

Probability and Statistics Review

2025. 8

Reference:

- Steven Skiena, The Data Science Design Manual, Springer, 2017
- Statistics and Managers Using Microsoft Excel, 4e, Prentice-Hall, inc., 2004.
- https://www.cs.cmu.edu/~epxing/Class.../Probability_Review.ppt
- ChatGPT, Gemini

Contents

- 확률 및 통계의 기초 (Fundamentals of Probability and Statistics)
 - 확률과 통계의 차이
 - 확률의 기본 개념
 - 조건부 확률과 베이즈 정리
- 기술 통계 (Descriptive Statistics)
 - 기술 통계의 정의
 - 중심 경향성 측정치
 - 변동성 측정치 및 분산의 이해
- 확률 분포 (Probability Distributions)
 - 확률 분포의 종류 (PMF, PDF, CDF)
 - 주요 이산 확률 분포
 - 주요 연속 확률 분포 (정규분포 등)
 - 중심 극한 정리
- 추론 통계 및 가설 검정 (Inferential Statistics and Hypothesis Testing)
 - 추론 통계의 정의 및 목적
 - 가설 검정의 핵심 개념 (가설, P-값 등)
 - 주요 가설 검정 방법 (t-test, F-test, Chi-squared)
 - 추정과 신뢰 구간
- 데이터 과학에서의 활용 (Applications in Data Science)
 - 계적 방법의 활용 예시 (피처 선택 등)

Fundamentals of Probability and Statistics

Probability and Statistics

- **Probability**

- deals with predicting the likelihood of **future** events
- theoretical branch of mathematics on the consequences of definitions
- For the dice game, “each face will come up with probability $1/6$.”

- **Statistics**

- analyzes the frequency of **past** events
- applied mathematics trying to make sense of real-world observations
- For dice game, “I will watch a while, and keep track of how often each number comes up.”

Probability

- Experiment: a procedure which yields one of a set of possible outcomes
- Sample space S : set of possible outcomes s of an experiment
- Event: specified subset of the outcomes of an experiment
- **Probability** $p(s)$ of an outcome s : *a number with:*
 - $0 \leq p(s) \leq 1$
 - $\sum_{s \in S} p(s) = 1$
- **Random variable**(확률 변수) V : numerical function(assignment) on the outcomes of a probability space
- Expected value (**기대값**) E of a random variable V on sample space S :

$$E(V) = \sum_{s \in S} p(s) \cdot V(s)$$

Probability (example)

- Experiment: tossing two six-sided dice
- Sample space S: 36 possible outcomes, namely
 - $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$
 $(3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$
 $(5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$
- Event: the event that the sum of the dice equals 7 or 11
 - $E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1), (5,6),(6,5)\}$
- Probability of the event: $p(E) = 8/36$
- Random variable V: $V(s) = 1, 2, 3, 4, 5, 6$ for each sample s
 - $P(V=7) = 1/6, p(V=12) = 1/36$
- Expected value E:
 - $E(V) = 1/6(1) + 1/6(2) + 1/6(3) + 1/6(4) + 1/6(5) + 1/6(6) = 21/6$

Compound Events and Independence

- Suppose half my students are female (event A), and Half my students are above median (event B). **What is the probability a student is both A & B?**
- Events A and B are **independent** iff

$$P(A \cap B) = P(A) \times P(B)$$

- Independence(독립성) (zero correlation) is good to simplify calculations, but bad for prediction (no information shared between events A and B)

Conditional Probability

- The conditional probability $P(A|B)$ is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

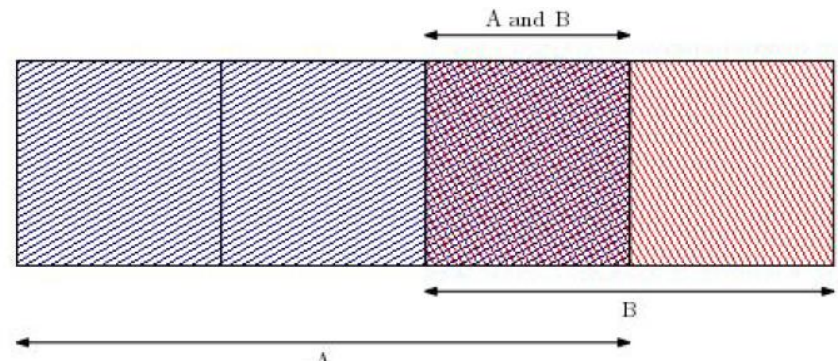
- Conditional probability get interesting only when events are ***not*** independent, otherwise:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes Theorem

- Bayes' theorem is an essential tool which reverses the direction of the dependences:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$



$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- $P(B|A)$: 사후 확률(Posterior probability). 증거 A를 관찰한 후 가설 B가 옳을 확률.
- $P(B)$: 사전 확률(Prior probability). 증거를 관찰하기 전, 가설 B가 옳을 초기 확률.
- $P(A|B)$: 우도(Likelihood). 가설 B가 옳다는 가정 하에 증거 A가 관찰될 확률.
- $P(A)$: 증거 확률(Evidence). 증거 A가 관찰될 확률.

Bayes Theorem

- 베이즈 정리의 핵심: 믿음(확률)을 업데이트하는 과정
 - 어떤 사건에 대해 가지고 있던 초기 믿음(사전 확률)을 새로운 정보(증거)를 통해 더 합리적이고 정확한 확률(사후 확률)로 수정한다.
- (예) 독감 검사 예시
 - 사전 확률: 내가 독감에 걸렸을 초기 확률은 매우 낮다. 예를 들어, 1만 명 중 1명에게만 독감이 발생한다고 가정한다
 - 새로운 증거: 독감 검사에서 양성 반응이 나왔다
 - 베이즈 정리의 역할: "검사에서 양성 반응이 나왔을 때, 내가 실제로 독감에 걸렸을 확률은 얼마인가?"라는 질문에 대한 답.
 - 이 정리를 통해 검사의 정확도($P(T=\text{positive} | F=\text{yes}) = 0.99$) 와 같은 정보를 활용하여 초기확률을 업데이트한다.

Bayes Theorem : Example

- A straightforward manipulation of probabilities:

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{\sum_{x_1} p(X_2|x_1) p(x_1)}$$

- Example:

- 내가 유행성 독감(1/10,000명에게만 발생)에 걸렸는지 알고 싶다. 그런데 독감에 대한 “정확한” 검사 방법이 있는데, 이 방법에 의하면 독감에 걸린 사람은 99% 정확도로 잘 판단하며, 걸리지 않은 사람도 99% 의 확률로 걸리지 않았음을 잘 판단할 수 있다. 내가 의사에게 가서 이 테스트에서 양성 반응을 보이면 독감에 걸릴 확률은 얼마일까?

- Two Random Variables, $T \in \{\text{positive, negative}\}$ and $F \in \{\text{yes, no}\}$

$$p(F=\text{yes}) = 1/10,000$$

$$p(F=\text{no}) = 9999/10,000$$

$$p(T=\text{positive} | F=\text{yes}) = p(T=\text{negative} | F=\text{no}) = .99$$

$$p(F=\text{yes} | T=\text{positive}) = ?$$

$$P(T=\text{positive}) = P(T=\text{positive} | F=\text{yes})P(F=\text{yes}) + P(T=\text{positive} | F=\text{no}) \times P(F=\text{no})$$

$$P(T=\text{positive} | F=\text{no}) = 1 - P(T=\text{negative} | F=\text{no}) = 1 - 0.99 = 0.01$$

Descriptive Statistics

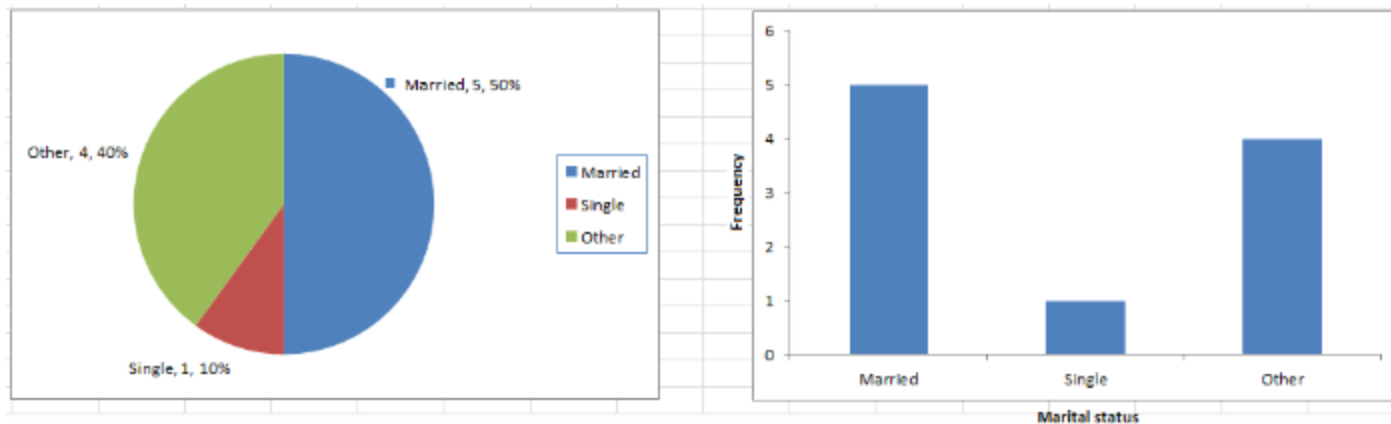
(기술통계)

Descriptive Statistics

- **Descriptive statistics** provide ways to capture the properties of a given data set/sample.
 - **Central tendency (중심경향성) measures** describe the center around which the data is distributed.
 - **Variation or variability (변동성) measures** describe data spread.

Uni-variate Descriptive Statistics

- Different ways you can describe patterns found in uni-variate data include central tendency : mean, mode and median and dispersion: range, variance, maximum, minimum, quartiles , and standard deviation.



Pie chart [left] & Bar chart [right] of Marital status from loan applicants table.

Centrality Measures

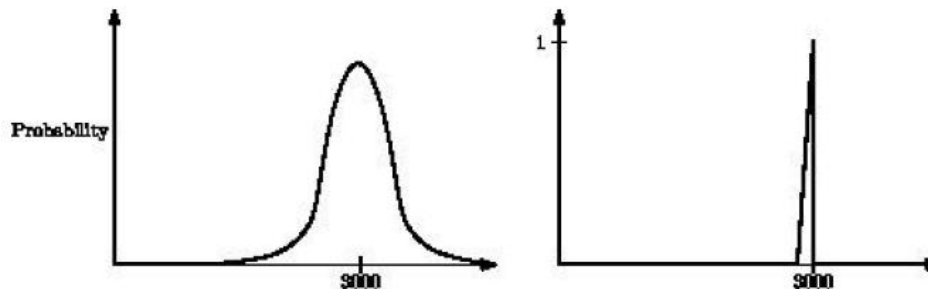
- Centrality measure
 - Mean (평균): arithmetic, geometric, harmonic
 - Median (중앙값)
 - Mode (최빈값)
- **Mean** is meaningful for symmetric distributions without outliers: e.g. height and weight.
- **Median** is better for skewed distributions or data with outliers: e.g. wealth and income.
- Bill Gates adds \$250 to the mean per capita wealth but nothing to the median.

Variability Measures

- The **variance** is the square of the **standard deviation** (σ).

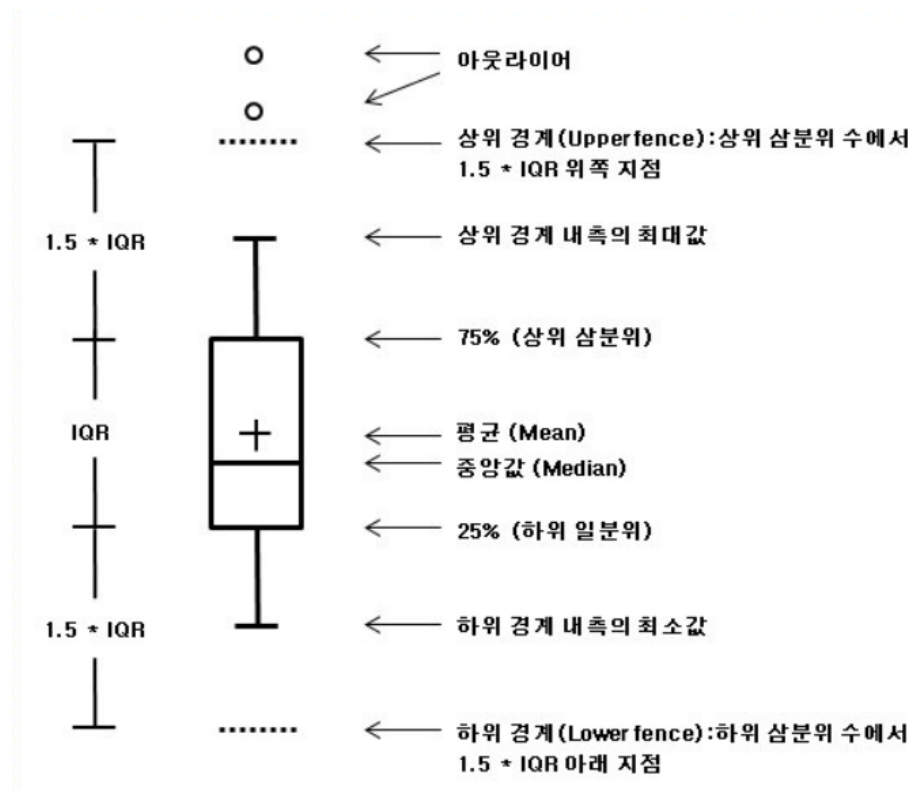
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

- Distributions with the same mean can look very different. But together, the mean and standard deviation fairly well characterize any distribution.



Boxplots

- Box plots are especially useful for indicating whether a distribution is **skewed** and whether there are potential unusual observations (**outliers**) in the data set.



Why We Divide by $n-1$?

- 왜 표본 분산(Sample Variance)을 계산할 때 n 이 아닌 $n-1$ 로 나누는가?
 - 표본 분산의 목적은 표본 자체의 변동성을 기술하는 것이 아니라, **모집단 분산(Population Variance)**을 추정하는 것이다. 우리가 모집단 평균(μ)을 모르기 때문에 표본 평균(\bar{x})을 사용하게 되는데, 이로 인해 실제 분산보다 작은 값(편향된 값)이 계산된다.
 - $n-1$ 로 나누는 것은 이러한 편향을 보정하여, 모집단 분산에 대한 비편향 추정량(unbiased estimator)을 얻기 위함이다

N-1 Correction: An Experiment

(ex) 실험: 모집단($\mu=170, \sigma=20$), 작은 표본(예: $n=10$)을 여러 번 추출

- estimating population mean and variance by sample mean and sample variance (모평균과 모분산의 추정)

```
1 z = np.random.normal(170, 20, size=1000)    # normal (mu=170, sigma=20)
2
3 n_sample = 10      # sample size
4 n_iter=1000        # number of sampling iteration
5
6 list_mean, list_sig_biased, list_sig_unbiased = [], [], []
7
8 for i in range(n_iter):
9     sample = np.random.choice(z, n_sample)
10    m = sample.mean()
11    sig_biased = ((sample - m)**2).sum() / n_sample
12    sig_unbiased = ((sample - m)**2).sum() / (n_sample-1)
13
14    list_mean.append(m)
15    list_sig_biased.append(np.sqrt(sig_biased))
16    list_sig_unbiased.append(np.sqrt(sig_unbiased))
17
18 print(sum(list_mean)/n_iter)
19 print(sum(list_sig_biased)/n_iter, sum(list_sig_unbiased)/n_iter)
```



169.42363026579284

18.782363563749385 19.7983495676017

Bi-variate Descriptive Statistics

- 두 변수 간의 관계를 분석하는 것으로, 데이터에 숨겨진 패턴을 파악하는 데 사용된다.
- 산점도(Scatter Plots): 두 연속형 변수(continuous variables)의 관계를 시각화하는 가장 간단한 방법.
- 공분산(Covariance): 두 변수가 함께 어떻게 변하는지를 측정하는 지표다. 공분산은 양수, 음수, 0의 값을 가질 수 있다.
- 상관관계(Correlation): 공분산을 표준화하여 두 변수 간의 선형 관계의 강도와 방향을 측정하는 지표다. 상관계수(r)는 -1부터 1까지의 값을 갖는다.

Covariance and Correlation

- Variance

$$\text{Var}(X) = E[(X - \mu)^2]$$

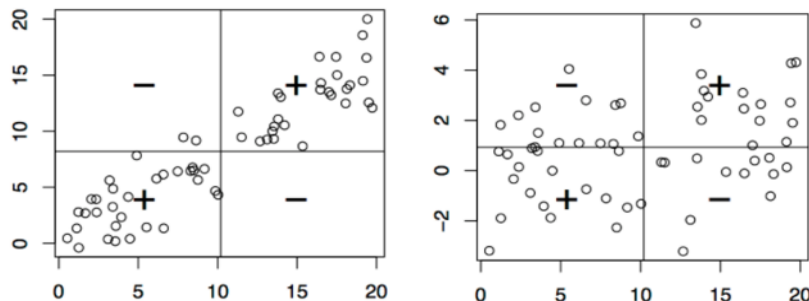
- Covariance

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$$

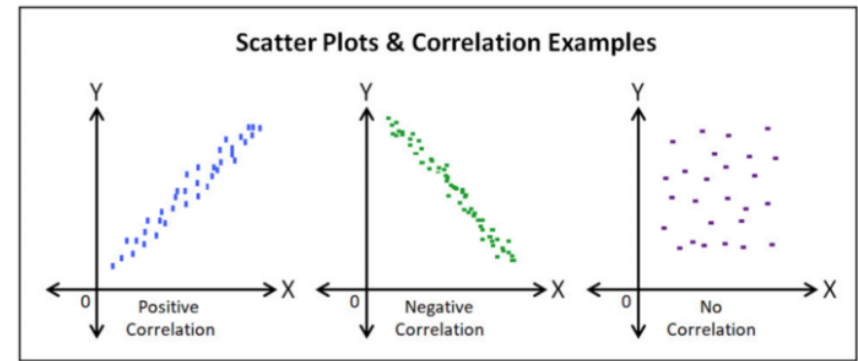
$$\text{cov}(X, X) = \text{var}(X) \equiv \sigma^2(X) \equiv \sigma_X^2$$

- Correlation

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$



positive and negative covariance



Positive correlation ($r > 0$), Negative correlation ($r < 0$), No correlation ($r = 0$)

Probability Distribution

Probability Distribution

- 확률분포
 - 확률변수 x 가 특정한 값을 가질 확률 정보
- Probability Density Function, PDF (확률밀도함수)
 - 연속 확률변수에서 확률변수의 분포
 - (ex) 키, 나이
- Probability Mass Function, PMF (확률질량함수)
 - 이산확률변수에서 특정값에 대한 확률
 - (ex) 주사위, 동전
- Cumulative Distribution Function, CDF (누적분포함수)

$$F_X(x) = P(X \leq x)$$

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

Discrete Distribution

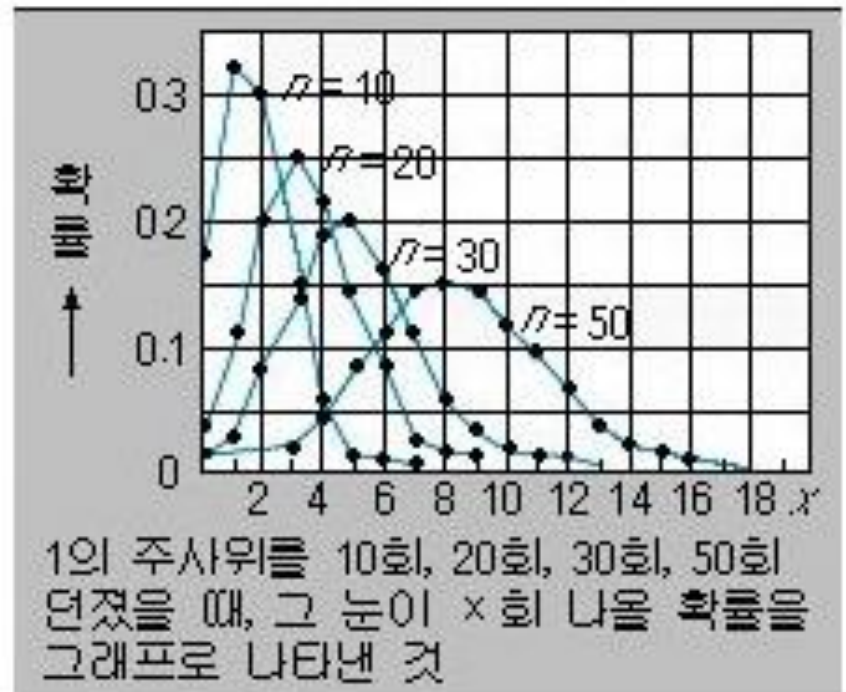
- Bernoulli Distribution (베르누이 분포): X

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1 - p) \end{cases}$$

- Binomial Distribution (이항분포): 여러 번의 연속 실험의 확률

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

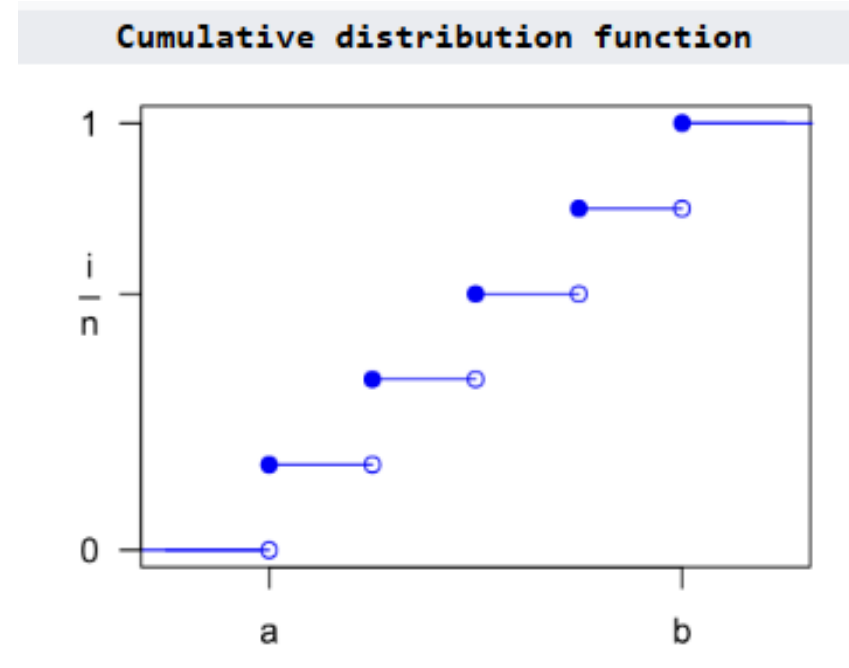
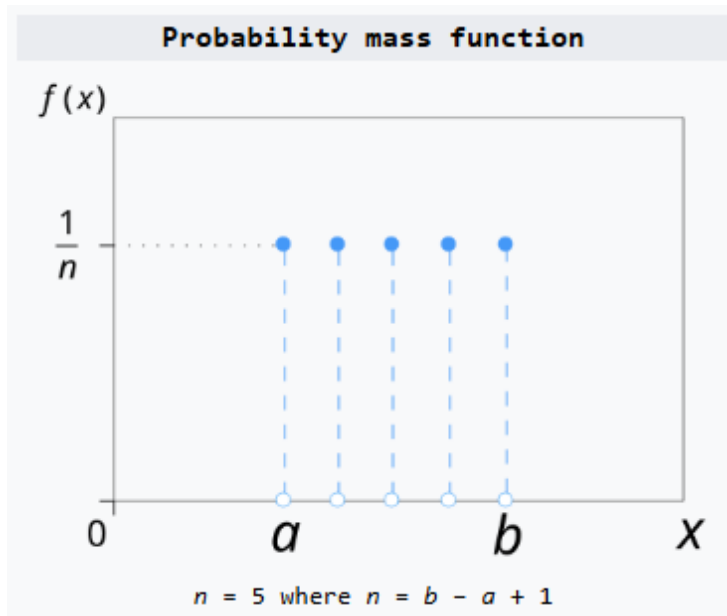
- p 가 0이나 1에 가깝지 않고 n 이 충분히 크면 이항분포는 정규분포(가우스분포)에 가까워지며, p 가 $1/2$ 에 가까워짐에 따라 그래프는 좌우대칭인 산 모양 곡선이 된다.



Discrete Distribution

- Discrete Uniform Distribution with parameter n:

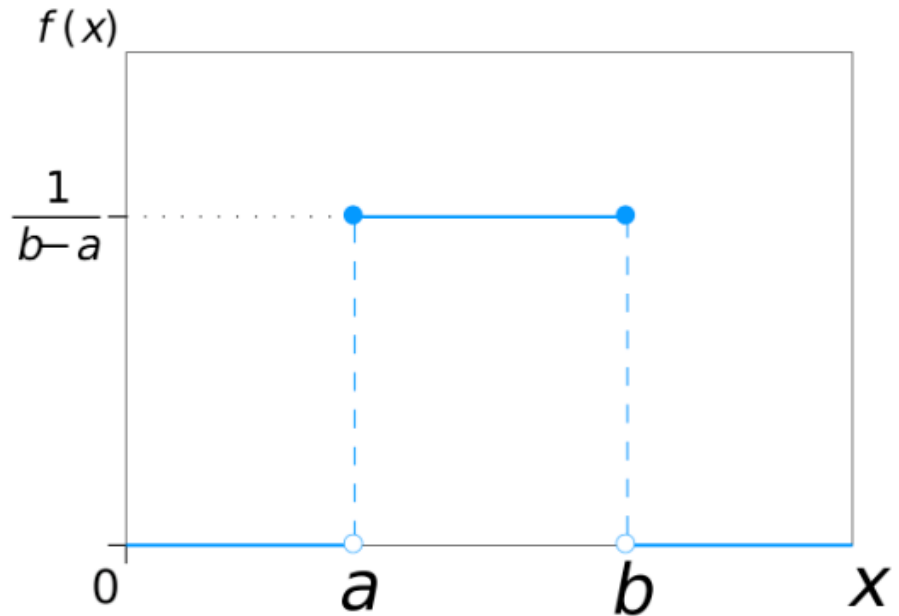
$$U\{a,b\} = 1/n \text{ for } k \in \{a, a+1, \dots, b-1, b\}$$



Expectation of Discrete Uniform Distribution: $E(X) = \frac{n+1}{2}$

Continuous Distribution

- Uniform Distribution(균일분포)



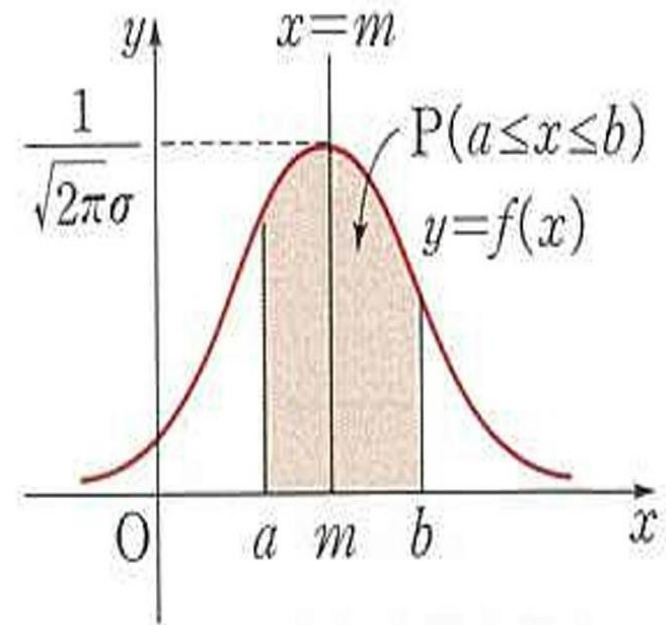
Continuous Distribution

- Normal or Gaussian Distribution (정규분포)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (x \text{는 모든 실수})$$

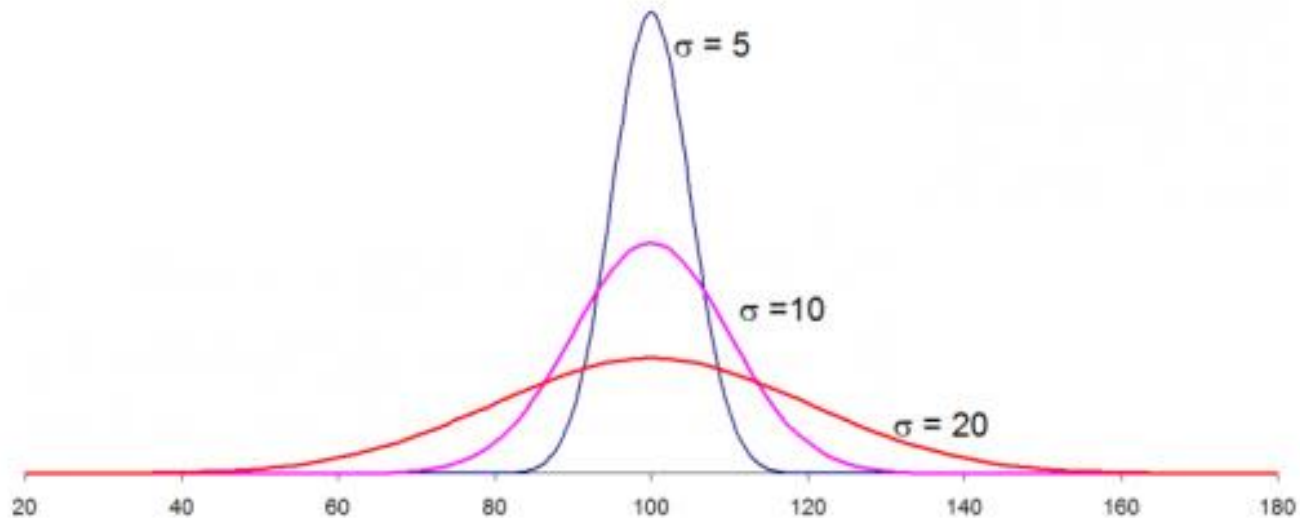
– 확률 $P(a \leq X \leq b) =$

$$\int_a^b f(x) dx = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$



Normal Distribution

- Standard Deviation (표준편차)



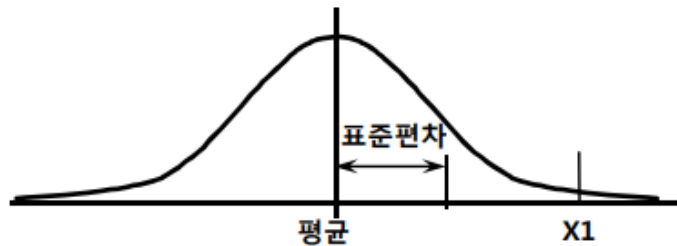
Three different data distributions with same mean (100) and different standard deviation (5,10,20)

Normal Distribution

- 표준정규분포 (Standard Normal Distribution)

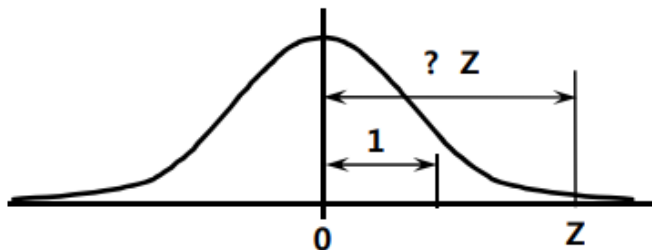
– 정규분포(평균 μ , 분산 σ^2)

확률변수 X 는 $X \sim N(\mu, \sigma^2)$



– 표준정규분포(평균0, 표준편차1)

확률변수 Z 은 $Z \sim N(0,1)$



정규분포

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



$$Z_i = \frac{x_i - \mu}{\sigma}$$

Z 변환

or

Standard Scaling



표준정규분포

$$Z \sim N(0, 1^2)$$

Normal Distributions

- Gaussian distribution (or Normal distribution)
 - a type of continuous probability distribution for a real-valued random variable
 - Probability density function (PDF), $N(x; \mu, \sigma^2)$:

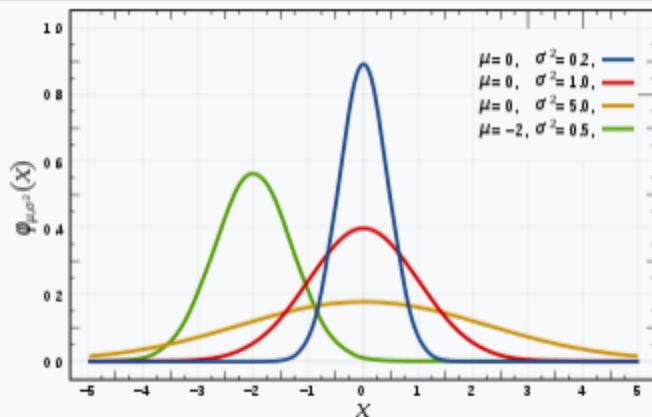
$$p(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \equiv N(x; \mu, \sigma^2)$$

, where μ is mean (or expectation) and σ is its standard deviation (σ^2 variance)

population mean

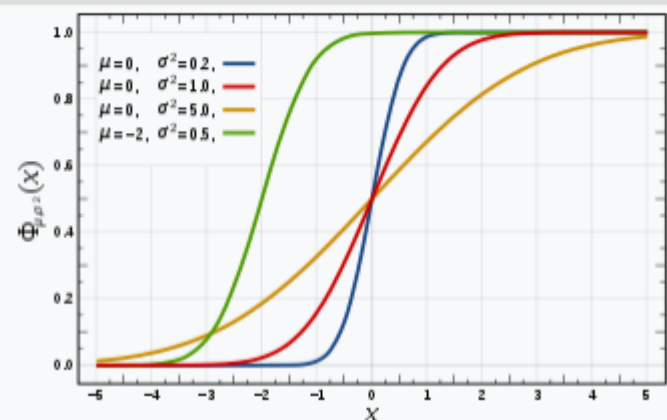
population variance

Probability density function



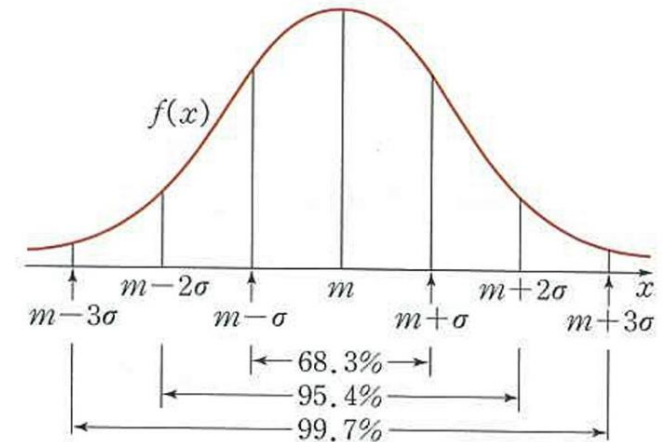
The red curve is the *standard normal distribution*

Cumulative distribution function



Normal Distributions

- Normal Distribution (continued)
 - Bell Shaped
 - Symmetrical
 - Mean, Median and Mode are Equal
 - The random variable X has an infinite theoretical range: $+\infty$ to $-\infty$
 - Often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.



Mean = Medium = Mode

Normal Distributions

- Standardized Normal Distribution
 - Any normal distribution can be transformed into the **Standardized normal distribution** (aka Z-distribution) (표준정규분포)

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

- Need to transform X units into Z units: (**Z-transform** or **Standard scaling**)

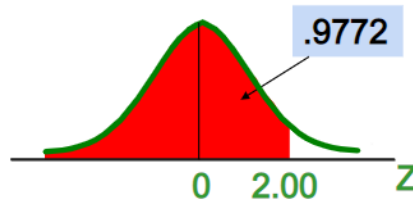
$$Z = \frac{X - \mu}{\sigma}$$

- Standardized normal table

- Gives the probability less than a desired value for Z

Example:

$$P(Z < 2.00) = .9772$$



row: value of Z to the first decimal point

column: value of Z to the second decimal point

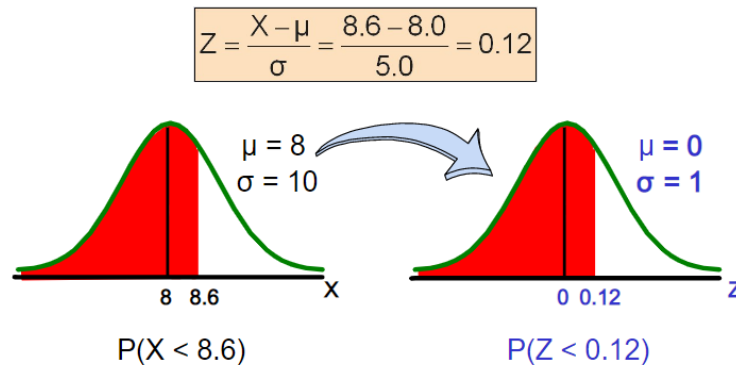
Z	0.00	0.01	0.02 ...
0.0			
0.1			
⋮			
2.0			

The value .9772 is located at the intersection of the row for 2.0 and the column for 0.00.

Normal Distributions

- Finding Normal Probability

- Suppose X is normal with mean 8.0 and standard deviation 5.0. How to find $P(X < 8.6)$?

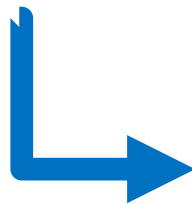


Statistics for Managers Using Microsoft Excel, 4e © 2004 Prentice-Hall, Inc.

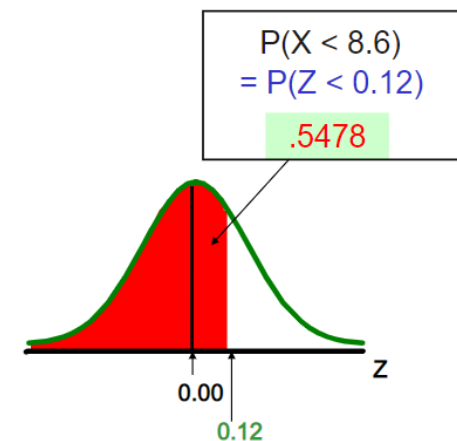
Chap 6-25

```
1 stats.norm(8, 5).cdf(8.6)
0.5477584260205838

1 # standardize it
2 z = (8.6 - 8.0)/5.0
3 stats.norm(0, 1).cdf(z)
0.5477584260205838
```



Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

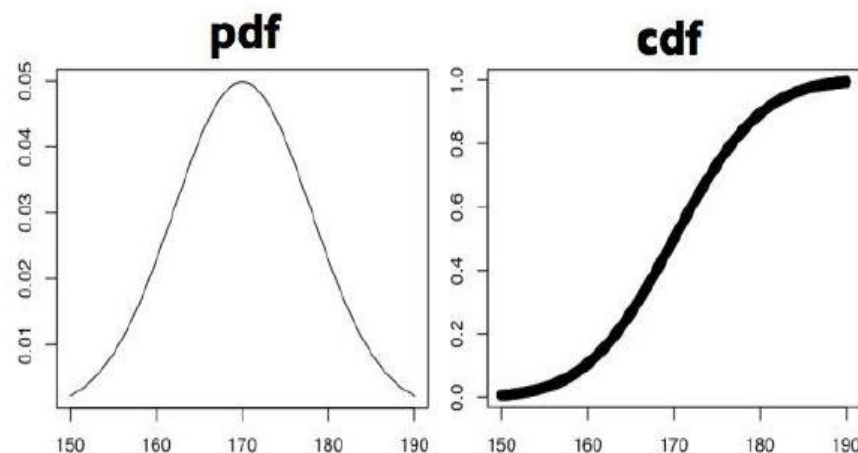


Cumulative Distribution Function

- The CDF is the running sum of the pdf:

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

- The PDF and CDF contain exactly the same information, one being the integral / derivative of the other.



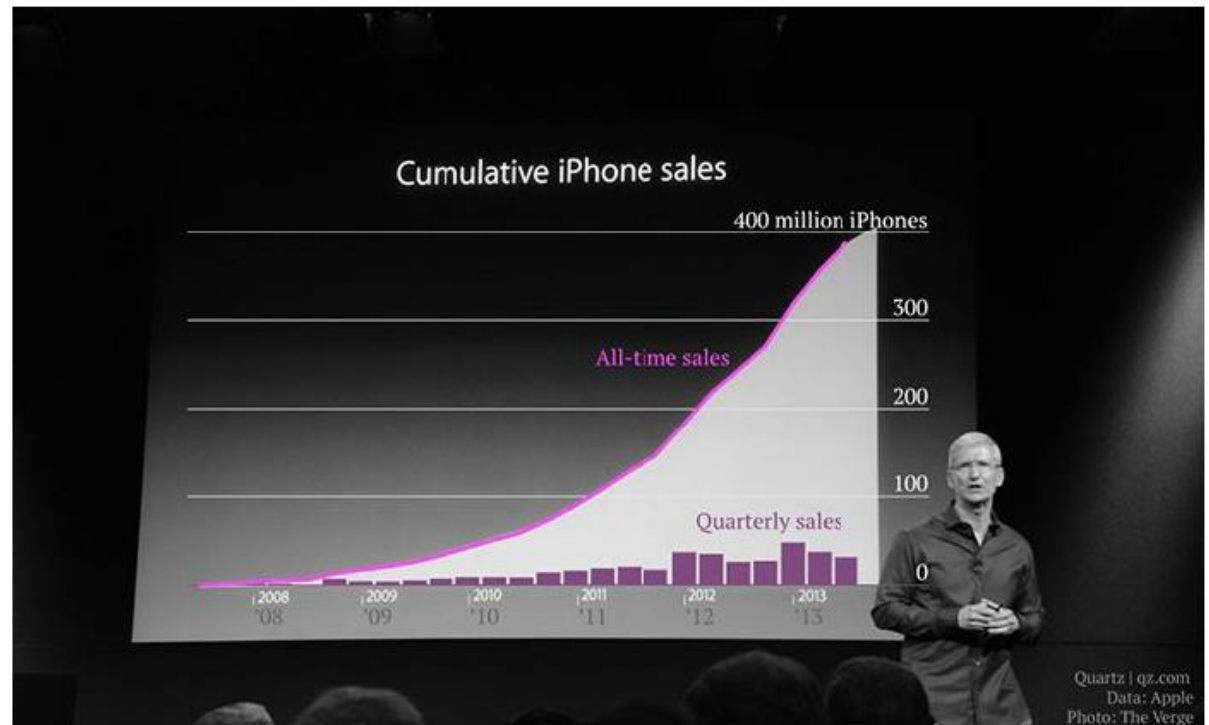
Visualizing Cumulative Distributions

- Apple iPhone sales have been exploding, right?



How explosive is that growth, really?

- Cumulative distributions present a misleading view of growth rate.
 - The incremental change is the derivative of this function, which is hard to visualize



Central Limit Theorem (CLT)

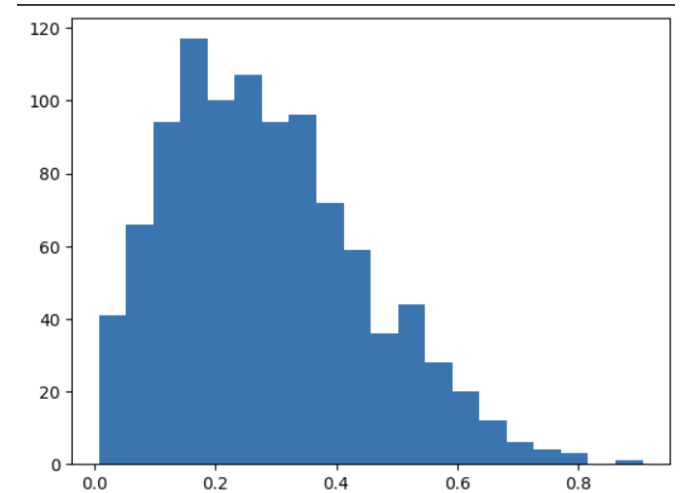
- **정의:** 모집단의 분포가 정규분포가 아니더라도, **표본의 크기(n)가 충분히 크면(보통 $n \geq 30$)**, **표본 평균의 분포**는 정규분포에 가까워진다는 정리.
- **중요성:** 이 정리는 우리가 모집단의 분포를 몰라도 통계적 추론을 가능하게 해준다. (통계적 검정(statistical testing)을 통해 **모집단의 평균**에 대해 유의미한 결론을 내리는 과정)
- **속성:** 표본 평균의 분포는 모집단 평균과 동일한 평균을 가지며, 표준편차는 σ / \sqrt{n} (분산: σ^2/n)이 된다.
- 즉, 우리는 CLT 덕분에 원래 데이터의 분포를 몰라도, 충분히 큰 표본을 사용해 정규성 가정을 기반으로 한 z-test나 t-test 같은 가설 검정을 수행할 수 있다.
- **실제적용;** 데이터 과학자는 모집단에 대해 알고 싶은 경우, **하나의 충분히 큰 표본(one sufficiently large sample)**만 추출한다. 이 하나의 표본 데이터를 바탕으로 표본 평균, 표본 표준편차 등을 계산하고, CLT 덕분에 이 표본 평균이 정규분포를 따르는 표본 분포에서 추출되었다고 가정하며, 이 가정 위에서 가설 검정(Hypothesis Testing)이나 신뢰구간(Confidence Interval)과 같은 추론 통계(Inferential Statistics)를 수행한다. (우리의 관심사는 평균)

CLT and Data Science

- CLT(중심 극한 정리)는 데이터 과학에서 추론 통계(Inferential Statistics)의 핵심
- 주요 활용 분야
 - **가설 검정 및 통계적 검정(Hypothesis Testing and Statistical Tests)**: 많은 통계적 검정 방법(예: Z-test, t-test)은 표본 평균의 분포가 정규분포(Normal Distribution)를 따른다는 가정을 전제로 한다. CLT는 원래 모집단의 분포가 무엇이든 관계없이, 표본의 크기(n)가 충분히 크면 표본 평균의 분포가 정규분포에 가까워진다고 보장한다. **이 덕분에 우리는 다양한 데이터에 대해 통계적 가설 검정을 적용할 수 있다.**
 - **신뢰 구간(Confidence Intervals)**: 신뢰 구간을 계산하여 예측의 불확실성을 나타내는 것도 데이터 과학의 중요한 부분이다. CLT는 표본 평균의 분포가 정규분포를 따른다는 가정을 정당화해주므로, 우리가 계산한 신뢰 구간이 통계적으로 유효함을 보장한다. 이는 모델의 예측 결과나 분석 인사이트를 제시할 때 신뢰도를 함께 전달하는 데 필수적이다.
 - **모집단에 대한 추론(Inference about the Population)**: 데이터 과학은 대부분 전체 모집단이 아닌, 표본(sample) 데이터만을 다룬다. CLT는 우리가 가진 작은 표본을 통해 더 큰 모집단의 특성을 추론하는 데 필요한 이론적 근거를 제공한다. 즉, 표본 데이터에서 얻은 결론을 모집단 전체에 일반화할 수 있는 통계적 다리 역할을 한다.

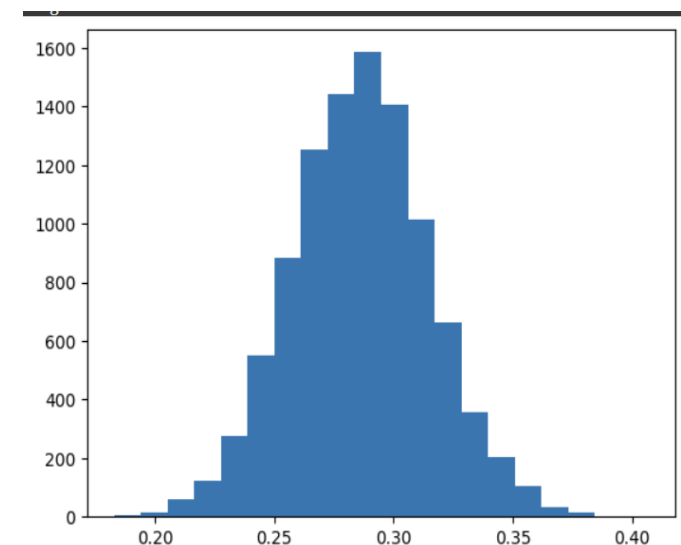
CLT: An Experiment

```
1 # let's try beta distribution for the population
2 z = np.random.beta(a=2, b=5, size=1000)
3 plt.hist(z, bins=20)
4 print(z.mean(), z.std()) # not normal
5 plt.show()
```



```
1 # now, sampling distribution is
2 n_sample = 30 # sample size
3 n_iter=10000 # number of sampling iteration
4 sample_means=[]
5 for i in range(n_iter):
6     sample = np.random.choice(z, n_sample)
7     sample_means.append(sample.mean())
```

```
1 plt.hist(sample_means, bins=20)
2 plt.figure()
```



Some common distributions

- Poisson Distribution:

- 평균 λ 회 발생하는 이벤트가 k 번 일어날 확률

- PMF:

$$f(k; \lambda) = \Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- k is the number of occurrences ($k = 0, 1, 2, \dots$)

- e is Euler's number ($e = 2.71828\dots$)

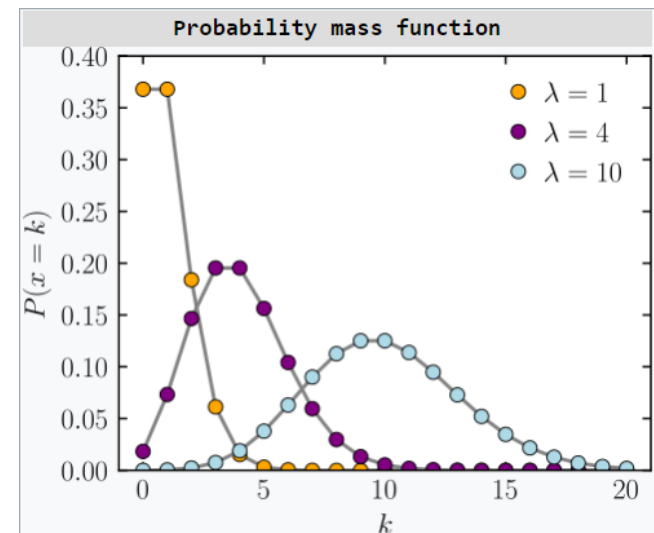
- $!$ is the factorial function.

- positive real number λ

- that is, λ : average number of events:

- 이산 확률 분포의 한 종류로, 정해진 시간이나 공간 안에서 어떤 사건이 평균적으로 λ 번 발생할 때, 그 사건이 k 번 발생할 확률을 나타낸다.
 - 시행 횟수(n)가 매우 크고, 발생 확률(p)이 매우 작을 때 이항 분포를 근사하는 데 사용된다. 포아송 분포는 n 의 크기를 알지 못해도 적용할 수 있다는 특징이 있다.
 - 사용 예시: 단위 시간당 교통사고 발생 건수, 콜센터에 걸려오는 전화 횟수, 웹사이트의 시간당 방문자 수 등 희귀한 사건의 발생 횟수를 모델링하는 데 주로 사용된다.

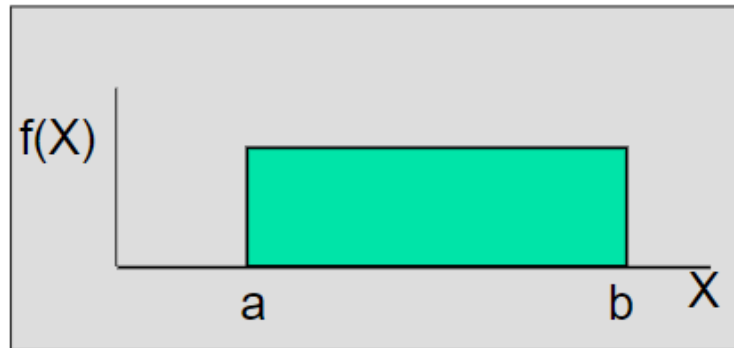
**N 이 없음 (즉, n 이 얼마든 상관없고
심한 경우 몰라도 됨)**



Some common distributions

- (Continuous) Uniform Distribution:

The Continuous Uniform Distribution:



where

$f(X)$ = value of the density function at any X value

a = minimum value of X

b = maximum value of X

$$\mu = \frac{a + b}{2}$$

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

For **[1, 6]**

$$\mu = \frac{1 + 6}{2} = 3.5$$

$$\sigma = \sqrt{\frac{(6 - 1)^2}{12}} = 1.44$$

Some common distributions

- Exponential Distributions:

- Used to model the **length of time between two occurrences** of an event (the time between arrivals)

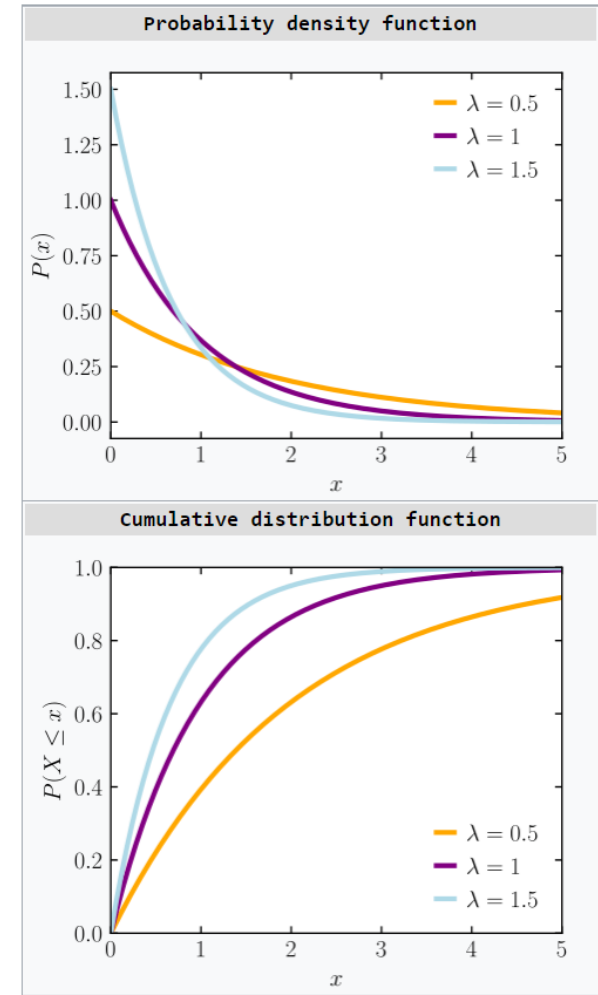
- PDF:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

- CDF:

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

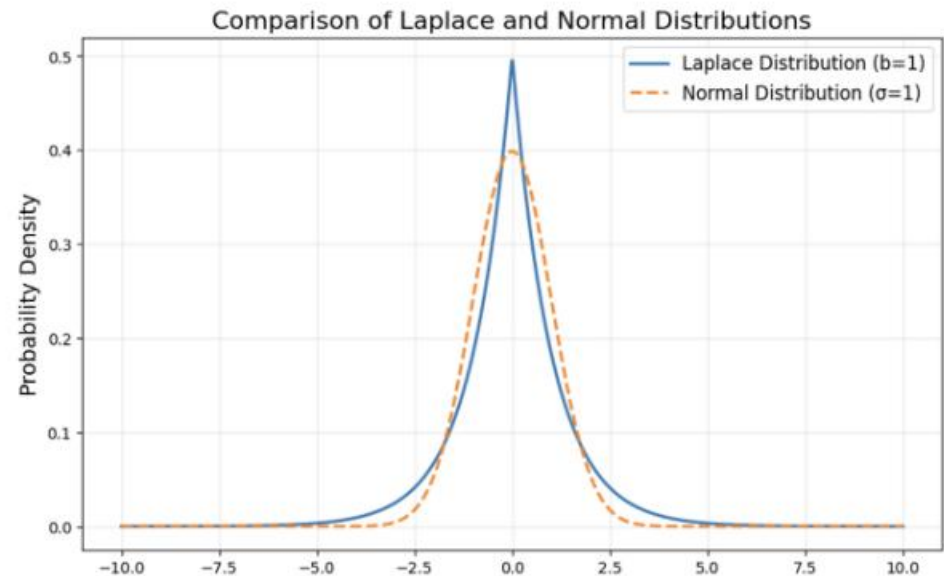
- 연속 확률 분포의 한 종류로, 포아송 분포를 따르는 사건들이 발생할 때, 한 사건이 발생하고 다음 사건이 발생하기까지 걸리는 시간 간격을 모델링하는 데 사용된다.
- 평균 발생 횟수(λ)에 반비례하는 평균($1/\lambda$)을 가지며, 시간이 지날수록 사건이 발생할 확률이 낮아지는 특징을 보인다.
- 사용 예시: 기계의 고장 발생 간격 시간, 고객이 매장에 도착하는 시간 간격, 전구의 수명 등 두 사건 사이의 '시간'을 분석하는 데 주로 사용된다.



Some common distributions

- Laplace Distribution
 - 이중지수분포라고 불림.
 - PDF:

$$f(x | \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$



- 정규분포와 비슷하게 종 모양을 띠지만, 꼬리(tails)가 더 두꺼워 평균으로부터 멀리 떨어진 이상치(outliers)에 더 큰 확률을 부여한다. 분산을 구할 때 제곱 차이(squared difference) 대신 절댓값 차이(absolute difference)를 사용한다는 특징이 있다.
- 사용 예시: 음성 인식에서 DFT 계수의 사전 확률(priors)을 모델링하거나, JPEG 이미지 압축에서 AC 계수를 모델링하는 데 사용된다.

Some common distributions

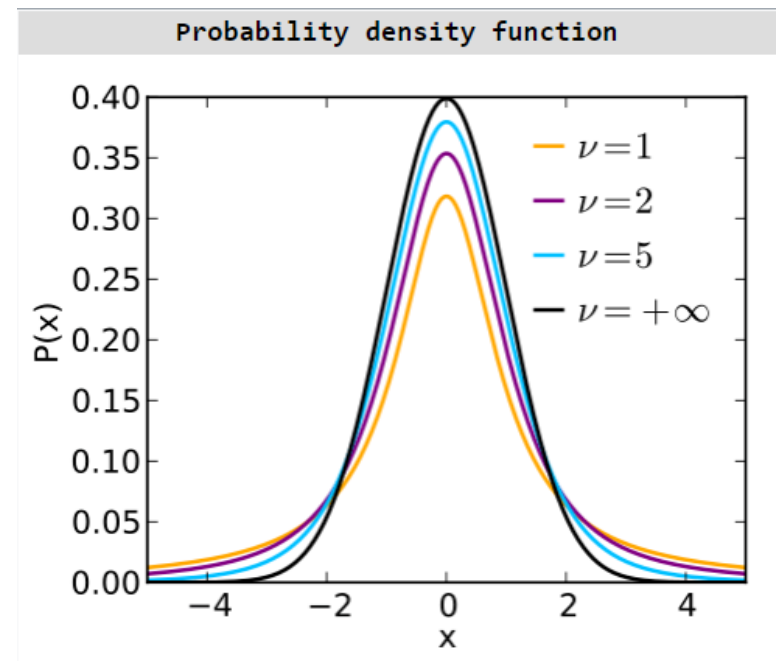
- Student's t-distribution (or t-distribution)

- PDF:
$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

ν : degrees of freedom, Γ : gamma function

- Looks like normal, but Heavier tails.

- t-분포는 정규분포와 유사하지만, 표본 크기(n)가 작을 때 사용하는 연속 확률 분포.
- 모집단의 표준편차(σ)를 모르는 상태에서 표본의 평균이 모집단 평균과 얼마나 떨어져 있는지를 설명하는 데 사용된다. 자유도(ν)가 커질수록 정규분포에 가까워진다.
- 사용 예시: 두 표본 평균의 차이를 검정하는 t-test에 사용되며, 모집단 표준편차를 모를 때 신뢰 구간 (confidence interval)을 구성하는 데도 활용된다

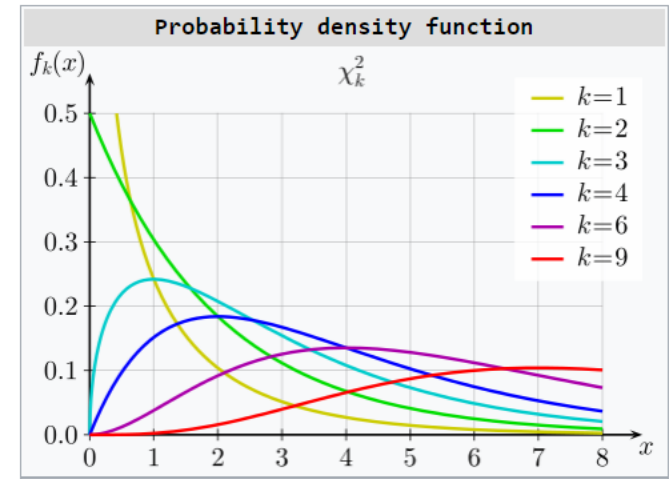


Some common distribution

- Chi-squared Distribution

- PDF:

$$f(x; k) = \begin{cases} \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



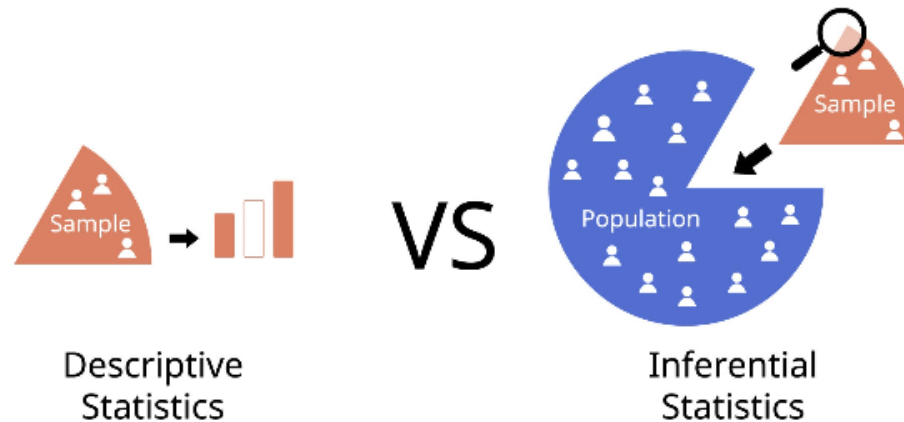
- 카이제곱 분포는 k 개의 독립적인 표준 정규 확률 변수들의 제곱 합이 따르는 분포다.
- 자유도(k)가 커질수록 분포의 모양이 정규분포에 가까워지는 특징이 있다.
- 사용 예시: 카이제곱 검정(Chi-squared Test)의 기반이 되는 분포로, 범주형 변수들 간에 연관성(relationship)이 있는지 검정하는 데 주로 사용된다. 또한, 표본 분산의 분포($(n-1)S^2/\sigma^2$)도 카이제곱 분포를 따르므로, 분산에 대한 추론 통계(inferential statistics)에도 활용된다.

Inference Statistics and hypothesis Testing

(추론통계 및 가설검정)

Inferential Statistics

- **Inferential Statistics (추론통계) : From Sample to Population**
 - 추론 통계는 표본(sample) 데이터를 사용하여 더 큰 모집단(population)에 대한 결론을 내리는 통계적 방법이다.
 - 추론 통계의 주요 목표는 표본의 특성을 바탕으로 모집단의 특성을 일반화(generalize)하는 것이다.
 - 가설 검정(Hypothesis Test)과 추정(Estimation)이 추론 통계의 가장 흔한 형태다.



<전체를 알 수 없기 때문에 표본을 사용한다.>

Hypothesis Testing

- **The Core of Hypothesis Testing:**

- 가설검정(Hypothesis Testing)은 통계적 의사 결정을 내리는 데 사용되는 방법이다.
- 귀무가설(H_0 , Null Hypothesis): "아무런 차이가 없다" 또는 "효과가 없다"는 기본 가정.
- 대립가설(H_a , Alternative Hypothesis, or H_1): 귀무가설과 반대되는 주장으로, 우리가 증명하고 싶은 가설.
- P-값(P-value): 귀무가설이 참일 때, 현재 관찰된 결과 또는 그보다 더 극단적인 결과가 나타날 확률을 측정한다.

The P-value

- **The P-value: An Intuitive Dice Example**

- 문제: 공정한 주사위를 10번 던졌을 때 6이 8번 나왔다. 이 주사위는 공정한가?
- 귀무가설(H_0): 주사위는 공정하다 ($p = 1/6$)
- 대립가설(H_a): 주사위는 공정하지 않다 ($p \neq 1/6$)
- p-값의 의미: "주사위가 공정하다는 가정 하에, 10번 중 8번 이상 6이 나올 확률은 얼마인가?"
- 결론: 이 확률이 매우 낮으면(p-값이 작으면), 우리는 귀무가설을 기각한다.

The P-value

- (문제) 주사위를 던졌더니 10회 중 8번 1이 나왔다. 이 주사위는 정상일까 아니면 무슨 장치가 되어 있을까 (비정상, 사기)?
- (가설 설정 및 검정)
 - 주사위를 정상이라고 가정하자. (H_0)
 - 그렇다면, 10회 중 8회 이상 나올 확률은

$${}_{10}C_8 (1/6)^8 (5/6)^{10-8} + {}_{10}C_9 (1/6)^9 (5/6)^{10-9} + {}_{10}C_{10} (1/6)^{10} = 0.0000194$$

- $p\text{-value} < 0.05$, So Reject H_0 .
- 즉, 이것은 있을 수 없는 수준으로 판단
 - > 원래의 가정이 틀렸다고 통계학적 결론.

The P-value

```
4 # 가설 설정: 주사위는 공정하다. (6이 나올 확률  $p = 1/6$ )
5 # 귀무 가설 ( $H_0$ ):  $p = 1/6$ 
6 # 대립 가설 ( $H_a$ ):  $p \neq 1/6$  (6이 비정상적으로 나온다)
7
8 # 실험 설정
9 num_rolls = 10          # 총 주사위를 굴린 횟수
10 p_success_null = 1/6    # 귀무 가설 하에서 6이 나올 확률
11 observed_successes = 8  # 관찰된 6이 나온 횟수 (극단적인 결과)
12
13 # p-값 계산 ('10번 중 8번 이상 6이 나올 확률'을 계산)
14 # stats.binom.sf(k, n, p)는 k회 초과 성공 확률을 계산하므로,
15 # k-1을 인자로 넣어 k회 이상 성공 확률을 구한다.
16 p_value = stats.binom.sf(observed_successes - 1, n=num_rolls, p=p_success_null)
17
18 alpha = 0.05    # 유의 수준 설정 (일반적으로 0.05 사용)
19
20 print(f"P-값: {p_value:.6f}")
21 print(f"유의 수준 (alpha): {alpha:.2f}")
```

```
23 # 결과 해석
24 if p_value < alpha:
25     print(f"P-값({p_value:.6f})이 유의 수준({alpha})보다 작으므로 귀무 가설을 기각한다.")
26     print(f"결론: 주사위는 공정하지 않다. (6이 나올 확률이 1/6보다 크다고 볼 수 있다.)")
27 else:
28     print(f"P-값({p_value:.6f})이 유의 수준({alpha})보다 크므로 귀무 가설을 기각할 수 없다.")
29     print(f"결론: 주사위가 공정하다는 것을 반박할 충분한 증거가 없다.")
```

```
P-값: 0.000019
유의 수준 (alpha): 0.05
P-값(0.000019)이 유의 수준(0.05)보다 작으므로 귀무 가설을 기각한다.
결론: 주사위는 공정하지 않다. (6이 나올 확률이 1/6보다 크다고 볼 수 있다.)
```

T-test and Z-test

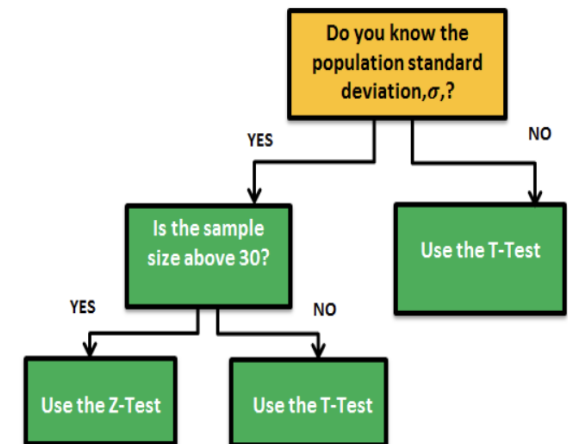
- T-test와 Z-test의 본질
 - 두 테스트 모두 가설 검정의 구체적인 방법으로 p-값을 계산하여 관찰된 결과가 우연에 의한 것인지 통계적으로 유의미한지 판단.
 - 이를 통해 “모집단 평균에 대한 가설이 옳은가?”라는 질문에 답을 제공.
- 왜 필요한가?
 - 모집단의 모든 데이터를 수집하는 것은 불가능하고 우리가 가진 것은 표본(sample) 데이터뿐임.
 - T-test와 Z-test는 이 작은 표본 데이터를 이용해 모집단 전체에 대한 결론을 내릴 수 있는 과학적인 방법을 제공.
 - 즉, 데이터에 숨겨진 의미 있는 차이를 발견하고, 이를 근거로 합리적인 의사 결정을 돕는다.
- 두 테스트의 주요 차이점
 - Z-test: 모집단 표준편차(σ)를 알고 있을 때 사용
 - T-test: 모집단 표준편차를 모를 때, 대신 표본 표준편차(s)를 사용. 현실의 대부분 상황에 더 유용.
 - T-분포는 이 불확실성을 반영해 정규분포보다 꼬리가 더 두껍다

t-test vs. Z-Test

- **Z-test and t-test**

- Z-test: to see whether the averages of the two datasets are different from each other when the standard deviation is available and the sample is large ($n \geq 30$).
- T-test: to see how averages of 2 data sets differ from each other when the standard deviation is not known.

	Z-test	T-test
Population std	known	unknown
Sample size	Large ($n \geq 30$)	Small ($n < 30$)
Key assumptions	<ul style="list-style-type: none">• All data points independent• Normal distribution for Z with mean 0, variance 1	<ul style="list-style-type: none">• All data points not dependent• Sample values are to be recorded and taken accurately
Based upon	Normal distribution	Student t-distribution



t-test

- **(Independent two-sample) t-test**

- when you want to compare two samples
- think of it as a ratio of signal (sample means) to noise (sample variability)
- **H0: $\bar{x}_1 = \bar{x}_2$** (H1: $\bar{x}_1 \neq \bar{x}_2$)
- t-statistic:

표본 평균 차이의 통계적 지표: 차이 / 불확실도

$$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference in means}}{\text{sample variability}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1, \bar{x}_2 = sample means
 s_1^2, s_2^2 = sample variances
 μ = population mean
 n_1, n_2 = sample sizes

- Degree of freedom = $(n_1-1)+(n_2-1) = n_1 + n_2 - 2$
- Look up the t-critical value (from t-table) based on degree of freedom (=n-1) and the alpha level (typically 0.05)

(*) 자유도(degree-of-freedom, DF): 어떤 값을 추정하기 위해 사용되는 독립적인 정보의 개수. 즉, 최종 결과를 결정하는 데 있어 자유롭게 변할 수 있는 값의 수를 의미

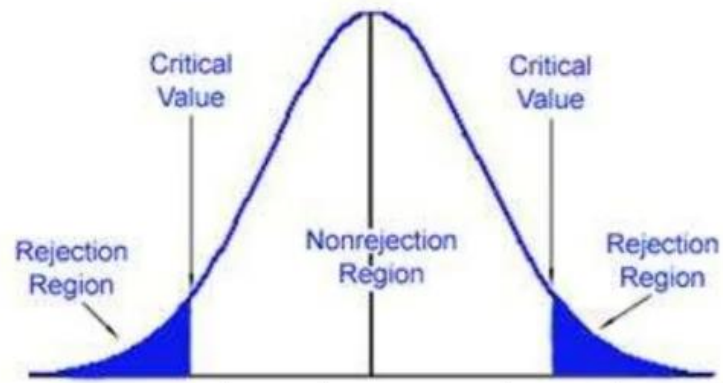
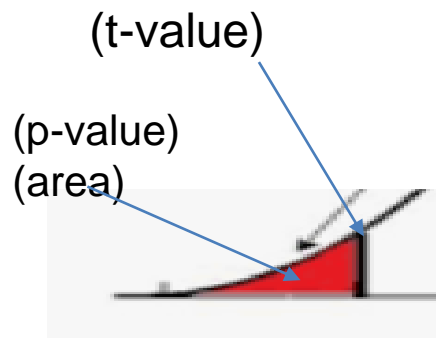
t-statistic

- **T-statistic: The Signal-to-Noise Ratio**
 - t-통계량(t-statistic)은 두 그룹 간 평균의 차이가 얼마나 큰지를, 표본 데이터의 변동성에 비해 측정하는 값이다.
 - 쉽게 말해, 우리가 발견한 '신호(signal)'가 데이터에 내재된 '잡음(noise)' 얼마나 큰지를 나타내는 비율.
- t-통계량의 구성 요소:
 - 분자 (Numerator): 두 표본 평균의 차이. 이는 우리가 확인하고자 하는 '신호'에 해당.
 - 분모 (Denominator): 표준 오차. 표본 데이터가 얼마나 퍼져 있는지를 나타내는 '잡음'에 해당.
- 결론
 - t-값이 클수록: 신호가 잡음보다 크다는 의미. 평균 차이가 우연일 가능성이 낮아 귀무가설(H_0)을 기각할 근거가 강해진다.
 - t-값이 작을수록: 신호가 잡음과 비슷하다는 의미. 평균 차이가 우연일 가능성이 높아 귀무가설을 기각하기 어렵다.

t-value, p-value

- **p-value in t-test**

- 귀무가설(H_0)이 참이라는 가정 하에, 현재 관찰된 결과(t-통계량) 또는 그보다 더 극단적인 결과가 나타날 확률
- t-분포 곡선에서 가로축(x-축)은 t-통계량(t-statistic)의 가능한 값들을 보여준다. 곡선의 가장 높은 지점인 중앙(0)은 귀무가설(H_0)이 참일 때 t-통계량이 가질 수 있는 가장 가능성이 높은 값이다.
- 입증하고자 하는 대립가설(H_1)이 "두 그룹의 평균이 같지 않다(\neq)"는 비방향성 주장을 할 때는 양측검정(two-sided-test), 한 그룹의 평균이 다른 그룹의 평균보다 크다($>$)" 또는 "작다($<$)"는 방향성 주장을 할 때는 단측검정 (One-tailed test)을 사용한다.



t-test Example

- Example
 - Compare the effectiveness of ammonium chloride and urea, on the grain yield of paddy, an experiment was conducted. The results are given below:

Ammonium chloride (X ₁)	13.4	10.9	11.2	11.8	14	15.3	14.2	12.6	17	16.2	16.5	15.7
Urea (X ₂)	12	11.7	10.7	11.2	14.8	14.4	13.9	13.7	16.9	16	15.6	16

- H₀: The effect of ammonium chloride and urea on grain yield of paddy are equal i.e., $\mu_1 = \mu_2$ (H₁: $\mu_1 \neq \mu_2$)

```
1 Ammonium_chloride=[13.4,10.9,11.2,11.8,14,15.3,14.2,12.6,17,16.2,16.5,15.7]
2 Urea=[12,11.7,10.7,11.2,14.8,14.4,13.9,13.7,16.9,16,15.6,16]
3 n1, n2 = len(Ammonium_chloride), len(Urea)
4 x1_m, x2_m = np.mean(Ammonium_chloride), np.mean(Urea)
5 x1_std, x2_std = np.std(Ammonium_chloride), np.std(Urea)
6 print("number of samples: ", n1, n2)
7 print("means of two samples: ", x1_m, x2_m)
8 print("std of two samples: ", x1_std, x2_std)
9 tval = (x1_m - x2_m) / np.sqrt(x1_std**2/n1 + x2_std**2/n2) # approx.
10 pval = scipy.stats.t.cdf(tval, n1+n2-2)
11 tval, (1-pval)*2
```

```
number of samples: 12 12
means of two samples: 14.066666666666665 13.908333333333333
std of two samples: 2.026217055390551 1.9956028050580497
(0.19286027404090295, 0.8488377215788003)
```

t-val, p-val computed by hand
(approximated)

```
1 t_value, p_value = ttest_ind(Ammonium_chloride, Urea)
2 t_value, p_value
```

```
(0.1846496543760765, 0.8551954147800473)
```

t-val, p-val by function

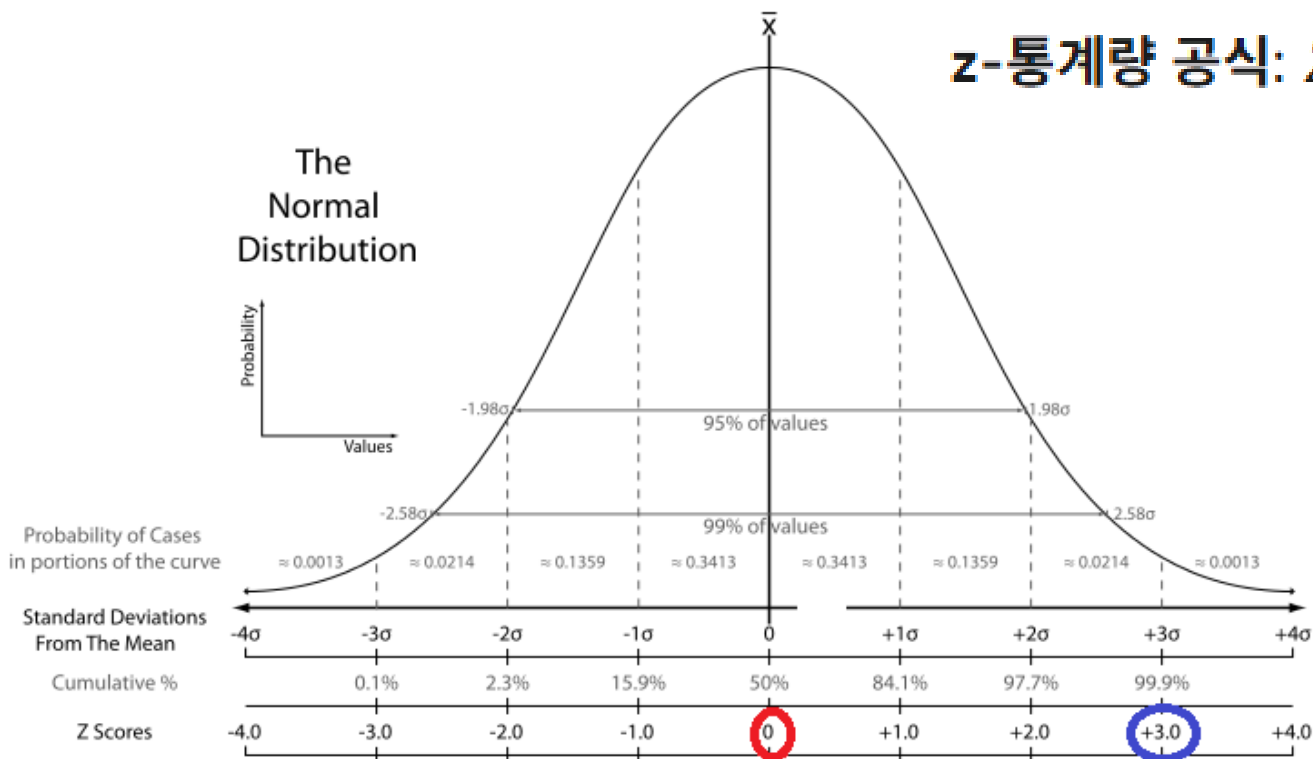
Z-test

- Z-test

- 모집단 표준편차(σ)를 알고 있을 때 사용한다. 하지만 현실적으로 모집단 표준편차를 알기 어려워 실제로는 거의 사용되지 않는다. (-test: 모집단 표준편차(σ)를 모르고, 대신 표본 표준편차(s)를 사용해 추정할 때 사용한다)
- z-test는 정규분포(Normal distribution)를 기반으로 한다. (t-test는 t-분포(Student's t-distributional distribution)를 기반으로 한다. 모집단 표준편차를 알기 때문에)를 기반으로 한다. 모집단 표준편차를 모르기 때문에 발생하는 추가적인 불확실성을 반영하기 위해 정규분포보다 꼬리가 더 두꺼운 t-분포를 사용한다.)
- 표본 크기(n)가 커질수록 t-분포의 모양은 정규분포에 매우 가까워진다. 따라서 표본 크기가 30 이상으로 충분히 크면 모집단 표준편차를 모르더라도 t-test 대신 z-test를 사용하기도 한다.

Z-values (Z-scores)

- z-통계량은 데이터 포인트가 평균으로부터 얼마나 떨어져 있는지를 표준편차(σ) 단위로 측정한다. t-통계량과 마찬가지로, z-통계량은 귀무가설이 참이라는 가정 하에 관찰된 결과가 얼마나 드문지 판단하는 데 사용된다.



Z-test Example

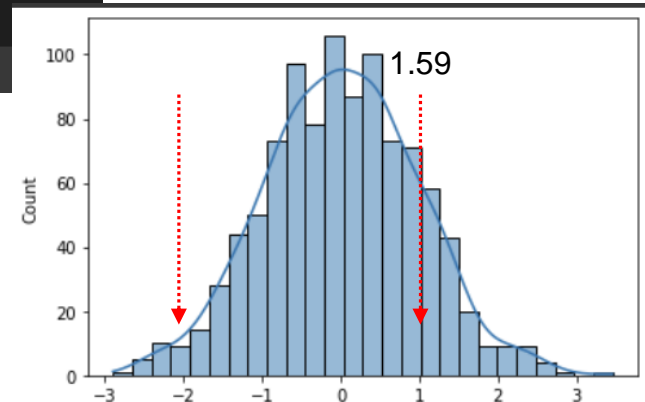
- Example:
 - Suppose the IQ in a certain population is normally distributed with a mean of $\mu = 100$ and standard deviation of $\sigma = 15$.
 - A researcher wants to know if a new drug affects IQ levels, so he recruits 20 patients to try it and records their IQ levels.

```
7 #enter IQ levels for 20 patients
8 data = [88, 92, 94, 94, 96, 97, 97, 97, 99, 99,
9         105, 109, 109, 109, 110, 112, 112, 113, 114, 115]
10
11 #perform one sample z-test
12 zvalue, pvalue = ztest(data, value=100) # value = known mean
13 zvalue, pvalue
```

```
(1.5976240527147705, 0.1101266701438426)
```

```
1 # calculation by hand
2 data = np.array(data)
3 n = len(data)
4 xbar = data.mean()
5 var = np.sum((data - xbar)**2).sum() / (n-1)
6 std = np.sqrt(var)
7 sigma_est = std / np.sqrt(n) # estimated population std
8 zval = (xbar - 100) / sigma_est
9 pval = (1 - stats.norm(0, 1).cdf(z)) * 2
10 zval, pval
```

```
1.5976240527147705, 0.11012667014384259)
```



```
1 ttest_1samp(data, 100)
```

```
Ttest_1sampResult(statistic=1.5976240527147705, pvalue=0.12662462536124378)
```

F-test (ANOVA)

- F-test

- 세 개 이상의 그룹 간 평균의 차이가 통계적으로 유의미한지 검정하는 분석 방법이다. 하지만 이 평균의 차이를 알아내기 위해 분산을 활용한다.

(분산분석: ANOVA, Analysis of Variance)

- 귀무가설(H_0): 모든 그룹 평균이 동일하다. ($H_0 : \bar{x}_1 = \bar{x}_2 = \dots = \bar{x}_k$)
- F-test는 '적어도 한 그룹의 평균이 다른 그룹과 다르다'는 결론을 내릴 수 있는지 알려준다.
- F-통계량:

$$F = \frac{\text{그룹 간 변동성}}{\text{그룹 내 변동성}} = \frac{\text{explained variance}}{\text{unexplained variance}} \text{ or } \frac{\text{between-group variability}}{\text{within-group variability}}$$

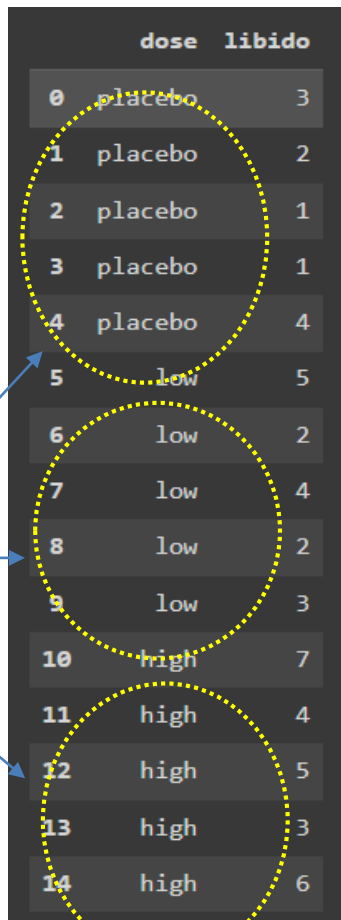
- F-값이 클수록: 그룹 간 변동성이 그룹 내 변동성보다 훨씬 크다는 의미다. 이는 그룹 간의 차이가 우연이라고 보기 어렵다는 강력한 증거가 된다. 따라서 H_0 를 기각하고 '그룹 평균에 유의미한 차이가 있다'고 결론 내린다.
- F-값이 작을수록: 그룹 간의 차이가 그룹 내 데이터의 무작위 변동성으로 충분히 설명될 수 있다는 의미다. 따라서 귀무 가설을 기각하기 어렵다.

ANOVA

- ANOVA (**A**nalysis of **V**ariance)
 - It tests for a difference overall between all groups.
 - Uses **f-test**.
- One-way ANOVA (or one-factor ANOVA)
 - a parametric test used to test for a **statistically significant difference** in an outcome between 3 or more groups. (it tests for a difference overall, i.e. at least one of the groups is statistically significantly different than the others.)
 - Why **one-way**? -> because those groups are under **one categorical variable**
 - If there are two variables being compared, it is called **two-way ANOVA** (if both variables are categorical), or it could be called an **ANCOVA** (if the 2nd variable is continuous). The "C" doesn't stand for continuous, it stands for the covariate.
- Parametric assumption assumptions:
 - Population distributions are normal
 - **Samples are equal variances**
 - **independence**

ANOVA Example

- Example: 리비도(성적욕구)를 증진하기 위해 개발된 새로운 약물의 효과를 검증하는 연구
(<https://www.pythonfordatascience.org/anova-python>)



	dose	libido
0	placebo	3
1	placebo	2
2	placebo	1
3	placebo	1
4	placebo	4
5	low	5
6	low	2
7	low	4
8	low	2
9	low	3
10	high	7
11	high	4
12	high	5
13	high	3
14	high	6

- 신약이 리비도(libido)에 미치는 영향을 분석하기 위한 실험
 - Dose: 복용량 (범주형: placebo, low, high)
 - Libido: 리비도 점수(정수형, 1~7 범위의 개인별 점수)
- 총 15명의 참가자가 있으며, 각 그룹(복용량)에 5명씩 배정되고 각 참가자는 약물 복용 후 리비도 점수를 평가받음

```
1 group1 = df['libido'][df['dose'] == 'high']  
2 group2 = df['libido'][df['dose'] == 'low']  
3 group3 = df['libido'][df['dose'] == 'placebo']
```

```
1 import scipy.stats as stats  
2 stats.f_oneway(group1, group2, group3)
```

```
F_onewayResult(statistic=5.11864406779661, pvalue=0.024694289538222603)
```

- 실험 분석 결과: $F = 5.12$, $p = 0.0247 \rightarrow$ 유의미한 차이 존재

Reject H_0 -> There is a statistically significant difference between the groups and their effects on the libido.

Chi-squared Test

- Chi-squared Test

- 카이제곱 검정은 두 범주형 변수(categorical variables) 간에 관계가 있는지 없는지 판단하는 데 사용된다. -> frequency test
- H0: “두 범주형 변수 사이에 관계가 없다(독립이다)”
- H1: “두 변수 사이에 관계가 있다”
- 카이제곱 통계량: 관찰된 빈도(observed frequency)와 기대 빈도(expected frequency)의 차이를 기반으로 계산한다.
 - 관찰된 빈도(O): 실제 데이터에서 각 범주 조합에 대해 관찰된 값.
 - 기대 빈도(E): 두 변수 사이에 아무 관계가 없다는 귀무가설(H0)이 참일 경우, 우리가 기대하는 값.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- 이 통계량이 크다는 것은 관찰된 빈도가 기대 빈도와 크게 다르다는 의미이므로, 귀무가설을 기각할 근거가 강해진다.

Chi2 Test in *scipy.stats*

- `scipy.stats.chi2_contingency()`
 - “Chi-square test of independence of variables in a contingency table”
 - Used when you don't know the underlying distribution but you want to test whether two (or more) groups have the same distribution
 - H0: "*the two groups have no significant difference*".

Chi2 test

- Example (chi2_contingency())

```
1 df.sample(frac=.1) # 데이터
```

	Age Group	Political Affiliation
735	45-65	Conservative
50	18-29	Conservative
473	30-44	Socialist
965	45-65	Socialist
848	45-65	Socialist
...
207	18-29	Socialist

original dataframe



```
2 data_crosstab = pd.crosstab(df['Age Group'],  
3                             df['Political Affiliation'])  
4 data_crosstab
```

Political Affiliation	Conservative	Other	Socialist
Age Group			
18-29	141	4	68
30-44	179	7	159
45-65	220	4	216
65 & older	86	4	101

cross table



```
1 # let's use the function chi2_contingency() - the same result  
2 tab = data_crosstab.values  
3 stats.chi2_contingency(tab)  
  
(24.367421717305202,  
 0.018121906378459784,  
 12,  
 array([[ 112.14297729,   3.40370059,   97.45332212,  213.        ],  
        [ 181.64003364,   5.51303616,  157.84693019,  345.        ],  
        [ 231.6568545 ,   7.03111859,  201.31202691,  440.        ],  
        [ 100.56013457,   3.05214466,   87.38772077,  191.        ],  
        [ 626.        ,   19.        ,   544.        , 1189.        ]]))
```

Chi2_contingency() returns:

- chi2-statistic
- p-value
- dof
- expected_frequency (E)

Estimation and Confidence Intervals

- 추정과 신뢰 구간

- 추정 (Estimation): 표본 데이터로 알 수 없는 모집단 모수(parameter)를 예측하는 과정.
- 점 추정 (Point Estimation): 표본 평균(\bar{x})처럼 하나의 값으로 모수를 예측. 간편하지만 불확실성 정보를 제공하지 못하는 한계가 있음
- 구간 추정 (Interval Estimation): 신뢰 구간(Confidence Interval)처럼 모수가 존재할 것으로 예상되는 일정 범위로 예측. 예측의 불확실성을 포함하여 보여준다.
- 신뢰 구간의 의미: 95% 신뢰 구간은 “우리가 이 과정을 반복했을 때, 생성된 구간의 95%가 실제 모집단 모수를 포함할 것”을 의미.

Applications in Data Science

Feature Selection and Feature Extraction

- **Feature Selection**

- Unsupervised: do not use the target variable (e.g. Correlations)
- Supervised: use the target variable
 - Wrapper: search for well-performing subsets of features (e.g. RFE (recursive feature elimination))
 - Filter: select subsets of features based on their relationship with the target
 - **Statistical Methods** (e.g. select features one by one and do ANOVA test with the target)
 - Feature importance Methods
 - Intrinsic: algorithms that perform automatic feature selection during training
 - Decision trees

- **Feature extraction**

- Dimension reduction: project input data into a lower-dimensional feature space
- T-SNE (statistical method)
- Autoencoders

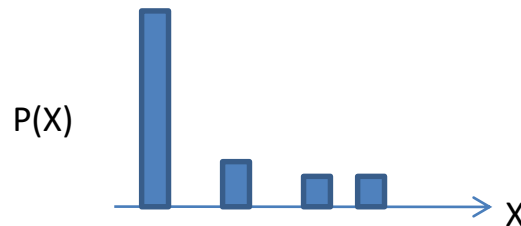
Statistics for Feature Selection

- Statistics for Feature Selection Methods

		output	
		Numerical (regression)	Categorical (classification)
Input	Numerical	Correlation coefficients <ul style="list-style-type: none">- Pearson's (linear)- Spearman Rank-based method (nonlinear)- more	Most common example of classification <ul style="list-style-type: none">- ANOVA correlation coefficient (linear)- Kendall's rank coefficient (nonlinear, ordinal categorical variable)
	Categorical	Not common case (nevertheless, you can still use the same "Numerical input, categorical output" methods, but in reverse.	Chi-squared test

Entropy

- $P(X)$ encodes our **uncertainty** about X
 - Some variables are more uncertain than others



- How can we quantify this intuition?
 - **Information** (or **Surprise**): $\log \frac{1}{p(x)}$
 - **Entropy**: average number of bits required to encode discrete R.V. X (or **expected value of the Surprise**)

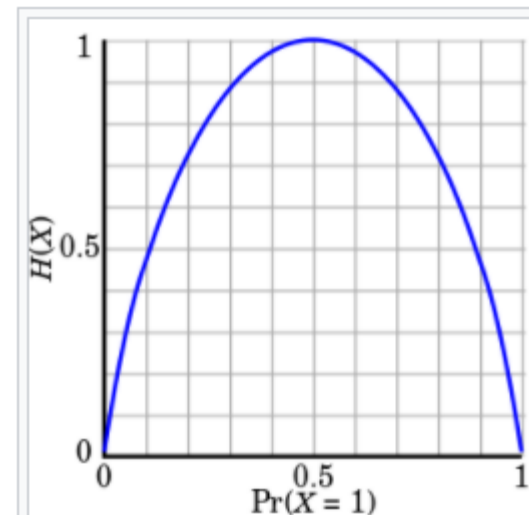
$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

Binary Entropy

$$H_p(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

If $\Pr(X = 1) = p$, then $\Pr(X = 0) = 1 - p$ and the entropy of X (in shannons) is given by

$$H(X) = H_b(p) = -p \log_2 p - (1 - p) \log_2 (1 - p),$$



Entropy of a Bernoulli trial as a function of binary outcome probability, called the **binary entropy function**.

Cross Entropy

- Cross Entropy (CE, 교차 엔트로피)
 - 교차 엔트로피는 참 분포(P)의 정보를 전달하기 위해 모델 분포(Q)의 인코딩을 사용했을 때 필요한 평균 비트 수(average number of bits)를 의미.
 - 손실 함수(loss function)로 사용된다. 머신러닝 모델의 목표는 모델이 예측한 확률 분포(Q)를 실제 정답의 확률 분포(P)와 최대한 가깝게 만드는 것인데, 이 둘의 차이를 교차 엔트로피로 측정하고 이 값을 최소화하는 방식으로 학습을 진행.
 - 두 확률 분포 간의 차이를 측정하는 지표로, 주로 머신러닝의 손실 함수(loss function)로 사용된다.
 - 즉, 교차 엔트로피가 낮을수록 모델의 예측(Q)이 정답(P)과 더 가깝다는 의미다.
 - For discrete distributions p and q with the same support (range of values) X ,

$$H(p, q) = - \sum_{x \in X} p(x) \log q(x)$$

KL Divergence

- KL 발산 (Kullback–Leibler Divergence)
 - KL 발산은 한 확률 분포 $P(x)$ 가 두 번째 분포인 $Q(x)$ 와 얼마나 다른지를 측정하는 지표.
 - $P(x)$ 가 '참' 분포이고, $Q(x)$ 가 이를 근사하는 '모델' 분포라고 가정할 때, KL 발산은 $P(x)$ 를 따르는 데이터를 설명하기 위해 $Q(x)$ 를 사용했을 때 발생하는 정보 손실(information loss) 또는 추가적인 '놀라움(surprise)'의 기댓값.
 - KL 발산 값은 항상 0 이상이다. 만약 두 분포 P 와 Q 가 완전히 일치하면 KL 발산 값은 0이 되며, 값이 낮을수록 두 분포가 더 비슷하다는 것을 의미.

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

$$H(p, q) = H(p) + D_{KL}(p \parallel q)$$

$$= - \sum_{i=0}^n p(x_i) \log(p(x_i)) + \sum_{i=0}^n p(x_i) \log(p(x_i)) - \sum_{i=0}^n p(x_i) \log(q(x_i))$$

$$= - \sum_{i=0}^n p(x_i) \log(q(x_i))$$

Naïve Bayes

- What is Naïve Bayes
 - Bayes' theorem with the “naive” assumption of conditional independence between every pair of features
 - given class variable y and dependent feature vector x_1 through x_n ,

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

Using the naive conditional independence assumption that

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),$$

for all i , this relationship is simplified to

$$P(y | x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x_1, \dots, x_n)}$$

Since $P(x_1, \dots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y | x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y)$$

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y),$$

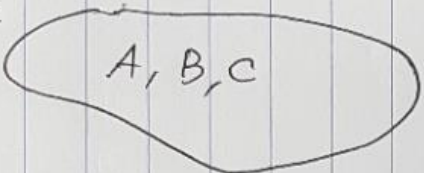
we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i | y)$.

Naïve Bayes

- Example:

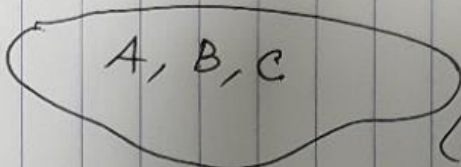
Naive Bayes

ham



A, B, C

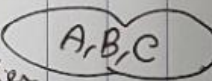
spam



A, B, C

(K = A ∩ B ∩ C)

Now, we have. $P[\text{ham}], P[\text{spam}]$
 $P(A|\text{ham}), P(B|\text{ham}), P(C|\text{ham})$
 $P(A|\text{spam}), P(B|\text{spam}), P(C|\text{spam})$.

We want to know ~~if~~ whether  is ham or spam?

① $P(\text{spam}|K) = \frac{P(K|\text{spam}) \cdot P(\text{spam})}{P(K)} = \frac{P(A, B, C|\text{spam}) \cdot P(\text{spam})}{P(A, B, C)}$

= $\frac{P(A|\text{spam}) P(B|\text{spam}) P(C|\text{spam}) \cdot P(\text{spam})}{P(A) P(B) P(C)}$

② $P(\text{ham}|K) = \frac{P(A|\text{ham}) P(B|\text{ham}) P(C|\text{ham}) \cdot P(\text{ham})}{P(A) P(B) P(C)}$

} choose bigger one!

통계의 머신러닝 적용

- 탐색적 데이터 분석 (EDA): 기술 통계(Descriptive Statistics)를 사용하여 데이터를 요약하고 시각화하여 데이터의 패턴과 특성을 이해한다.
- 모델 파라미터 추정: 통계에서 모집단 모수를 추정하듯이, 머신러닝 모델은 데이터를 통해 최적의 모수(parameters)를 학습(fitting)한다..
- 분류 모델: 나이브 베이즈(Naïve Bayes)는 베이즈 정리(Bayes' theorem)를 기반으로 한 분류(classification) 알고리즘이다.
- 피처 선택(Feature Selection): 카이제곱 검정(Chi-squared Test)이나 ANOVA를 사용하여 모델 예측에 가장 중요한 특성(feature)을 선별한다.
- 모델 성능 평가: P-값(P-value)을 통해 회귀 모델의 변수가 통계적으로 유의미한지 판단한다.
- 도메인 엔지니어와의 원활한 대화