기초 확률 및 통계
01. 변수의 종류

1. 변수간의 관계 기준

y (종속변수, 반응변수) x(독립변수, 설명변수)

폐암 발병 ← **흡연 / 음주 / 나이 / 성별**

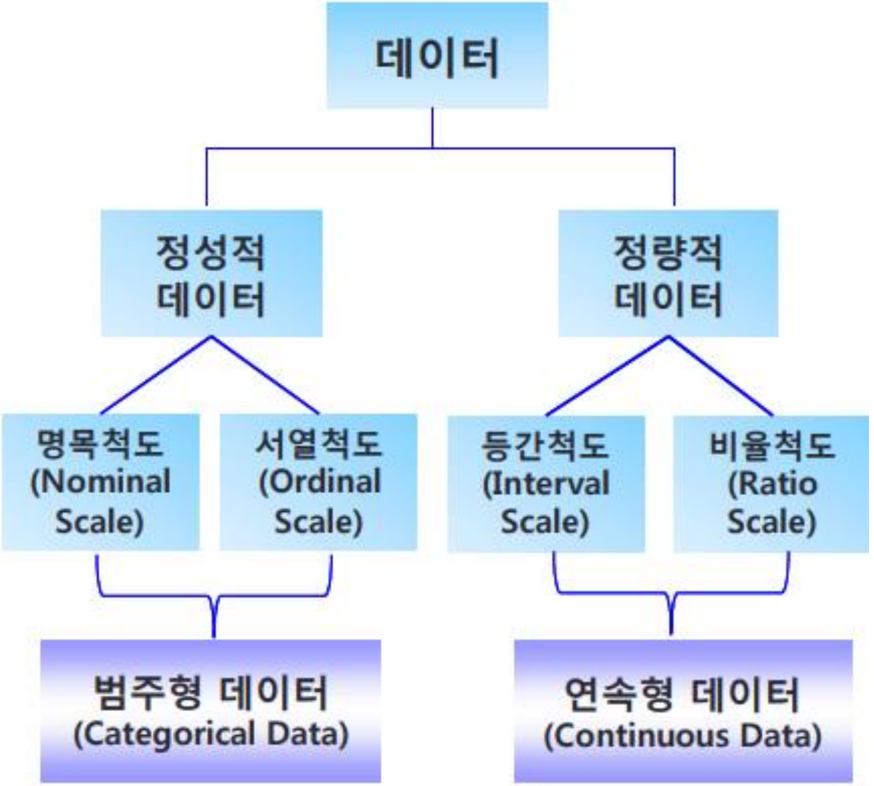
(a) 종속변수(반응 변수)

종속변수는 독립변수에 의해서 변동이 일어나는 변수이다.
독립변수가 얼만큼 영향을 줄 때마다 현상이 얼만큼 변하는지 보여주는 변수다.

(b) 독립변수(설명 변수)

일반적으로 현상을 설명하기 위한 요인들이다. 어느 정도의 영향을 주었기 때문에 이러한 현상이 일어났다라는 말을 하기 위해 설정된 변수다.

기초 확률 및 통계
01. 변수의 종류



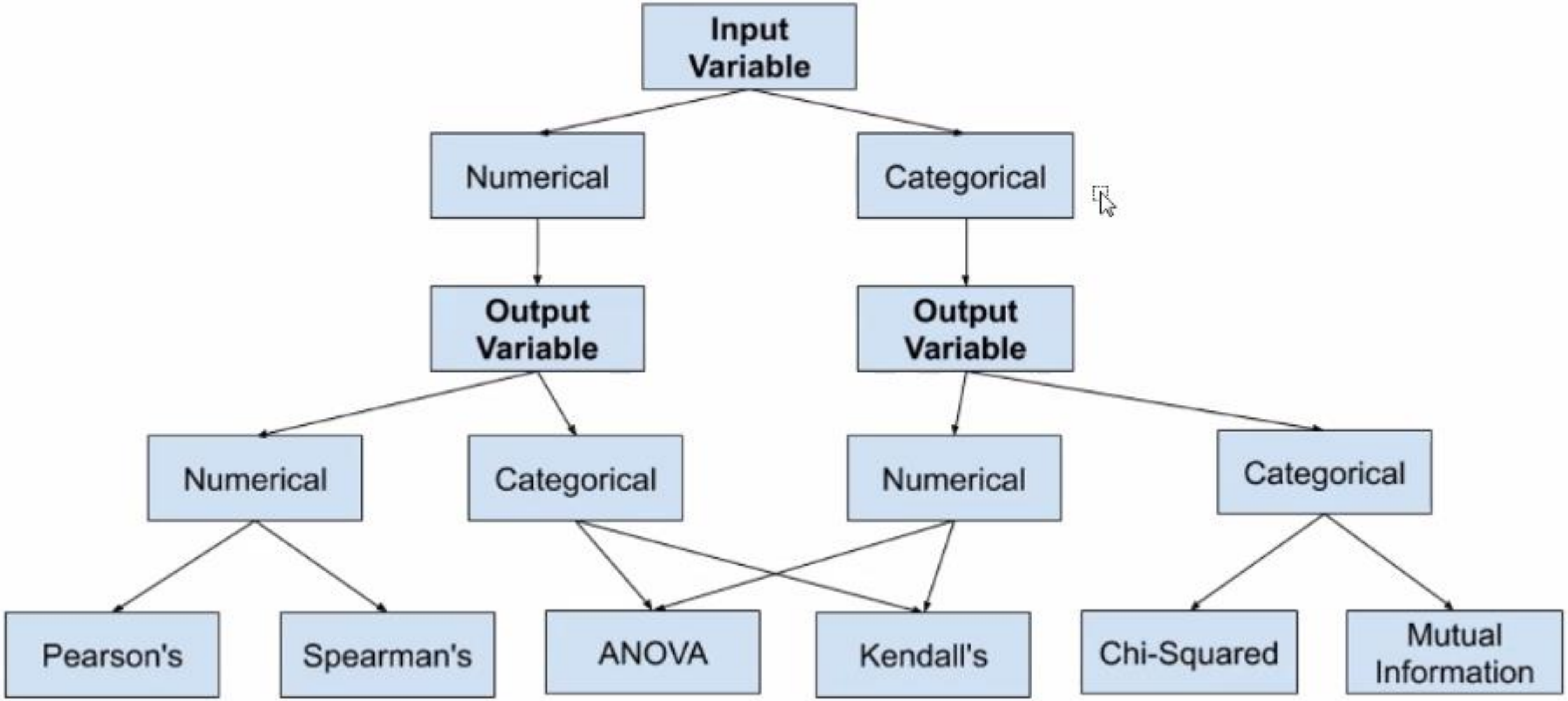
범주형 데이터 ☞ 빈도분석, 교차분석

- 명목척도 (Normal Scale)
  - ➢ 측정된 현상을 상호 배타적인 범주(category)로 수치를 부여한 척도
    (예시) 성별 (남자=1, 여자=2), 실험군 (대조군=1, 대조군=2), 혈액형 (A=1, B=2, AB=3, O=4) 등
  - ➢ 순서, 간격의 개념이나 가감승제의 수학적 연산기능을 가지지 못하는 척도
- 서열척도 (Ordinal Scale) = 순서척도, 순위척도
  - ➢ 명목척도의 기능 뿐만 아니라 각 범주간의 대소관계, 서열성에 관하여 수치를 부여한 척도
    (예시) 건강상태 (나쁨=1, 보통=2, 양호=3), 치료의 정도 (반응=1, 중간반응=2, 무반응=3) 등
  - ➢ 수학적 의미 : A > B, A < B , A = B

연속형 데이터 ☞ 기술통계, 평균비교, 회귀분석 등

- 등간척도 (Interval Scale) = 구간척도
  - ➢ 절대적 원점(Absolute zero)이 없으며, 대상이 갖는 양적인 정도의 차이에 따라 등 간격으로 수치를 부여한 척도
    (예시) 온도 (섭씨 0℃, 50℃, 100℃), 물가지수, 생산지수 등
  - ➢ 수학적으로 가감의 조작이 가능하지만, 승제의 조작은 불가능한 척도
- 비율척도 (Ratio Scale)
  - ➢ 절대적 원점이 존재하며, 비율계산이 가능한 수치를 부여한 척도
    (예시) 광고비, 판매량, 매출액, 무게, 가격, 소득 등
  - ➢ 수학적으로 가감승제의 조작이 모두 가능한 척도

How to Choose a Feature Selection Method

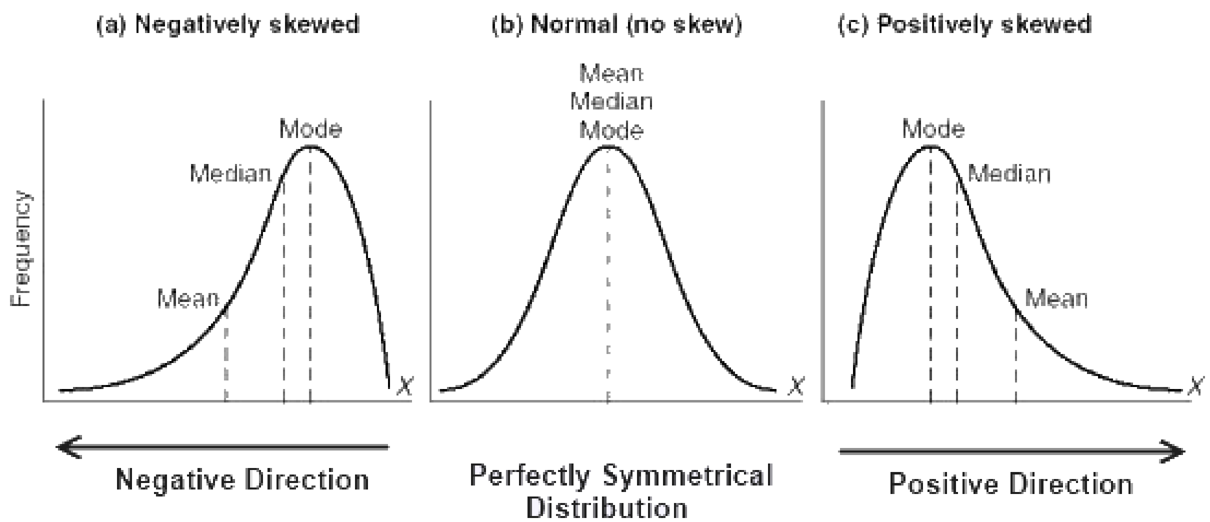


기초 확률 및 통계
02. 데이터의 지표

데이터의 중심

2 , 5 , 8 , 1 0 , 1 1 , 1 6 , 9 , 3 , 2 0 , 8

- Mean(산술평균) = 총합 / 총 n = 9.2
- Median(중앙값) = (9+8) / 2 = 8.5
- Mode(최빈값) = 8

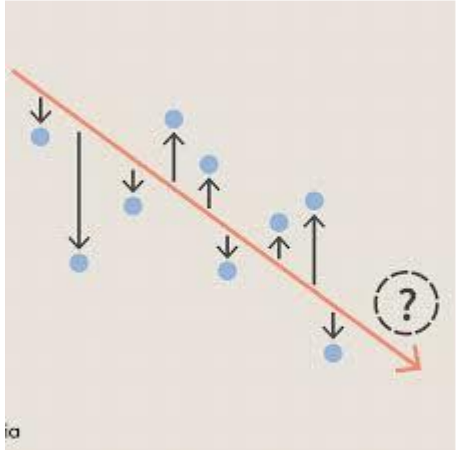


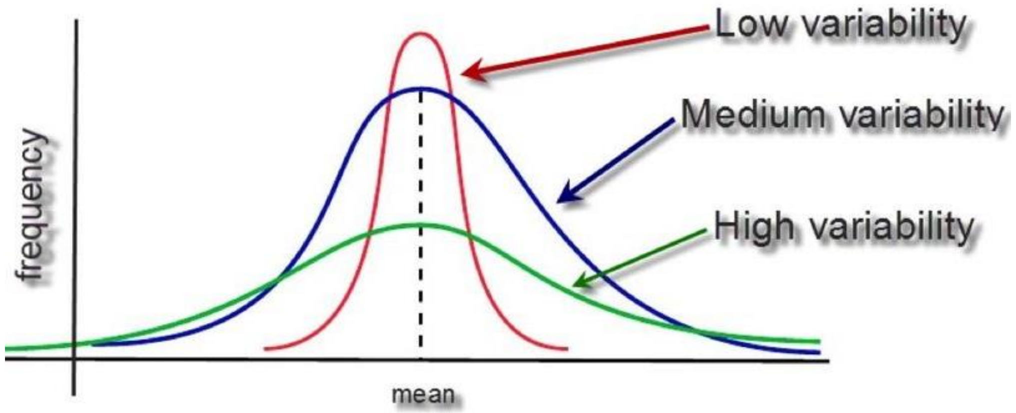
데이터의 분포

- 편차(deviation) : 평균과 관측값 간의 차
- 분산(variance) : 편차의 제곱 합 / n
- 표준편차(Standard deviation) : 분산의 루트값

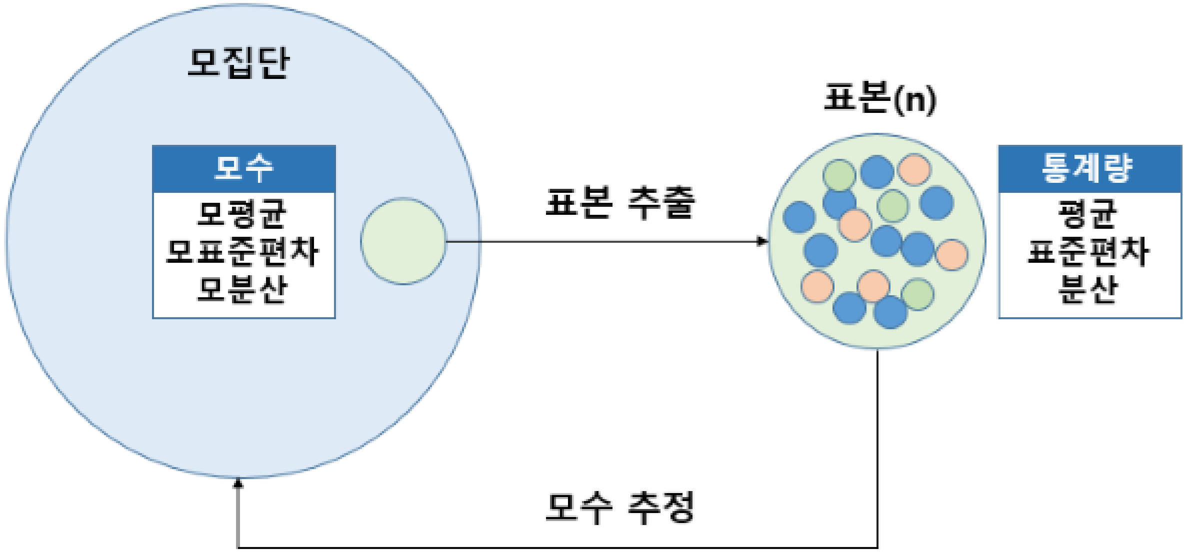
- 분산과 표준편차?

분산의 경우 편차의 제곱값으로 실제 값에서
동 떨어진 값이 나오기 때문에 근사값으로 만들기 위해
표준편차도 사용





기초 확률 및 통계
03. 모수와 샘플의 관계



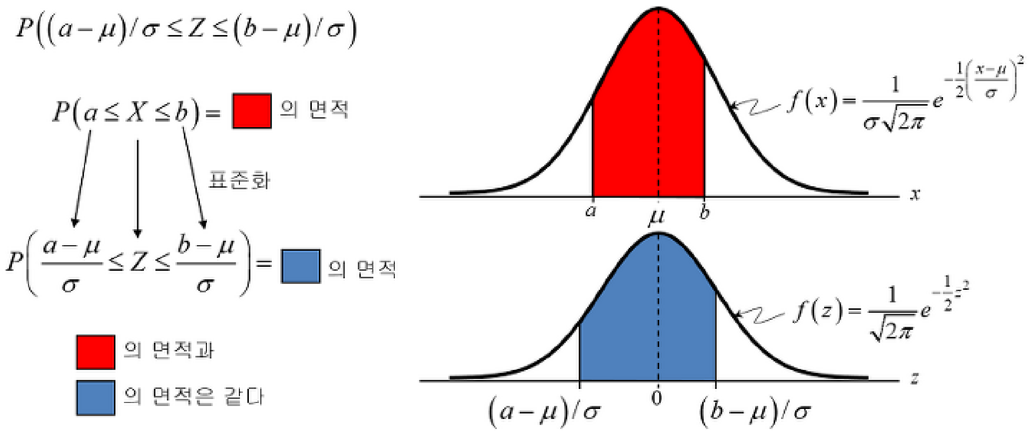
	Dataset ($X \in \mathbb{R}^N$)	Samples ($S \in \mathbb{R}^n$)
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	Biased: $\sigma_S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ Un-biased: $\sigma_{S'}^2 = \boxed{\frac{1}{\boldsymbol{n-1}}}\sum_{i=1}^{n}(x_i - \mu)^2$

Degree of freedom

표준정규분포

Standard Normal Distribution

$$P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \qquad \mu = 0 \qquad \sigma^2 = 1$$



정규화

Standardizing

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

Normal Approximation to Binomial Dist'n

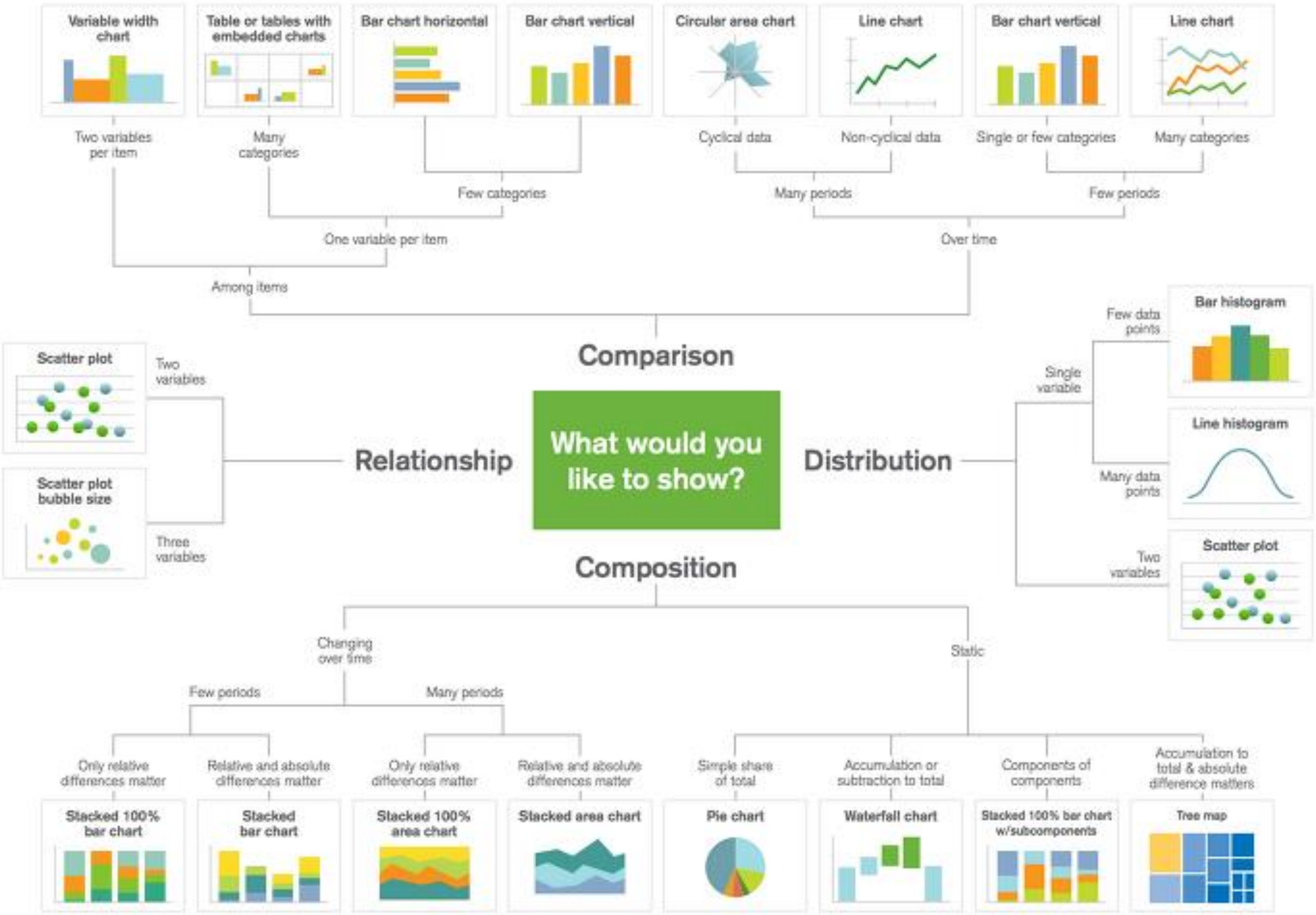
$$P(X \le x) = P(X \le x + 0.5) \approx P\left(Z \le \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

* The approximation is good for $np > 5$ and $n(1-p) > 5$

Normal Approximation to Poisson Dist'n

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \qquad E(X) = \lambda \quad V(X) = \lambda$$

* The approximation is good for $\lambda > 5$



Source: ©A. Abela, 2010. www.ExtremePresentation.com

기초 확률 및 통계

*. Probability & Random Variable

1. Event & Sample Space

Sample Space(샘플) = {Event(샘플값)}

2. Probability

일종의 Likelihood(가능성)의 집합, [0,1]

3. Conditional Probability(조건부 확률)

P(B|A) = P(A ∩ B) / P(A)

P(A ∩ B) = P(A)P(B|A)

4. Independence(독립조건)

두 확률이 독립

P(A|B) = P(A)

P(B|A) = P(B)

P(A ∩ B) = P(A) × P(B)

5. Baye's Theorem

3개 값으로 미지의 목표값 추정

유방암 발병률 : 0.1%

검사

검출력 : 99%

위양성 : 2%

검사결과 양성인데, 실제 양성일 확률?

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \Rightarrow p(B|A)p(A) + p(B|A^c)p(A^c)$$

양성을 양성으로 나올 확률 + 음성이 양성으로 나올 확률

p(b) : 검사결과 양성이 나올 확률

= p(양성결과 | 양성)p(양성일때) + p(양성결과|음성)p(음성일때) = 0.020079

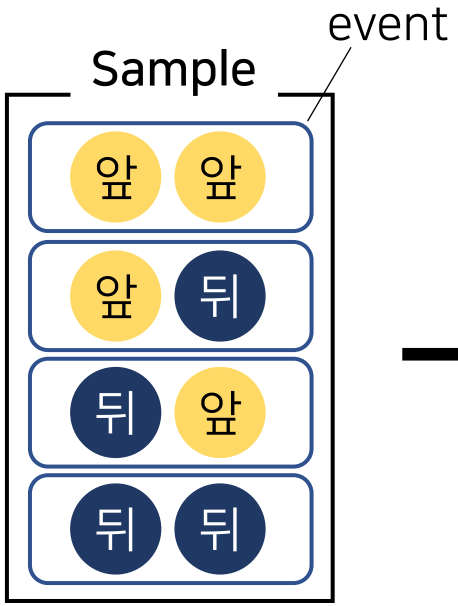
p(b|a) = 0.99

p(a) = 양성일때 = 0.001

p(a|b) = 0.049

사전확률 = 0.001 >> 사후확률 = 0.049(증가했음) | 사전확률 p(a) = 0.049 >> 사후확률 =0.718

p(b) = 0.06753



S = { 앞앞, 앞뒤, 뒤앞, 뒤뒤}

확률분포
Probability distribution

P(앞 = 0) = 1/4
P(앞 = 1) = 2/4
P(앞 = 2) = 1/4

X = Random variable

Discrete

확률질량함수
Probability Mass Functior

$(1)\ f(x_i) \geq 0$
$(2)\ \sum_{i=1}^{n} f(x_i) = 1$
$(3)\ f(x_i) = P(X = x_i)$

누적 확률 분포 함수
Cumulative Distribution Functions

$P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$

Continuous

확률질량함수
Probability Density Function

$(1)\ f(x_i) \geq 0$
$(2)\ \int_{-\infty}^{\infty} f(x)\,dx = 1$
$(3) P(a \leq X \leq b) = \int_a^b f(x)\,dx$

누적 확률 분포 함수
Cumulative Distribution Functions

$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du$

pdf from cdf

$$\frac{d}{dx}\int_{-\infty}^{x} f(u)\,du = f(x) = \frac{x}{dx}F(x)$$

Mean and Variance of a Discrete Random Variable

$$\mu = E(X) = \sum xf(x) = 1$$

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

- Expected Value

$$E[\mathrm{h(X)}] = \sum_x h(x)f(x)$$

$$IF\ h(x) = aX + b, \qquad \mathrm{E(aX+b) = aE(x) + b}$$

$$\mathrm{V(aX+b)} = a^2 V(x)$$

Mean and Variance of a Contimuous Random Variable

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (X-\mu)^2 f(x)\mathrm{dx} = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

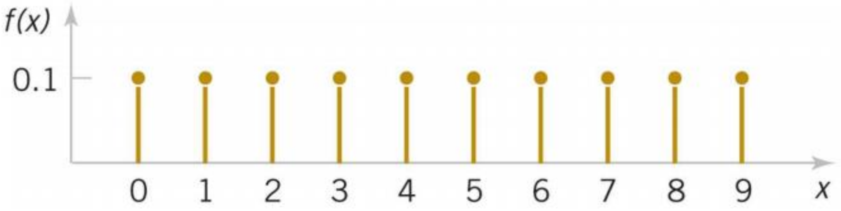
- Expected Value

$$E[\mathrm{h(X)}] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

이산 균일분포

Discrete Uniform Distribution

$$f(x_i) = 1/n$$



$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = \frac{(b-a+1)^2-1}{12}$$

포아송분포 = 범위안에서 Event가 일어난 횟수

Poisson Distribution

$$f(x) = \frac{e^{-\lambda T}(\lambda T)^x}{x!} \quad x = 0,1,2, \ldots$$

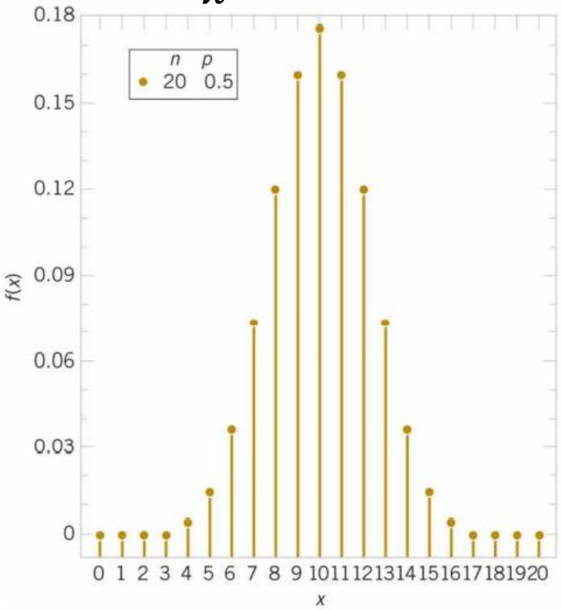
$$\mu = E(X) = \lambda T$$

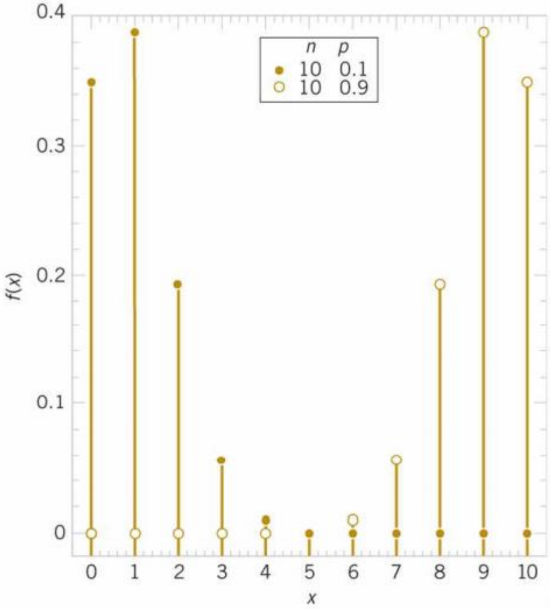
$$\sigma^2 = V(X) = \lambda T$$

이항분포 : 결과 = Y or N, 결과가 전체 횟수 중 몇번

Binomial Distribution

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0,1, \ldots, n$$





$$\mu = E(X) = np$$

$$\sigma^2 = V(X) = np(1-p)$$

*베르누이 시행

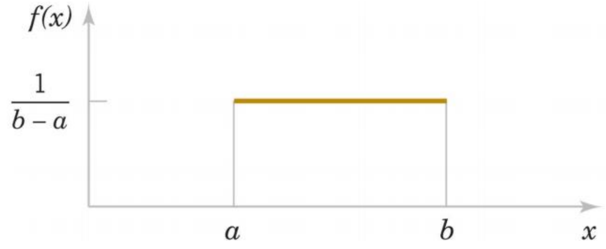
Bernoulli Trial

$$P\{X = x\} \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

연속 균일분포

Continuous Uniform Distribution

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b$$



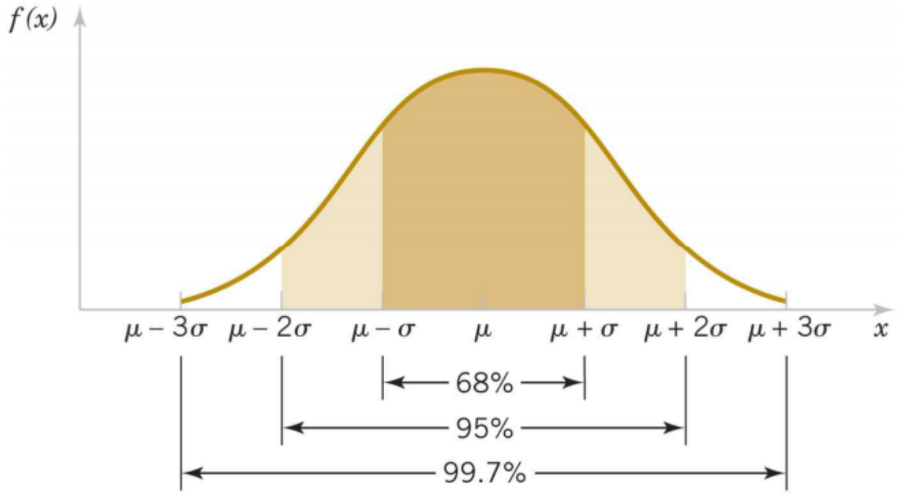
$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$

정규분포

Normal Distribution

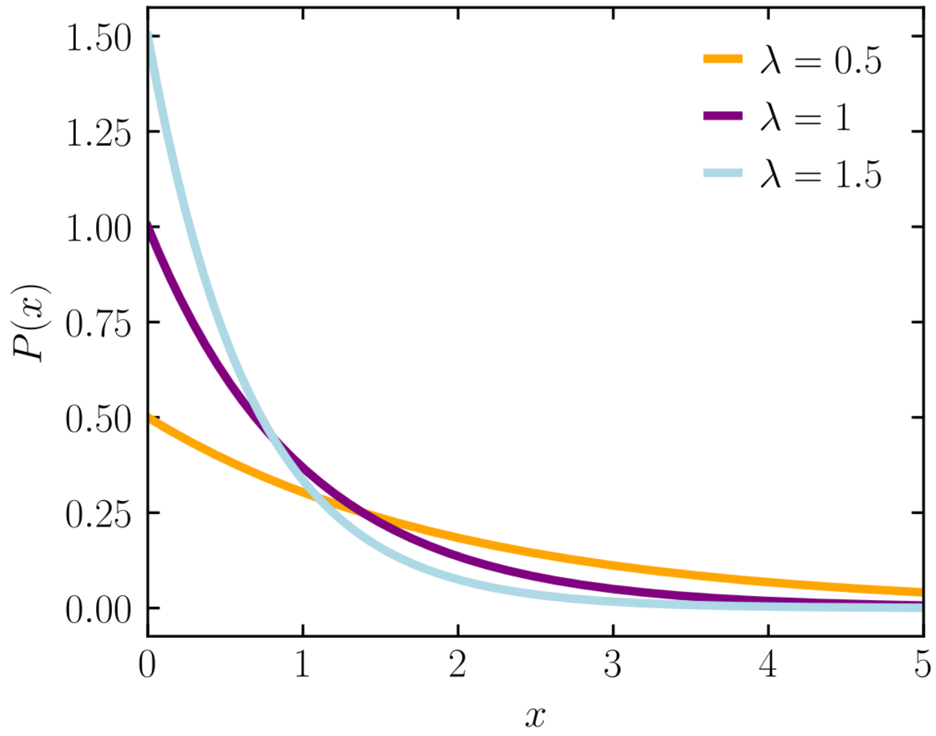
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad -\infty \le x \le \infty \qquad \mu = E(X) \quad \sigma^2 = V(X)$$



지수분포 : 일정 시간동안 발생하는 사건의 횟수가 푸아송 분포를 따른다면, 다음 사건이 일어날 때까지 대기 시간

Exponential Distribution

$$f(x) = \lambda e^{-\lambda x} \quad 0 \leq x < \infty$$



$$\mu = E(X) = \frac{1}{\lambda}$$

$$\sigma^2 = V(X) = \frac{1}{\lambda^2}$$