



Parrot 정규 세션 : 통계

Estimate of Location 위치 추정

- Mean 평균(average)

자료를 모두 더한 후 자료의 개수로 나눈 값

- Median 중앙값

크기의 순서대로 정렬된 데이터 중 가운데 위치한 값

- Trimmed Mean 절사평균

자료를 크기 순으로 나열하여 이상치를 제외한 후 남은 자료들의 평균

* Outlier 이상치(extreme value) : 관측한 데이터 범위에서 크게 벗어난 통계적 관측치

- Weighted Mean 가중평균

중요도나 영향도에 해당하는 가중치를 곱하여 구한 평균

Estimate of Variability

변이 추정

- Standard Deviation 표준편차

관측치들의 산포도를 나타내는 수치

* 이상치 : 평균 $\pm 3\text{std}$

- IQR Inter-Quartile Range 사분범위(사분위수 범위)

분포의 양끝 ¼을 제외한 범위 = $Q3 - Q1$

* 이상치 : $Q1 - 1.5 \text{ IQR}$, $Q3 + 1.5 \text{ IQR}$

- $Q1$ (1사분위수) : 데이터의 25% 지점
- $Q2$ (2사분위수) : 데이터의 50% 지점 = 중앙값(median)
- $Q3$ (3사분위수) : 데이터의 75% 지점

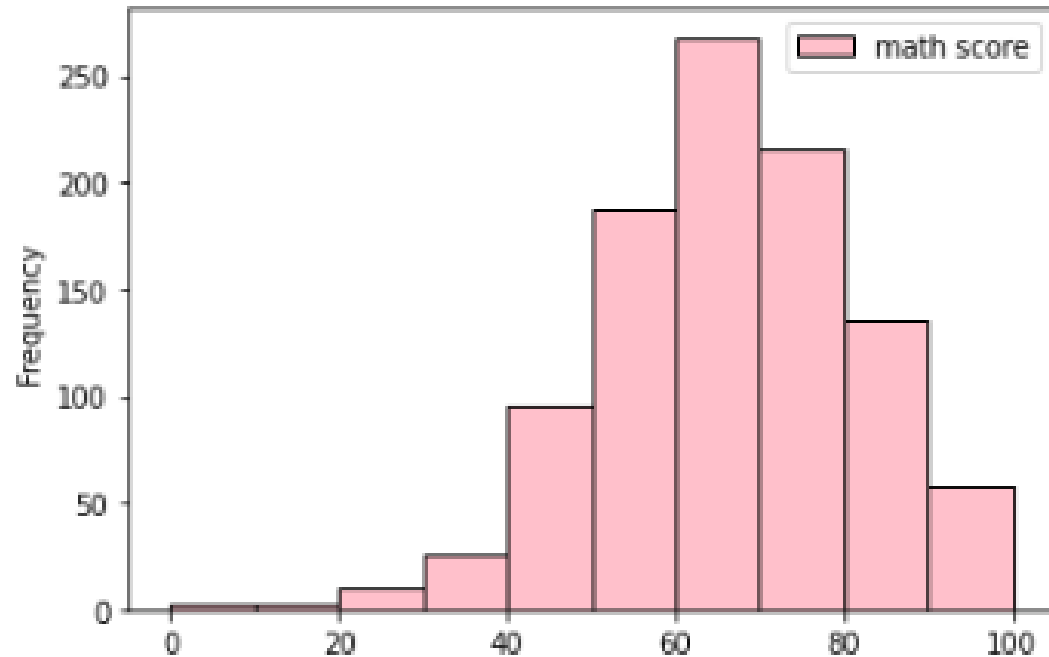
- Mad median absolute deviation from the median 중위절대편차(중앙값 절대편차)

|중앙값으로부터의 편차들|의 중앙값

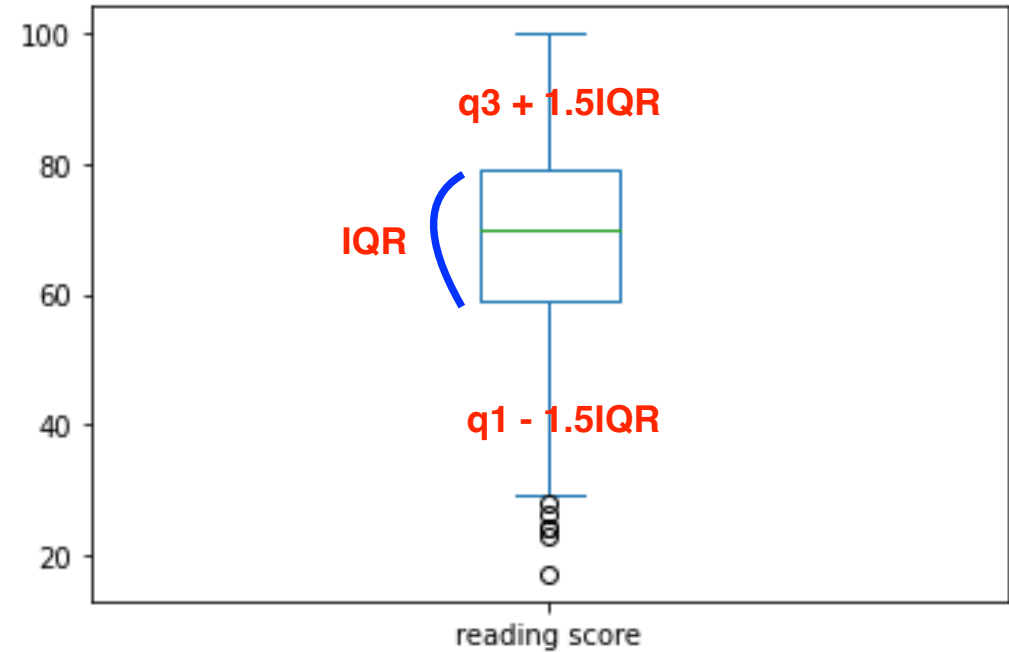
* 이상치 : 중앙값 $\pm 3\text{mad}$

Histogram & Boxplot 히스토그램 & 박스플롯

[히스토그램]



[박스플롯]



Correlation 상관계

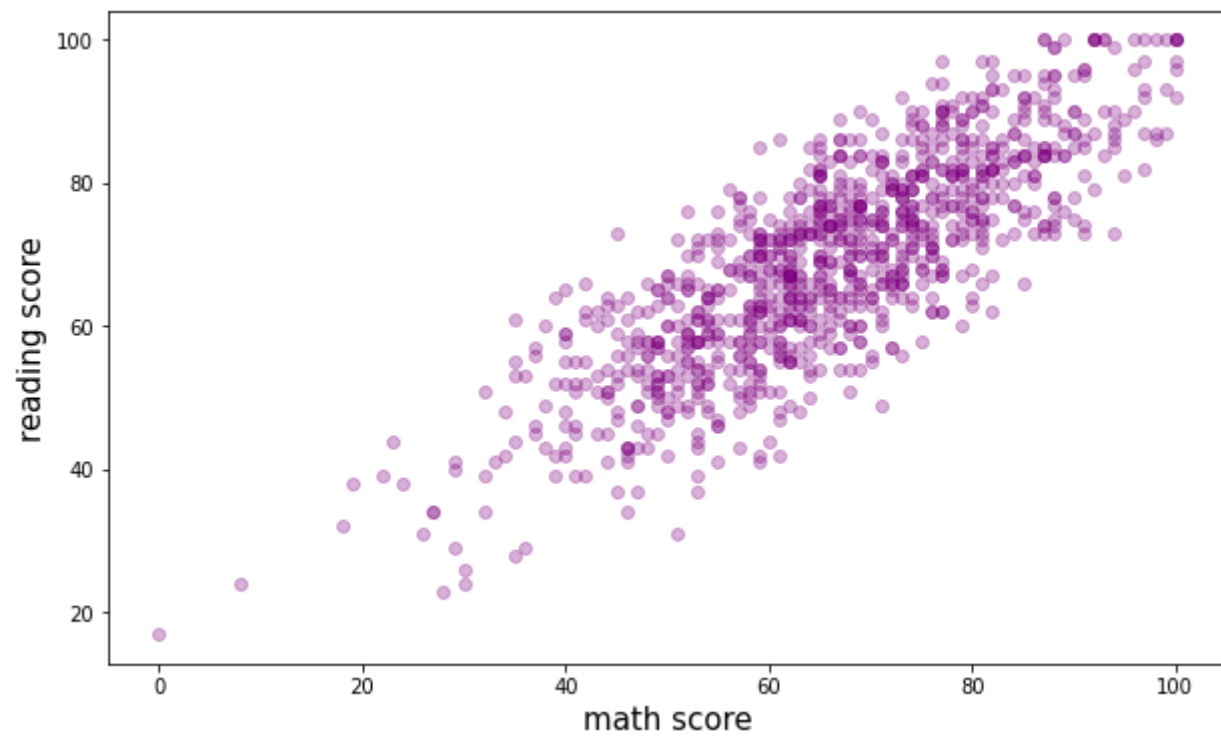
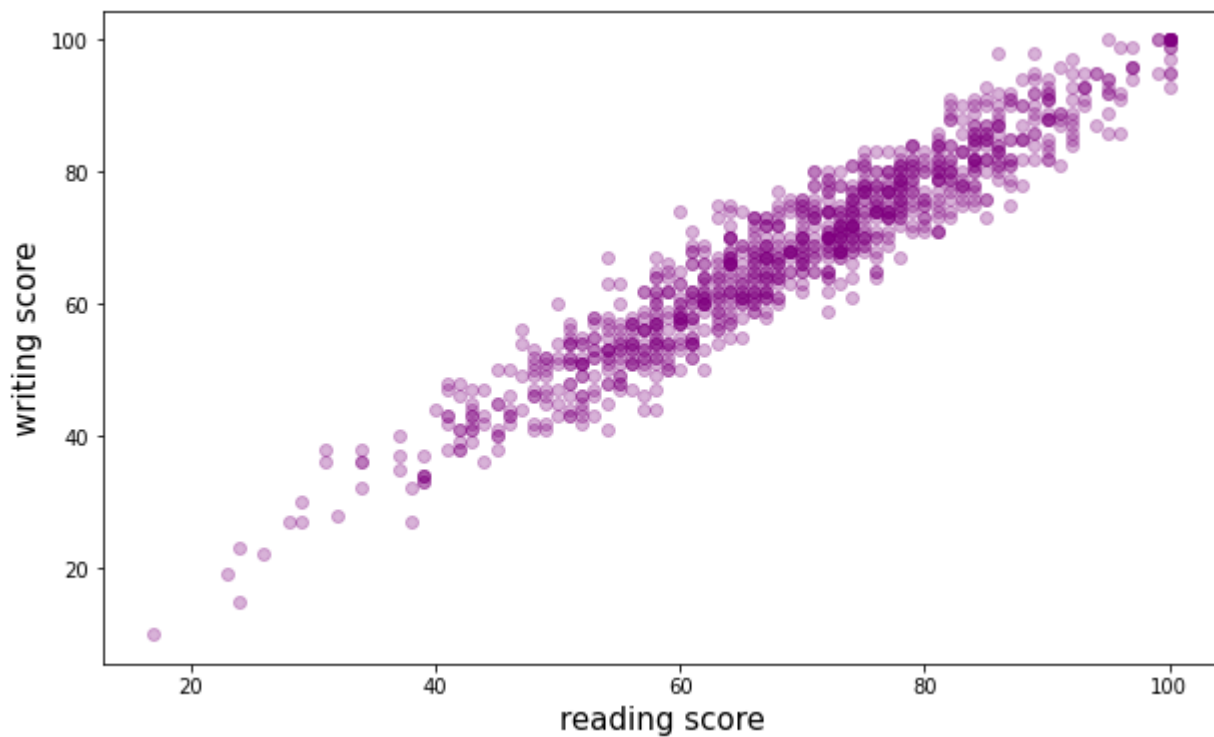
- Correlation coefficient 상관계수

두 변수 간 선형관계의 방향과 강도를 숫자로 나타낸 것

- 상관계수가 0인 경우 : 두 변수는 연관이 없음
- 상관계수가 1에 가까운 경우 : 두 변수는 정적 상관을 나타냄
- 상관계수가 -1에 가까운 경우 : 두 변수는 부적 상관을 나타냄

★ 두 변수의 관계를 나타내는 것이지 인과관계를 나타내는 것이 아님 ★

Scatter Plot 산점도(산포도)



Distributions 분포

- Normal Distribution 정규분포

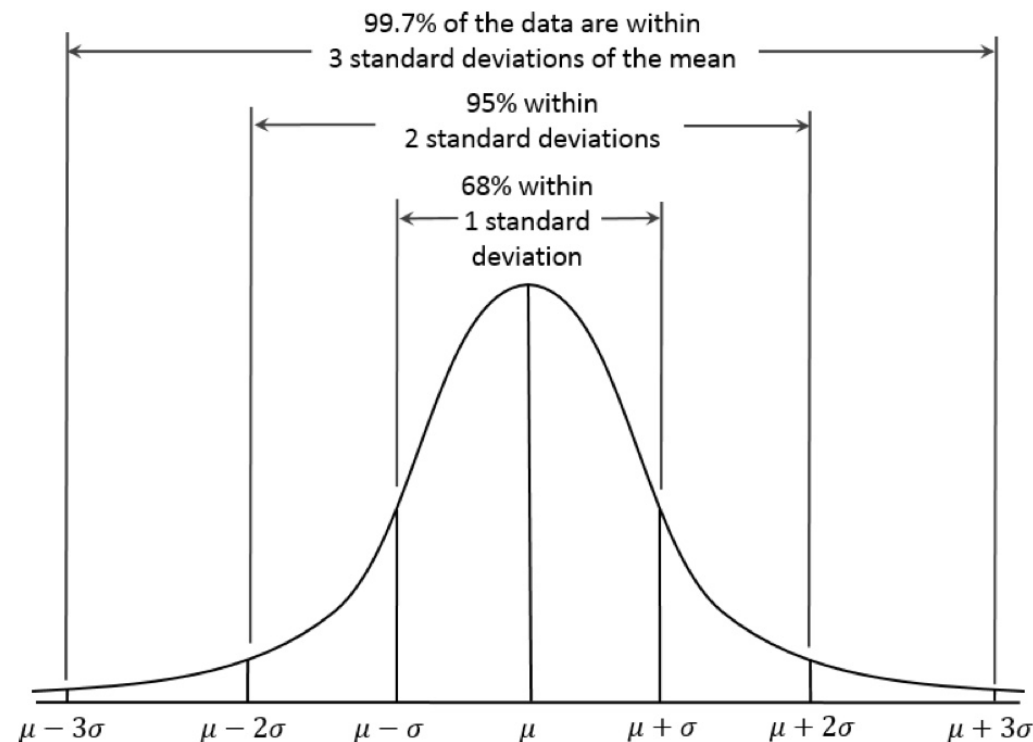
정규분포 곡선은 히스토그램의 일종
따라서 정규분포곡선을 이용하여 데이터를 표준화하면 해당 데이터의 히스토그램을 매끄럽게 만들 수 있음

특징

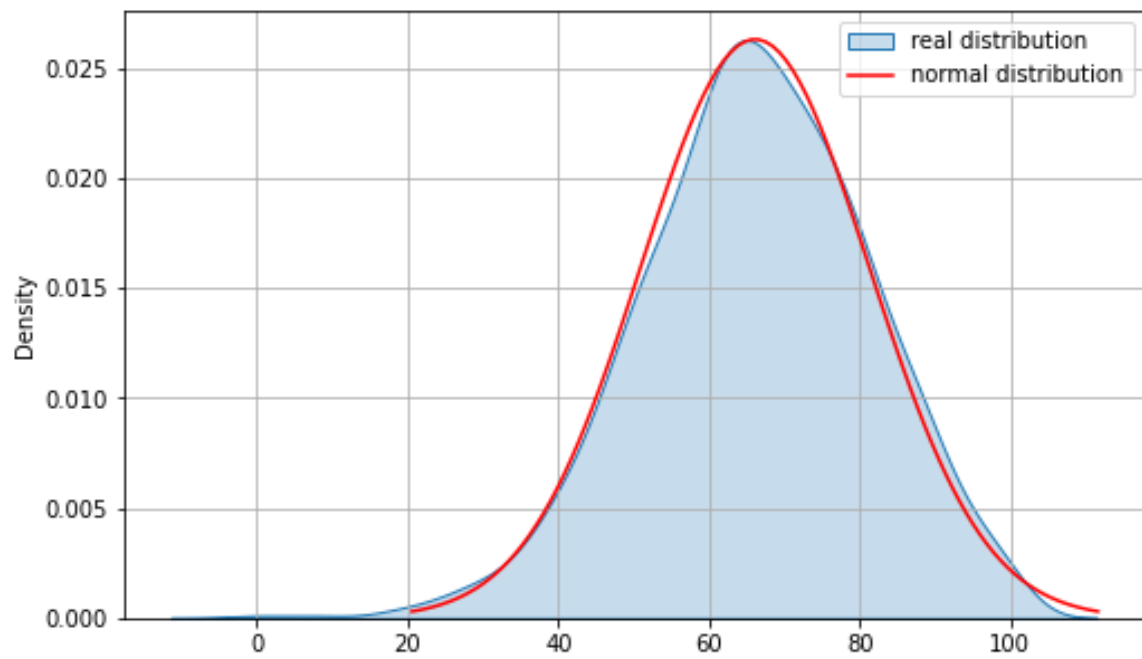
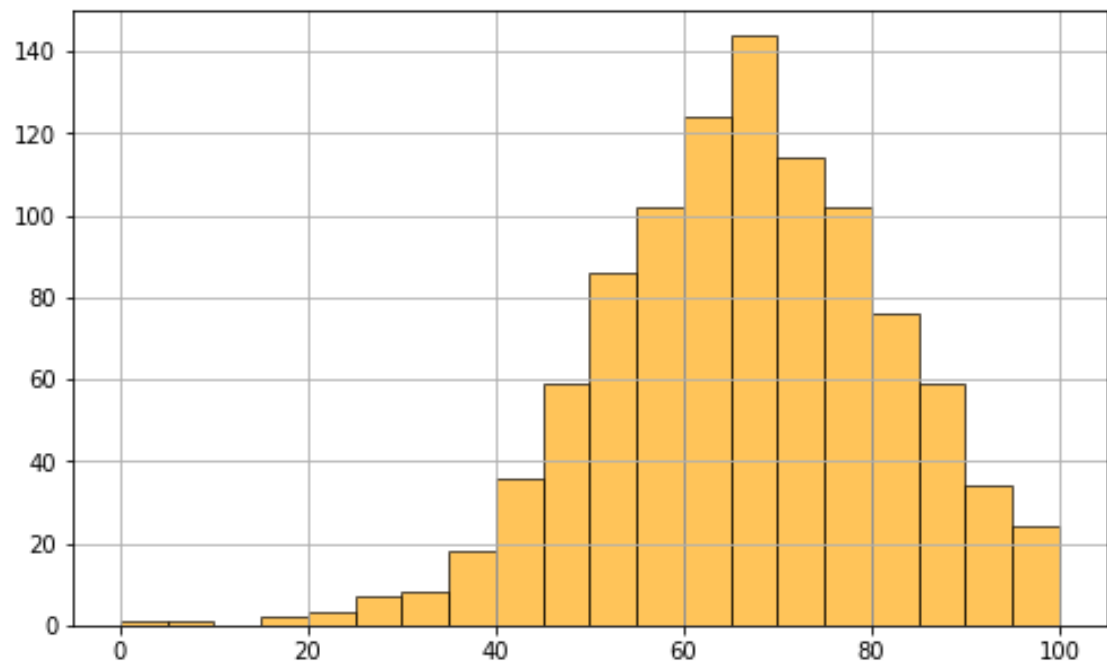
- 1) 평균을 중심으로 좌우대칭
- 2) 평균과 표준편차를 통해 정규분포곡선 아래 면적의 크기를 구할 수 있음

* 중심극한정리
: 확률 히스토그램(가능한 값을 확률로 계산하여 나타낸 히스토그램)은 시행 횟수가 증가할수록 정규분포곡선에 수렴

분포가 고르지 않은 데이터도 중심극한정리가 성립



정규분포 가정



Distributions 분포

- Standard Normal Distribution

표준정규분포

평균이 0이고 표준편차가 1인 정규분포

표준화를 해주면 평균과 표준편차가 다른 집단간의 비교가 용이해짐

- Long-Tailed Distributions

긴꼬리분포

일반적으로 데이터는 극단값이 존재 → 극단값이 분포의 꼬리 형성

(대부분의 데이터는 정규분포를 따르지 않음 따라서 정규분포를 가정하면 극단값을 과소평가할 가능성 있음)

- Student's t-Distribution

t-분포

표본 통계량의 분포를 묘사하는 데 사용 **n이 30이상이면 정규분포로 근사 씹가능**
정규분포와 비슷하지만 더 두꺼운 꼬리를 갖고 있음

**t분포는 평균검정할때 사용을 함
한 집단은 일반적으로 하고
두 집단일 경우는 둘다 $n < 30$ 에
동분산일 경우만 가능**

- F-Distribution

F-분포

두 집단의 분산에 대한 분포

선형회귀분석 혹은 분산 분석(ANOVA)할 때 사용

Distributions 분포

- Binomial Distribution 이항분포

베르누이 시행을 독립적으로 n 번 반복한 분포

* 베르누이 시행 : 결과가 두 가지로 나타나는 시행 ex) 동전의 앞뒤, 양품/불량품 등

** n 이 충분히 크고, p 가 0이나 1에 가깝지 않을 때 이항분포는 정규분포로 근사

- Chi-Square Distribution 카이스퀘어분포

표준정규확률변수를 제공해서 더한 분포

데이터가 평균으로부터 얼마나 떨어져 있는지를 확률적으로 나타내는 분포

* 범주형 데이터 분석, 적합도 검정, 교차 분석 등에 사용

- Poisson Distributions 포아송분포

시간이나 공간(영역)에서 특정한 사건의 분포

* λ : 시간 혹은 공간 단위당 사건 발생횟수의 평균

Hypothesis Tests 가설 검정(유의성 검정)

통계에서 나타나는 현상이 우연인지 실질적인지 밝혀내는 통계법

- Null hypothesis 귀무가설

기존의 사실(모집단으로부터 알려진 정보)로 검정하려는 가설

- Alternative hypothesis 대립가설

귀무가설에 대립되는 가설

자료로부터 정확한 증거에 의하여 입증하고자 하는 가설

- One-way test 단측검정

확률 결과를 한 방향으로만 계산하는 가설 검정

- Two-way test 양측검정

확률 결과를 두 방향으로 계산하는 가설 검정

Hypothesis Tests 가설 검정(유의성 검정)

통계에서 나타나는 현상이 우연인지 실질적인지 밝혀내는 통계법

- Type 1 Error 제1종오류

귀무가설이 옳음에도 불구하고 이를 기각하는 오류

- Type 2 Error 제2종오류

귀무가설이 거짓임에도 불구하고 이를 기각하지 않는 오류

	귀무가설 참	귀무가설 거짓
귀무가설 기각	제1종오류	옳은 결정
귀무가설 기각 X	옳은 결정	제2종오류

Hypothesis Tests 가설 검정(유의성 검정)

통계에서 나타나는 현상이 우연인지 실질적인지 밝혀내는 통계법

- Alpha(Significance Level) 유의수준

제1종오류를 범할 확률

α

- Power 검정력(검출력)

귀무가설이 거짓일 때 이를 기각할 확률

$1 - \beta$

- P-value 유의확률

귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률

* 가설 검정은 귀무 가설이 참이라고 가정하고 결과의 합리성 판단

표본 크기가 크면 정규분포곡선을 이용하여 p값을 구하고 Z-검정

표본 크기가 작으면 t-분포를 이용하여 p값을 구하고 t-검정

Anova 분산분석(analysis of variance)

여러 집단을 대상으로 한 실험의 결과를 분석하기 위한 통계법

- 정규성 검정

가정 1 : 각 집단은 정규분포를 따름

Shapiro-wilks test, Kolmogorove-Smirnov test, Quantile-Quantile plot

- 등분산 검정

가정 2 : 각 그룹의 분산은 동일함

Levene's test, Bartlett's test

- 독립성 검정

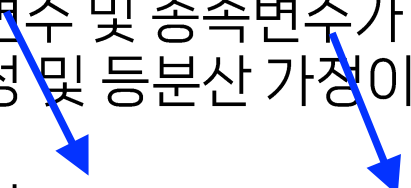
가정 3 : 각 그룹은 독립적으로 추출됨

Anova 분산분석(analysis of variance)

여러 집단을 대상으로 한 실험의 결과를 분석하기 위한 통계법

- One-Way Anova 일원분산분석

독립변수 및 종속변수가 모두 1개인 경우
정규성 및 등분산 가정이 만족할 때 사용



회귀식 : $Y = B_0 + B_1X_1$

- Kruskal-Wallis test 크러스칼-왈리스 검정

정규성 가정이 만족하지 않을 때 사용

<https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/>

- Welch's Anova <https://www.statisticshowto.com/welchs-anova/>

정규성은 만족하나 등분산성을 만족하지 않을 때 사용

Anova 분산분석(analysis of variance)

여러 집단을 대상으로 한 실험의 결과를 분석하기 위한 통계법

- Two-Way Anova 이원분산분석

독립변수는 2개 이상이고 종속변수는 1개인 경우
정규성 및 등분산 가정이 만족할 때 사용

$$\text{회귀식} : Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

[추가 자료] 데이터 분석 방법

		독립변수	
		범주형	연속형
종속변수	범주형	교차분석	로지스틱회귀분석 판별분석
	연속형	ANOVA t-test 분석	상관분석 회귀분석