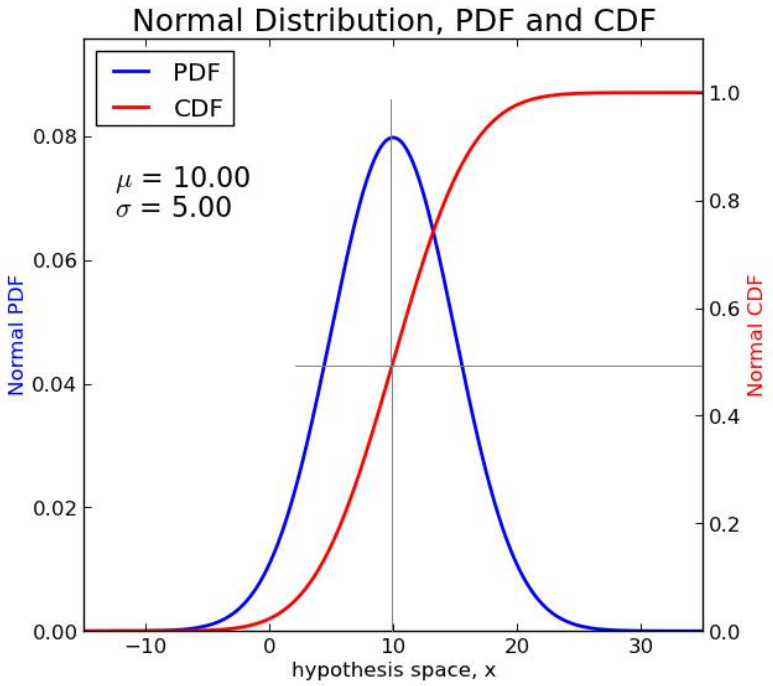
Cumulative Distribution Function

누적 분포 함수(cumulative distribution function, cdf)는 어떤 확률 분포에 대해서, 확률 변수가 특정 값보다 작거나 같은 확률을 나타낸다.

CDF



CDF vs PDF



PROBABILITY EXERCISE

Programming Tips

- To use normal distribution
  - Import scipy.stats
  - scipy.stats.norm(mean,sd)
- To use cauchy distribution
  - Import scipy.stats
  - scipy.stats.cauchy(mean,sd)
- To calculate pdf probability
  - Scipy.stats.norm(mean,sd).pdf(x)
- To calculate cdf probability
  - Scipy.stats.norm(mean,sd).cdf(x)
- 이전값과의 차이를 계산하려면
  - dataframe.diff()

Scipy Reference
http://docs.scipy.org/doc/scipy/reference/stats.html

Quiz 1

- 주식 KR모터스의 값을 예측할 수 있는 Alpha Model을 개발하였다.
  - 주식 KR모터스의 값은 Normal Distribution을 따른다.
  - KR 모터스 = 000040.data

- Q1) Alpha Model이 KR 모터스의 값을 790원으로 예측하였다면, 그 예측이 적중할 확률은 어떻게 되는가?
- Q2) Alpha Model이 주식 KR모터스의 값을 780원으로 예측하였다면, 주식값이 780원 이상이 될 확률은 어떻게 되는가?

Quiz 2

- 주식 KR모터스의 값을 예측할 수 있는 Alpha Model을 개발하였다.
  - 주식 KR모터스의 값은 Cauchy Distribution을 따른다.
  - KR 모터스 = 000040.data
  - 전날의 종가가 770원이었고, Alpha Model은 오늘의 가격을 785원으로 예측하였다.

- Q1) Alpha Model의 예측확률을 주가의변동 관점에서 계산하라.
- 주가변동은 (전날종가-당일종가)로 계산한다.

수고하셨습니다.