

1 데이터분석 의사결정을 위한 수학/통계적 언어를 이해하기

[Open in Colab](#)

"데이터분석이 뭐길래?"

- 사람이 비이성적, 비효율적으로 처리하는 방식을 정량적, 체계적, 합리적 의사결정 하도록 다양한 도구들을 활용하는 것

비즈니스/사람 일상 반복	사람	분석	분석도구
문제/호기심	어떤 현상을 보고	어떤 현상을 보고 알고 싶은 것을 정의	
관찰	왜 그런일이 벌어지는지 경험/직감/인터넷으로 짧게 공들이고	왜 그런일이 벌어지는지 정량적 및 체계적으로 대량 수집	데이터 시각화, 기술적 분석, 데이터집계, 상관분석 등
주장/의사결정	그 이유를 대충 근거를 들어 주장한 후	통계적 검정을 통해 주장한 후	통계추론, 확률적 검정
검증	나와 의견이 같으면 좋은 사람 아니면 유탄	주장이 충분히 합리적이며, 알고리즘을 이용하여 그 현상을 재현해내며 주장을 설명	확률통계, 기계학습, 딥러닝, 지도학습, 비지도학습, 강화학습, 검증지표 등
해결	???	알고 싶은 것을 해결하고 미래에도 사용	

1.1 확률? 통계? 데이터사이언스?

1) 배경:

"우리는 뜬금없이 경우의 수? 순식간에 확률 공식? 곧바로 통계? 부터 첫만남을 해서 당황스럽고 충격을 받아 내려놓게 되었는데..."

"어느날 갑자기 데이터사이언스? 라는게 화두가 되면서 호기심을 가져볼까 했는데 또다시 경우의 수... 확률... 통계... 부터 나오니 이 글자를 어떻게 벗어날 수 있을지 걱정 될 수 있습니다."

"데이터 분석에 이 글자들이 어떻게 쓰이는지 정리 해 보면.."

데이터 분석	목적	대응
데이터 시각화	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 데이터를 한눈에 확인 - 점그림, 선그림, 영역그림 - 막대그림, 등고선그림, 분포그림(히스토그램) → 통계 사용 (2) 데이터를 뿐이고 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/컴퓨터 사용
기술적 분석	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 특성을 몇 개의 숫자들로 확인 → 통계 사용 (2) 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/컴퓨터 사용
상관관계/인과관계	여러종류 데이터끼리의 관계를 알고 싶다	(1) 각 데이터를 몇 개의 숫자들로 표현 → 통계 사용 (2) 표현된 숫자들을 비교 → 확률/통계 사용
통계추론	일부 데이터로 전체를 알고 싶다	(1) 일부 데이터의 특성을 확인 → 통계 사용 (2) 빈번적으로 실험 진행 및 통계치 재확인 → 컴퓨터 사용 (3) 전체 특성을 추론 → 통계 사용
알고리즘학습	전체 데이터로 미래를 알고 싶다	(1) 데이터의 관계를 수학적으로 표현 → 확률/통계/함수/컴퓨터 사용 (2) 미래를 예측한 후 정확성 확인 → 확률/통계/컴퓨터 사용
가설검정(A/B Test)	뭔가 진실과 가까운 의사결정을 하고 싶다	(1) 기존 데이터와 새로운 데이터 비교를 위해 숫자들로 표현 → 통계 사용 (2) 표현된 숫자들을 비교 → 확률/통계/컴퓨터 사용

2) 확률? 통계?:

- 따로 배워서 다른 것 같지만, 일반적으로 통계가 확률을 사용
- 데이터를 통계적으로 표현하는 이유는 이를 사용하여 무언가를 추론하거나 예측을 할 때 확률을 이용해서 설명하기 때문
- 확률은 알려진 모델을 통해 데이터를 예측하는 것이고, 통계는 주어진 데이터로부터 모델을 예측하는 것

확률 vs 통계





- 예시:

"휴대폰 매장에 스마트폰이 빨간색 7개와 검은색 3개가 있다"

(1) 통계: 랜덤하게 비복원(뽑은걸 다시 매장에 반납)으로 계속 선택했더니, 100번중 빨간색이 69번 검은색이 31번이 선택되었다면 매장이 보유한 빨간색은 몇개?

→ 스마트폰 색상 재고량 추정 알고리즘 생성

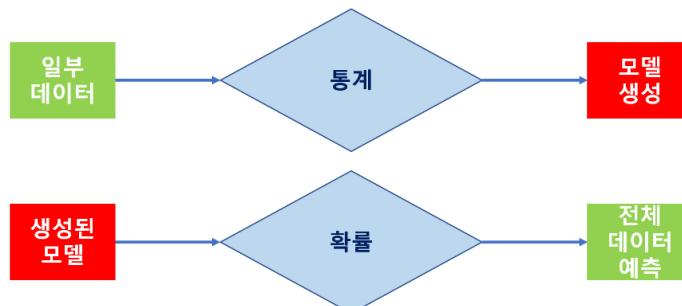
(2) 확률: 랜덤하게 1개를 선택해서 구매한다면 빨간색을 구매하게 될 확률은?

→ 스마트폰 색상 재고량 추정 기반 비즈니스 예측

3) 데이터분석/데이터사이언스: 확률과 통계를 사용하여 분석 목적에 따라 통계추론 또는 알고리즘학습 사용 의사결정 프로세스

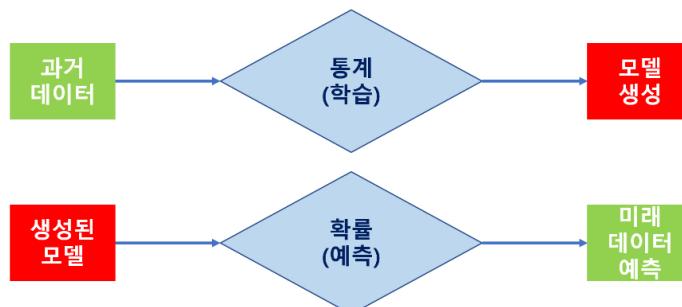
- 통계추론 기반 데이터분석 목적: 일부집단(스몰데이터) → 전체집단(빅데이터) 추론

"통계추론"은
 (1) 일부 데이터로 전체특성을 확인(모델로 반영)하고
 (2) 추론된 모델을 전체로 가정하여 확률적으로 전체 데이터를 예측



- 알고리즘학습 기반 데이터분석 목적: 과거의 특성으로 전체집단(빅데이터) 가정하고 일부 미래특성(스몰데이터) 추론

"알고리즘학습"은
 (1) 과거 데이터를 학습하여 과거패턴을 확인(모델로 반영)하고
 (2) 추론된 모델을 전체로 가정하여 확률적으로 일부 미래 데이터를 예측

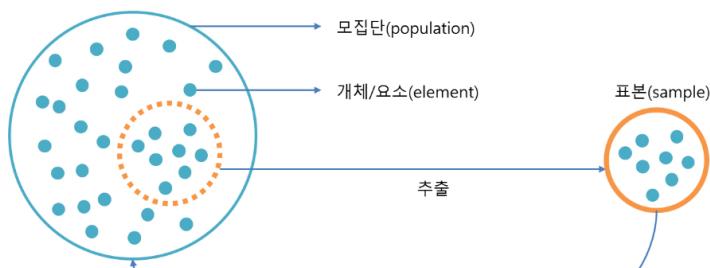


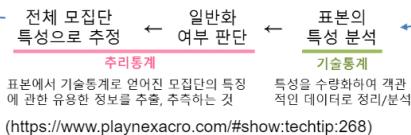
1.2 통계와 기계학습/인공지능의 차이?

1) 통계: 통계추론 방식을 사용하여 일부집단(스몰데이터) → 전체집단(빅데이터) 추론

- 데이터가 거의 없던 또는 사람이 직접 데이터를 수집하기 위해 제품 팔던 시절
- 예를 들어 일부 마을의 세금정보만으로 제품없이 전체 마을의 세금정보 알고 싶음
- 기술통계 + 추론통계 + 가설검정 방식의 데이터분석 방법론 등장

- 기술통계(**Descriptive Statistics**): 일부 마을의 세금정보 정량적 요약
- 통계추론(**Inferential Statistics**): 전체 마을의 세금정보 통계적 추론
- 가설검정(**Hypothesis Testing**): 추론된 세금정보 믿을만 한지 확률적 의사결정





Q. 빅데이터가 존재하는 디지털시대에는 통계가 필요없는 것 아닌가요?

A. 단순히 양만 많은 데이터를 빅데이터로 부르는 경향이 있는데 전체집단을 반영하지 못하면 결국 스몰데이터와 차이가 없음

예를 들어 1초에 수백개의 데이터가 생성되는 SKT 스마트폰 데이터는 결국 일부집단(한국)이지 글로벌 특성 대표성 없음

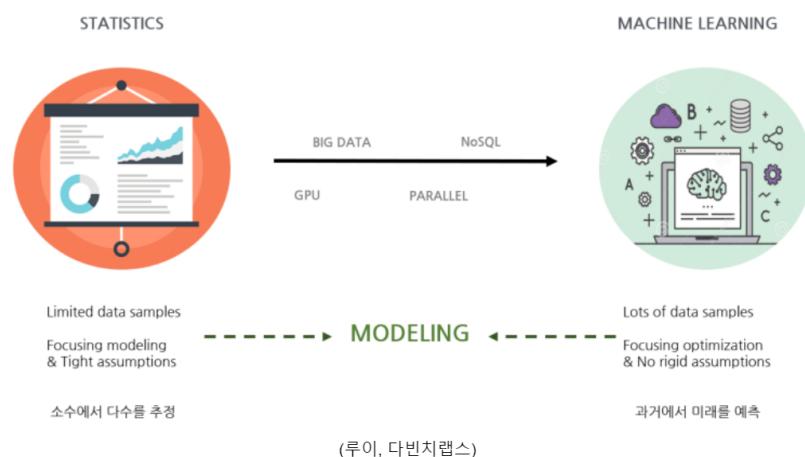
분석 목적상 한국 데이터밖에 없는데 미국 또는 모든 사람들의 특성을 알고 싶으면 기술통계 + 통계추론 + 가설검정 분석과정 필요

만약 모든 사람들의 데이터가 있더라도 일반화 및 분석결과의 실제 활용을 위해 통계추론 + 가설검정 필요

2) 기계학습/인공지능: 알고리즘학습 방식을 사용하여 과거의 특성으로 전체집단(빅데이터) 가정하고 일부 미래특성(스몰데이터) 추론

- 데이터의 생성과 수집이 발전하여 빅데이터를 손쉽게 다루게 되면서 컴퓨터기반 방식의 정교한 데이터분석 방법론 등장

- 기계학습(Machine Learning):** 전체 데이터를 기계가 학습하여 정교한 모델 생성
- 인공지능(Artificial Intelligence):** 알고리즘학습 방식이 고도화되어 스스로 더욱 정확한 모델 생성



Q. 더더욱 통계가 필요없고 인공지능이 다 해주는 것 아닌가요?

A. 단순히 양만 많은 데이터가 아닌 질적으로 다양한 패턴/특성이 반영된 빅데이터인 경우에만 효과 한계성

개인이 다루는 데이터나 일반적인 상황에선 여전히 스몰데이터인 경우가 많기 때문에 인공지능 <<< 통계/기계학습 활용도 높음

통계/기계학습/인공지능 모두 데이터분석 프로세스는 동일하고 인공지능 <<< 통계/기계학습이 단순하여 이해하기 쉽기 때문에 여전히 중요

2 데이터의 종류 및 용어

2.1 데이터 관점에 따른 분류

횡단면 데이터(Cross Sectional)	시계열 데이터(Time Series)	시계열 횡단면 데이터(Pooled Cross Section)	패널 데이터(Panel)
정의	특정시점 + 다수독립변수	다수시점 + 특정독립변수	다수독립변수 + 다수시점 (동일 변수 및 시점)
예시	2016년 16개 시도의 GRDP와 최종소비	연도별 전국 GRDP와 최종소비	연도별 16개 시도의 GRDP와 최종소비
특징	값 독립적, 모집단 중 특정 시점 표본추출	값 Serial-correlation/Trend/Seasonality 등	시점/변수 불일치로 공백 가능 시점/변수 일치로 연구자들이 가장 선호

Sales

Annual sales figures for each company in millions of KRW.

Year	A	시계열 B	C	D 횡단면
2000	1,881	11,296	24,855	6,929
2001	1,900	12,007	23,130	5,693
2002	1,994	12,659	23,519	6,145
2003	15,24	13,091	20,761	6,769
2004	2,107	13,636	22,505	7,902

2.2 데이터 변수구분 및 용어정리

- 원데이터(Raw Data): 수집된 차례로 기록되어 처리되지 않고 순서화되지 않은 그대로 보존된 데이터 (ex. Log, Table)
- 변수(Variable): 정보가 수집되는 특정한 개체나 대상 (보통 열(Column) 값들을 의미)
- 데이터 특성에 따라:

대분류	소분류	의미/예시
질적변수(Qualitative Variable)	-	내부 값이 특정 범주(Category)로 분류된 변수(색상, 성별, 종교)
명목형 변수(Nominal Variable)		값이 순위가 존재하지 않는 경우(혈액형)
순위형 변수(Ordinal Variable)		값이 순위가 존재하는 경우(성적)
양적변수(Quantitative Variable)	-	내부 값이 다양한 숫자 분포로 구성된 변수(키, 몸무게, 소득)
이산형 변수(Discrete Variable)		값이 셀 수 있는 경우(정수)
연속형 변수(Continuous Variable)		값이 셀 수 없는 경우(실수)

- 데이터 관계에 따라: $Y = f(X)$

대분류	의미/예시
독립변수(Independent Variable)	다른 변수에 영향을 미치는 변수 (X)
종속변수(Dependent Variable)	다른 변수에 의해 영향을 받는 변수 (Y)

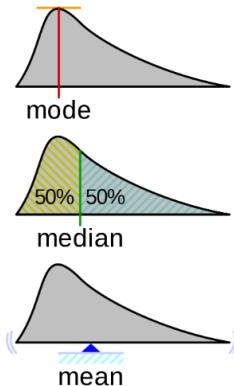
3 데이터 특성확인(Descriptive Statistics)

3.1 중심 통계량: 데이터의 중심경향을 측정

$$E(X) = \mu$$

이산형	$\sum_{x=0}^n E(X = x) = \sum_{x=0}^n xf(x)$	
	$E(X = x) = \int_{-\infty}^{\infty} xf(x)$	

- 평균(Average): 표본데이터의 중심 무게 (산술평균, 기하평균, 조화평균, 가중평균)
- 중앙값(Median): 순서를 가진 표본데이터의 가운데(50%)에 위치한 값
- 최빈값(Mode): 표본데이터 중 가장 빈번한 값



3.2 변동 통계량: 데이터의 변동성을 측정

$$Var(X) = \sigma^2$$

이산형	$E[(X - \mu)^2] = \sum_{x=0}^n (x - \mu)^2 f(x)$	
	$E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)$	

- 범위(Range): 최대값과 최소값의 차이
- 편차(Deviation): 관측값과 평균의 차이
- 변동(Variation): 편차 제곱의 합
- 분산(Variance): 편차 제곱의 합을 데이터의 수로 나눈 값
- 표준편차(Standard Deviation): $\sqrt{\text{분산}}$

Example

Find the standard deviation and variance

x	$x - \bar{x}$	$(x - \bar{x})^2$
30	4	16
26	0	0
22	-4	16

$\left\{ \text{Sum} = 0 \right.$

78

J

32

The variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = 32 \square 2 = 16$$

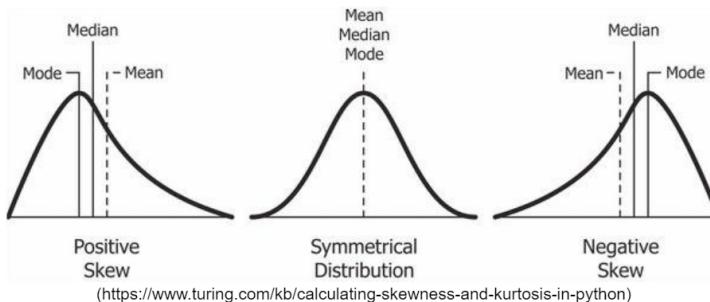
The standard deviation

$$S = \sqrt{16} = 4$$

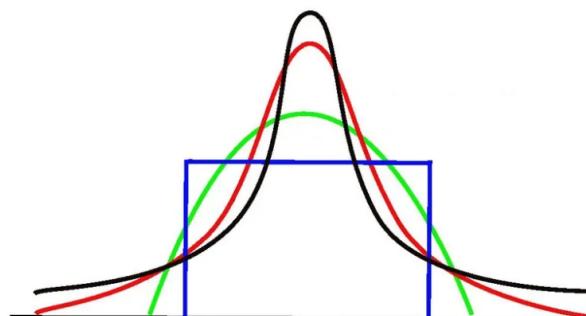
6

3.3 형태 통계량: 데이터의 분포형태와 왜곡을 측정

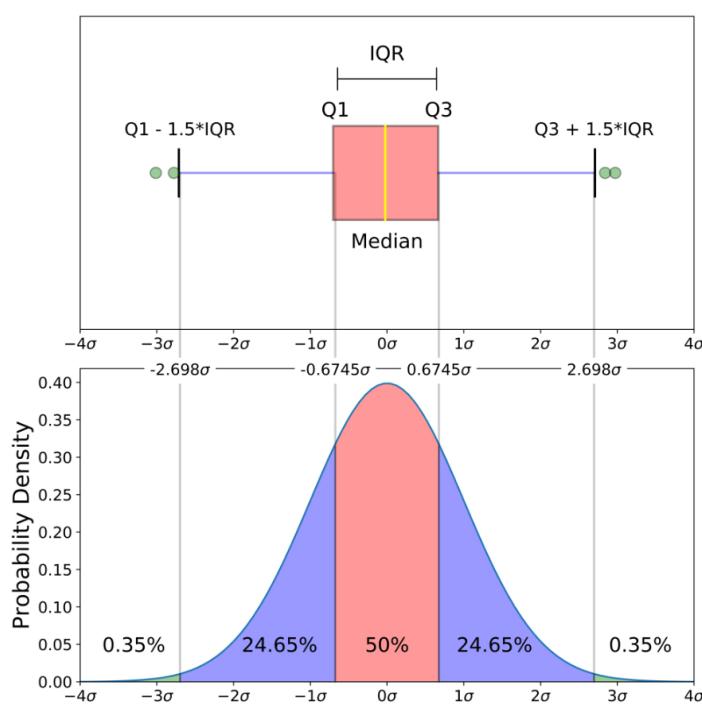
- **왜도(Skewness):** 평균을 중심으로 좌우로 데이터가 편향 되어 있는 정도



- **첨도(Kurtosis):** 뾰족함 정도



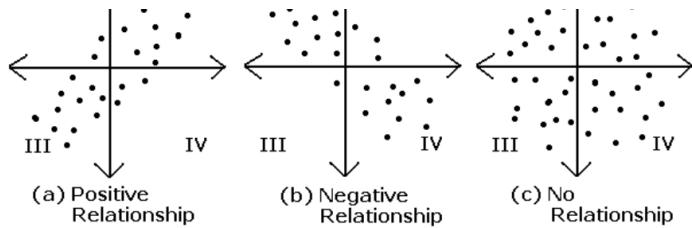
- **이상치(Outlier):** 통계 분포적으로 오류로 판단되는 값이지만, 일반적으로 이상치의 기준이 불명확



3.4 관계 통계량: 데이터간의 관계를 측정

- **공분산(Covariance):** 사건이 2개 이상인 데이터, 즉 확률변수가 2개 이상인 경우 서로 어떤 관련성으로 퍼져있는지 확인





- $Cov(X_1, X_2) > 0$: X_1 이 증가 할 때 X_2 도 증가
- $Cov(X_1, X_2) < 0$: X_1 이 증가 할 때 X_2 도 감소
- $Cov(X_1, X_2) = 0$: X_1, X_2 는 아무런 선형관계가 없이 서로 독립

$$E(X_1) = \mu, E(X_2) = \nu$$

$$Var(X_1) = E[(X_1 - \mu)^2]$$

$$Var(X_2) = E[(X_2 - \nu)^2]$$

$$Cov(X_1, X_2) = E[(X_1 - \mu)(X_2 - \nu)]$$

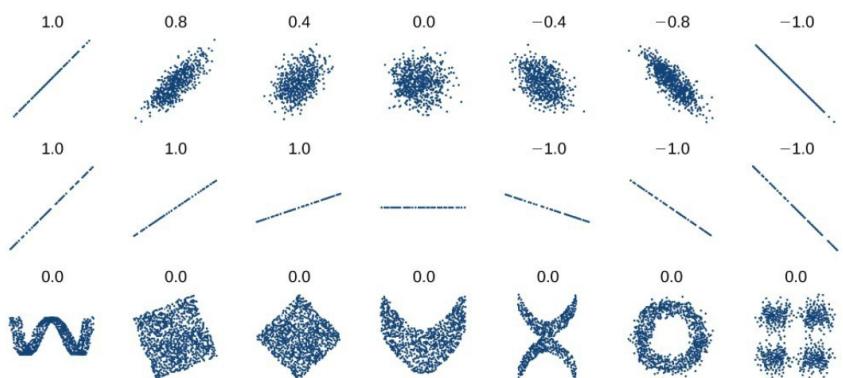
- 이슈: 단위 크기에 영향을 받아, 100점 만점 인 두과목의 공분산 > 10점 만점 인 두과목의 공분산

- 상관관계(Correlation): 각 변수들의 크기에 영향을 받지 않도록 스케일 조정

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum(x - \bar{x})^2}{n-1} \times \frac{\sum(y - \bar{y})^2}{n-1}}}$$

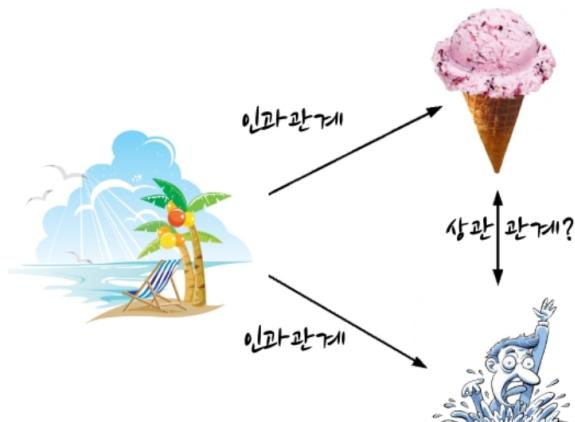
공분산
루트 빼면,
변수 x의 분산
루트 빼면,
변수 y의 분산

- 두 변수가 함께 변하는 정도를 각 변수가 변하는 정도로 나눈 값
- 공분산은 $-\infty \sim \infty$ 의 범위지만 상관계수는 $-1 \sim 1$ 범위
- 변수들이 서로 독립 이면 상관계수는 0
- 양 의 선형관계면 1 이, 음 의 선형관계면 -1 이 출력
- 이상치에 영향을 받고 선형관계만을 측정하기 때문에 데이터가 비선형 이면 부적절한 지표



(<https://www.coursehero.com/study-guides/ivytech-collegealgebra/distinguish-between-linear-and-nonlinear-relations/>)

- 인과관계(Causality): A변수와 B변수중 하나는 원인 이 되고 다른 하나는 결과가 되는 관계 분석



- 예시:

- 아이스크림 판매량 vs 의사자의 수
- 화재 현장에 출동하는 소방대원 수 vs 화재의 규모
- 해적의 수가 감소됨과 동시에 지구 온난화가 증가됨

3.5 예시 및 함정

- 통계량 사용 예시:

```
In [1]: # 데이터로딩
import numpy as np
import pandas as pd
from sklearn.datasets import load_iris
df = pd.DataFrame(load_iris().data, columns=['X1', 'X2', 'X3', 'X4'])
display(df)
executed in 1.36s, finished 12:34:09 2022-07-09
```

	X1	X2	X3	X4
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

150 rows × 4 columns

```
In [17]: # 중심통계량
## 평균
pd.DataFrame(df.mean()).T
executed in 20ms, finished 19:35:35 2022-07-02
```

```
Out[17]: X1      X2      X3      X4
0  5.843333  3.057333  3.758  1.199333
```

```
In [18]: ## 중앙값
pd.DataFrame(df.median()).T
executed in 16ms, finished 19:35:42 2022-07-02
```

```
Out[18]: X1      X2      X3      X4
0  5.8  3.0  4.35  1.3
```

```
In [3]: ## 최빈값
df.mode()
executed in 15ms, finished 12:35:36 2022-07-09
```

```
Out[3]: X1      X2      X3      X4
0  5.0  3.0  1.4  0.2
1  NaN  NaN  1.5  NaN
```

```
In [19]: # 변동통계량
## 분산
pd.DataFrame(df.var()).T
executed in 25ms, finished 19:36:02 2022-07-02
```

```
Out[19]: X1      X2      X3      X4
0  0.685694  0.189979  3.116278  0.581006
```

```
In [20]: ## 표준편차
pd.DataFrame(df.std()).T
executed in 20ms, finished 19:36:12 2022-07-02
```

```
Out[20]: X1      X2      X3      X4
0  0.828066  0.435866  1.765298  0.762238
```

```
In [23]: # 혼태통계량
## 정규분포의 skewness = 0
pd.DataFrame(df.skew()).T
executed in 14ms, finished 19:36:48 2022-07-02
```

```
Out[23]: X1      X2      X3      X4
0  0.314911  0.318966 -0.274884 -0.102967
```

```
In [24]: ## 정규분포의 kurtosis = 3
pd.DataFrame(df.kurt()).T
executed in 18ms, finished 19:37:12 2022-07-02
```

	X1	X2	X3	X4
0	-0.552064	0.228249	-1.402103	-1.340604


```
In [26]: # 관계분석
## 공분산
df.cov()
executed in 34ms, finished 19:40:57 2022-07-02
```

	X1	X2	X3	X4
X1	0.685694	-0.042434	1.274315	0.516271
X2	-0.042434	0.189979	-0.329656	-0.121639
X3	1.274315	-0.329656	3.116278	1.295609
X4	0.516271	-0.121639	1.295609	0.581006

```
In [27]: ## 상관계수
df.corr()
executed in 17ms, finished 19:41:19 2022-07-02
```

	X1	X2	X3	X4
X1	1.000000	-0.117570	0.871754	0.817941
X2	-0.117570	1.000000	-0.428440	-0.366126
X3	0.871754	-0.428440	1.000000	0.962865
X4	0.817941	-0.366126	0.962865	1.000000

- 통계를 이용한 조작: 특정하게 `skew` 된 sample 수집, 임의로 `outlier` 를 정해서 값변경

- [Sample을 편향되게 만드는 방법](#)
- [편향으로 인한 인공지능 알고리즘 이슈](#)
- [Sampling 과정에서 생기는 Bias 제거하는 방법](#)

4 데이터 의사결정(Inferential Statistics + Hypothesis Testing)

4.1 성공적인 데이터기반 의사결정을 위한 3요소

"성공한 사람들 이 기본적으로 갖추고 있는 3요소"

"모든 사람은 비합리적이나 특정 분야에 합리적인 사람들 이 가진 3요소"

"세상에서 선호하는 인재상 이 가진 3요소"

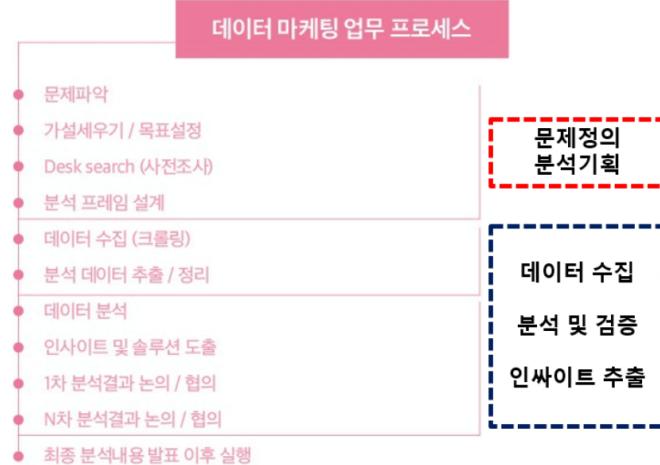
"내일의 내가 오늘의 나보다 발전 하기 위한 3요소"

"나를 돌아보기 보다 남탓을 하는 사람들이 가지지 못한 3요소"

- 성공적 데이터 의사결정 필수 3요소: 문제정의, 분석기획, 성능검증

"데이터분석에 이 굴레들이 어떻게 쓰이는지 정리 해 보면.."

데이터 분석	목적	대응
데이터 시각화	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 데이터를 한눈에 확인 - 점그림, 선그림, 영역그림 - 막대그림, 등고선그림, 분포그림(히스토그램) → 통계 사용 (2) 데이터를 뿌리고 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/컴퓨터 사용
기술적 분석	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 데이터 특성을 몇 개의 숫자들로 확인 → 통계 사용 (2) 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/컴퓨터 사용
상관관계/인과관계	여러종류 데이터끼리의 관계를 알고 싶다	(1) 각 데이터를 몇 개의 숫자들로 표현 → 통계 사용 (2) 표현된 숫자들을 비교 → 확률/통계 사용
통계추론	일부 데이터로 전체를 알고 싶다	(1) 일부 데이터의 특성을 확인 → 통계 사용 (2) 반복적으로 실험 진행 및 통계치 재확인 → 컴퓨터 사용 (3) 전체 특성을 추론 → 통계 사용
알고리즘학습	전체 데이터로 미래를 알고 싶다	(1) 데이터의 관계를 수학적으로 표현 → 확률/통계/함수/컴퓨터 사용 (2) 미래를 예측한 후 정확성 확인 → 확률/통계/컴퓨터 사용
(1) 기존 데이터와 새로운 데이터 비교를 위해 숫자들로 표현		



데이터 마케터가 일하는 법 - 특히 마케팅 전략기획 수립 시 유용한 순서입니다

디지털과 내가 연결되는 순간
DIGITAL INSIGHT

• 우선순위: 문제정의 > 분석기획 > 성능검증

"문제정의 + 분석기획 이 전체 데이터분석 단계에서 가장 중요"

1) 문제정의 : 현실에서 어떤 문제를 풀 것인지 정하는 것

- 우리는 일상 생활 매 순간마다 의사결정이 필요한 문제들을 해결하고 있고, 마찬가지로 비즈니스 연구 등에서도 모든 순간들에는 해결해야 할 문제 존재

2) 분석기획 : 풀어야 하는 문제를 데이터분석 과정에서 어떻게 증명할지 기획

- 데이터분석 모든 단계에서 가장 중요한 것이 문제정의를 증명하는 분석기획 단계
- 문제정의에서 해결해야 할 문제가 없다면 분석기획 어떻게?
- 고민해야 하는 질문 리스트 10가지

- (1) 내가 줄어야 하는 문제가 무엇인지 명확하게 정의하고 제시 할 수 있는가?
- (2) 사람들은 그 문제를 해결하기 위해서 어떤 해결책을 시도하는가?
- (3) 데이터로 그 문제를 해결하려면 어떤 데이터가 필요 한가?
- (4) 해당 데이터를 얻기(수집) 위해서는 어떤 방식이나 채널을 활용하는 것이 가장 적절한가?
- (5) 적절한 방식이나 채널에서는 데이터는 수집이 용이 해서 바로 사용할 수 있는가?
- (6) 데이터는 어떤 고객을 대상으로 어떤 상황 어떤 기간동안 수집할 것인가?
- (7) 수집되는 데이터의 양은 예상했던 수만큼 적절한가? 너무 많거나 적지 않은가?
- (8) 정제된 데이터를 사용해서 문제를 해결하려면 어떤 실험을 기준으로 분석할 것인가?
- (9) 데이터분석 결과, 어떤 인사이트 또는 어떤 정답 후보가 예상 하는가?
- (10) 데이터분석 결과는 실제 비즈니스에 적용 가능 한 솔루션인가?

"데이터 수집 + 분석 및 검증 + 인사이트 추출은 빅데이터와 인공지능을 활용한 다소 수동적 절차"

3) 성능검증 : 데이터분석 과정에서 도출된 의사결정 후보들을 현실에서 검증하는 것

- 데이터분석 산출물이 현실에서 검증되지 못한다면 의미없던 데이터분석 될 것
- 분석기획을 마련하지 못하면 성능검증 어떻게?

4.2 A/B Test: (확률적) 의사결정을 위한 필수과정

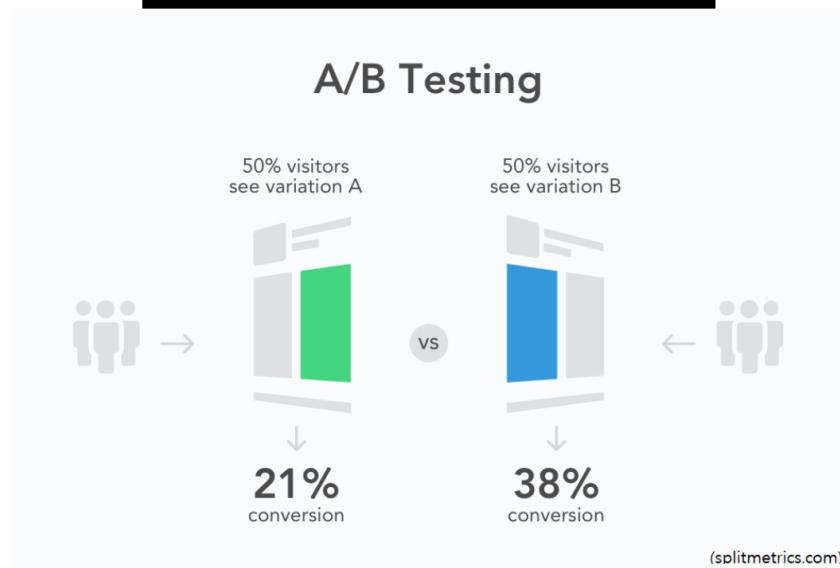
1) A/B Test: 분석기획 & 성능검증 단계에서 필수적인 사고

- 우리는 일상 생활 매 순간마다 의사결정이 필요한 문제들을 해결하고 있고, 마찬가지로 비즈니스 연구 등에서도 모든 순간들에는 해결해야 할 문제 존재
- 모든 (데이터)분석은 누구나 할 수 있는 비교(A/B Test) 기반 분석기획 포함됨
- 정답을 모르는 상황에서, 비교 후보/보기를 두어 더 나은 후보/보기 선택하는 방법
- 비교대상은 제외한 다른 조건들은 모두 동일하게 두어야 공평한 비교 방법



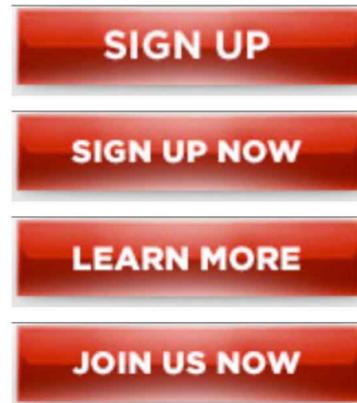


A/B Testing



2) 예시:

가설1: 어떤 버튼이 기부를 증가시킬까?



(<https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/>)

가설2: 어떤 컨텐츠가 기부를 증가시킬까?



가설1: 어떤 버튼을 넣을까?

가설2: 어떤 컨텐츠를 넣을까?

Combinations (24)		Page Sections (2)		Download: XML CSV TSV Print		
Relevance Rating	Variation	Est. conv. rate		Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Button	Original	7.51% ± 0.2%	—	—	—	5851 / 77858
	Learn More	8.91% ± 0.2%	+	100%	18.6%	6927 / 77729
	Join Us Now	7.62% ± 0.2%	+	73.5%	1.37%	5915 / 77644
	Sign Up Now	7.34% ± 0.2%	+	13.7%	-2.38%	5660 / 77151
Media	Original	8.54% ± 0.2%	—	—	—	4425 / 51794
	Family Image	9.66% ± 0.2%	+	100%	13.1%	4996 / 51696
	Change Image	8.87% ± 0.2%	+	92.2%	3.85%	4595 / 51790
	Barack's Video	7.76% ± 0.2%	—	0.04%	-9.14%	3992 / 51427
	Sam's Video	6.29% ± 0.2%	—	0.00%	-26.4%	3261 / 51864
	Springfield Video	5.95% ± 0.2%	—	0.00%	-30.3%	3084 / 51811

(<https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/>)

가설1: 어떤 버튼을 넣을까?

가설2: 어떤 컨텐츠를 넣을까?

Combinations (24)		Page Sections (2)		Download: XML CSV TSV Print		
Combination	Status	Est. conv. rate		Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Original	Enabled	8.26% ± 0.5%	→	—	—	1088 / 13167
★ Top high-confidence winners. Run a follow-up experiment ↗						
Combination 11	Enabled	11.6% ± 0.6%	→	100%	40.6%	1504 / 12947
Combination 7	Enabled	10.3% ± 0.6%	→	100%	24.0%	1340 / 13073
Combination 3	Enabled	9.80% ± 0.6%	→	99.7%	18.7%	1277 / 13025
Combination 10	Enabled	9.23% ± 0.6%	→	95.9%	11.7%	1203 / 13031
Combination 8	Enabled	9.03% ± 0.6%	→	91.6%	9.28%	1178 / 13046
Combination 9	Enabled	8.77% ± 0.6%	→	81.8%	6.10%	1111 / 12672
Combination 6	Enabled	8.64% ± 0.5%	→	75.3%	4.58%	1108 / 12822

(<https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/>)

가설: 어떤 화면이 기부 결제율을 높일까?

CONTROL

"SEQUENTIAL"

↑ +5%

<http://bit.ly/obama>

(<https://www.optimizely.com/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/>)

가설: 어떤 이미지가 가입율을 높일까?

CONTROL

IMAGE VARIATION

↑ +19%

<http://bit.ly/obama>

(<https://mirakle.mk.co.kr/view.php?year=2017&no=347786>)

CONTROL

IMAGE VARIATION

↑ +19%

<http://bit.ly/obama>

- 성능검증을 위해선 결국 산출 결과의 해석 능력이 필수적

OLS Regression Results

Dep. Variable:	Model:	R-squared:
count	OLS	0.390
Method:		Adj. R-squared:
Logit		0.390

Method:		Least Squares		F-statistic:		893.0					
Date:		Sat, 16 Feb 2019		Prob (F-statistic):		0.00					
Time:		02:13:49		Log-Likelihood:		-81290.					
No. Observations:		13003		AIC:		1.626e+05					
Df Residuals:		12988		BIC:		1.627e+05					
Df Model:		14									
Covariance Type:											
nonrobust											
		coef	std err	t	P> t	[0.025 0.975]					
season		12.3801	2.305	5.371	0.000	7.862 16.899					
holiday		-21.4220	8.249	-2.597	0.009	-37.592 -5.252					
workingday		0.2884	3.991	0.072	0.942	-7.535 8.112					
weather		-7.1833	1.938	-3.706	0.000	-10.982 -3.384					
temp		1.8075	1.184	1.526	0.127	-0.514 4.129					
atemp		4.5848	1.083	4.235	0.000	2.463 6.707					
humidity		-1.6066	0.069	-23.305	0.000	-1.742 -1.471					
windspeed		0.4438	0.148	3.001	0.003	0.154 0.734					
Year		-0.0033	0.004	-0.742	0.458	-0.012 0.005					
Quater		-32.9994	4.584	-7.199	0.000	-41.985 -24.014					
Quater_ver2		20.8358	0.662	31.487	0.000	19.539 22.133					
Month		5.5936	1.376	4.066	0.000	2.897 8.290					
Day		-0.2107	0.126	-1.667	0.096	-0.458 0.037					
Hour		6.8639	0.169	40.695	0.000	6.533 7.195					
DayofWeek		0.6025	0.915	0.658	0.510	-1.191 2.396					
Omnibus:	2395.271	Durbin-Watson:	0.552								
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4484.367								
Skew:	1.146	Prob(JB):	0.00								
Kurtosis:	4.739	Cond. No.	1.57e+04								

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.57e+04. This might indicate that there are strong multicollinearity or other numerical problems.

3) 사용 직군: 2013년 댄 시로커와 피트 쿠메이 쓴 A/B 테스트 책 인용

- (1) 디지털 마케터와 마케팅 매니저
- (2) 디자이너
- (3) 프로덕트 매니저
- (4) 소프트웨어 엔지니어
- (5) 창업가
- (6) 카피라이터
- (7) 그로스해커
- (8) 데이터 사이언티스트

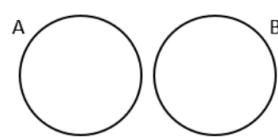
=> 제품이나 서비스를 제작/생산/유통하는 모든 사람은 A/B 테스트로 성능 검증

4.3 가설설정: 분석기획 필수조건 3가지

- 우선순위: 문제정의 > 분석기획 > 성능검증

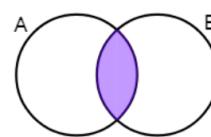
1) 상호배반적(Mutually Exclusive): 대중주장(A) 과 나의주장(B) 은 모호함 없이 독립적 이어야 하며 더하면 다른주장은 없어야 함

Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B)$$

Non-Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

문제정의	양치기들이 거짓말쟁이인가?
분석기획 대중주장(A)	양치기들은 거짓말쟁이가 아니다!
나의주장(B)	양치기들은 거짓말쟁이다!

- 이슈: 거짓말쟁이 를 어떻게 정의하지? 어느 수준이 거짓말쟁이라는거지?

문제정의	양치기들이 거짓말쟁이인가?
분석기획 대중주장(A)	모든 양치기들이 다 거짓말쟁이는 아니다!
나의주장(B)	양치기들은 다 거짓말쟁이다!

2) 증명가능성(Demostrable): 성급한 일반화에 빠지지 않으려면 증명 가능한 것 이나 범위로 제시

- 이슈: 모든 양치기들을 확인하기도 어렵고 일부 양치기들 중에는 거짓말쟁이가 아닌 양치기 도 있을 수 있음
- 이슈: 모든 양치기를 조사후 거짓말쟁이가 없다 하더라도 과거에는 거짓말 했을 수도 있음

문제정의	양치기들이 거짓말쟁이인가?
분석기획 대중주장(A)	현재 대한민국에 있는 양치기들은 일반적으로 더 거짓말을 하는 경향이 있는 않다!
나의주장(B)	그들은 일반적으로 거짓말하는 경향이 있다!

3) 구체적(Specific): 충분히 구별되고 실현가능한 표현으로 정의되어야 함

문제정의	양치기들이 거짓말쟁이인가?
분석기획 대중주장(A)	현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 하지 않는다!
나의주장(B)	현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 한다!

=> 만약 분석기획(가설설정)이 변경되면 모든 데이터, 분석 과정 및 성능검증 과정이 모두 변경됨!

4.4 가설검정: 분석기획 추론방법

분류/방법	용어	의미/예시
데이터 분류	모집단 (Population)	전체 집단으로 관심(연구) 대상
	샘플 (Sample)	모집단에서 선택된 일부 집단
데이터 분류 별 조사방법	전수조사 (Population Scale Test)	모집단을 조사/분석 하는 것 시간과 비용이 가장 많이 소요되는 방식 (ex. 인구주택총조사)
	샘플조사 (Sample Scale Test)	샘플을 조사/분석 하는 것 시간과 비용을 줄일 수 있으나 편향(Bias) 문제 존재 (ex. 출구조사, 여론조사)

1) 통계추론(Statistical Inference): 샘플을 분석하여 모집단의 특성을 추론하고 신뢰성 검정하는 것

분류/방법	용어	의미/예시
데이터 분류	모집단 (Population)	현재 전 세계 사람
	샘플 (Sample)	현재 대한민국 사람
	샘플1 (Sample)	현재 대한민국 양치기들
샘플2 (Sample)	현재 대한민국 일반인들	

- 최근엔, 보유한 데이터는 샘플, 보유하지 못한 미래의 데이터를 모집단으로 인식
- 샘플으로 모집단을 추정하기 때문에 샘플의 특성이 모집단을 잘 반영해야
- 샘플의 기초통계로 데이터 분포 확인하고, 분포에 따라 분석 방법론 달라짐
- 통계량(Statistic): 샘플의 특성을 측정한 수치
- 모수(Parameter): 샘플의 통계량으로 추론한 모집단의 특성
- 샘플오차(Sampling Error): 모집단의 특성과 샘플의 특성 차이

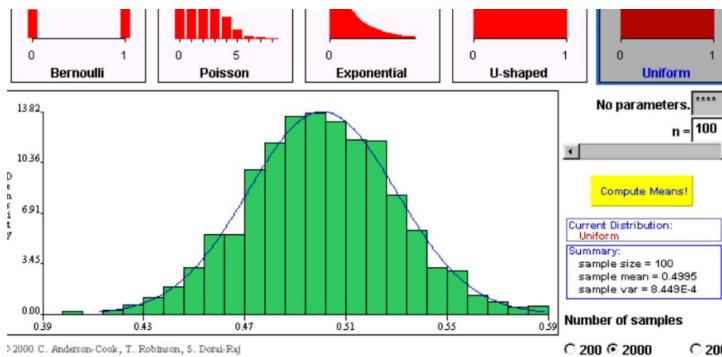
- 샘플의 평균으로 모집단의 평균을 추론할 경우, 모평균-샘플평균
- 모평균 추론시 여러번의 샘플추출의 평균 통계량으로 추론
- 절대 샘플평균이 모평균 자체/같음을 의미하지 않음!

2) 중심극한정리(Central Limit Theorem): 통계학과 데이터 분석의 기반 이 되는 이론으로 분석을 통해 인사이트를 추론할 수 있는 기반

"전체 데이터에서 추출된 샘플들의 평균들을 구하면 전체 데이터의 특성을 알 수 있고, 이 평균들의 분포는 정규분포"

(1) 데이터가 어떻게 생겼든(정규분포든 아니든) 어떠한 사건이라도 반복적으로(빅데이터 일수록) 추정하면 정규분포의 확률! [Simulation](#)





(2) 실험반복 횟수가 증가하면 기대되는 발생확률(평균)은 이론적/일반화/상식적 수치와 같아짐

(3) 실험반복 횟수가 증가하면 기대되는 발생확률을 오차(분산)는 작아져서 평균 정확성 향상

- 동전을 1회 던지면 어떤 면이 나오게 될까?

$$\bar{X} \xrightarrow{\text{a.s.}} N(\mu, \frac{\sigma^2}{n})$$

- 추론통계상 나의주장이 맞다면:

문제정의	양치기들이 거짓말쟁이인가?
분석기획 대중주장(A) 현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 하지 않는다!	
나의주장(B) 현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 한다!	

- 모든 양치기들을 조사하지 않더라도 Sample로 추출한 양치기의 거짓말 횟수 평균은 일반인의 거짓말 횟수 평균과 비교했을 때 차이가 있어야 함
- 일반인의 거짓말 횟수 평균 점에서 양치기의 거짓말 횟수가 발생할 확률은 낮아서, 양치기의 거짓말 횟수 데이터는 주목해야 하고 횟수 차이가 있다고 판단해야
- 일반인과 양치기의 거짓말 횟수 차이가 높을 확률이 낮아서, 양치기의 거짓말 횟수 데이터는 주목해야 하고 횟수 차이가 있다고 판단해야



(https://bookdown.org/mathemedicine/Stat_book/)

3) 가설검정 방법 3단계:

(1) 가설 설정

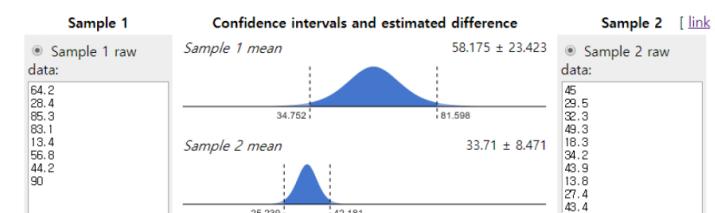
가설종류	주장종류	양치기들이 거짓말쟁이인가?
귀무가설 (Null Hypothesis, H_0): 현재의 상황이나 통념	대중주장(A)	현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 하지 않는다!
대립가설 (Alternative Hypothesis, H_1): 나의주장(B) 새로운 현상이나 주장		현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 한다!

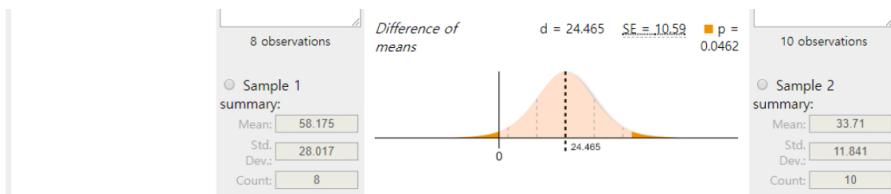
(2) 검정통계량 추정 및 유의수준 설정

- 검정통계량(Test Statistics): 귀무가설(대중주장)과 대립가설(나의주장)을 비교하기 위한 검증(Evaluation) 통계량으로, 1회의 추정치 라 점 추정이라고도 함

$$\frac{\text{양치기 거짓말 빈도 평균} - \text{일반인 거짓말 빈도 평균}}{\text{일반인 거짓말 빈도 평균}}$$

- 양치기와 일반인 거짓말 빈도 차이가 없다면, 검정통계량은 0에 가까울 것
- 양치기와 일반인 거짓말 빈도 차이가 있다면, 검정통계량은 0보다 커질 것
- 검정통계량에 두 집단의 차이를 반영하는 이유: 두 샘플 집단을 각각 분석하기 보다, 차이만 분석하면 훨씬 단순한 모형
- 신뢰구간(Confidence Interval): 검정통계량을 여러 횟수로 추정한 범위로 구간추정이라고도 함





- 유의수준(Significant Level, α): 분석가가 직접 설정 한 오류 허용치

"귀무가설이 참인데 잘못된 데이터 또는 실험으로 분석결과상 귀무가설이 틀렸다고 주장하게 될 분석가가 직접 설정한 허용오류 최대치"

- 유의수준 5%: 양치기와 일반인의 거짓말 차이가 없다는 가정하에, 혹시나 데이터 또는 실험으로 귀무가설이 틀렸다고 주장할 오류로 통상 5%를 많이 사용하여 100번 중에 5번을 의미

(3) 의사결정

- 유의확률(p-value): 컴퓨터가 직접 계산 해주는 오류치

"그동안의 실험 데이터에서 대립가설이 발생할 확률"

- 유의확률 10% > 유의수준 5%: 대립가설 발생확률 > 허용오류 최대치

→ 대립가설은 귀무가설과 차이가 없는 정도니 귀무가설 참! (거짓말 차이 없음)

- 유의확률 1% < 유의수준 5%: 대립가설 발생확률 < 허용오류 최대치

→ 대립가설은 귀무가설과 차이가 있으니 대립가설 참! (거짓말 차이 있음)

4.5 예시: 분석기획 및 추론

1) 이해문제1: 양치기들이 거짓말쟁이인가?

(1) 가설 설정

- 대중주장: 현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 하지 않는다!
- 나의주장: 현재 대한민국에 있는 양치기들은 일반인 대비 거짓말을 많이 한다!

(2) 검정통계량 추정 및 유의수준 설정

- 검정통계량(점추정): $\frac{\text{양치기 거짓말 빈도} - \text{일반인 거짓말 빈도}}{\text{양치기 거짓말 빈도 표준편차}}$ (1회성)
- 신뢰구간(구간추정): 실험을 여러번 반복해서 거짓말차이(검정통계량)의 히스토그램 또는 분포 (반복성)
- 유의수준: 양치기와 일반인이 거짓말 차이가 없는데, 차이가 있다고 오류를 범할 확률

(3) 의사결정: (유의수준 5% 기준)

- 유의확률: 양치기와 일반인이 거짓말 차이가 없는데, 실제 실험에서 차이가 있다고 나타날 확률
- 나의주장 참: 허용오류가 5% 인데, 나의 실험과 데이터에서 나타날 오류는 3%로 더욱 낮기에 나의 주장은 신뢰성이 있고 양치기들은 거짓말쟁이!
- 대중주장 참: 허용오류가 5% 인데, 나의 실험과 데이터에서 나타날 오류는 7%로 더욱 높기에 나의 주장은 신뢰할 수 없고 대중의 주장대로 양치기들은 거짓말쟁이가 아님!

2) 이해문제2: (논문읽기 A/B Test) 내 알고리즘의 성능은 좋은가?

(1) 가설 설정

- 대중주장: 지금까지 존재하는 알고리즘의 정확성은 최대 80%
- 나의주장: 내가 만든 알고리즘의 정확성은 90%

(2) 검정통계량 추정 및 유의수준 설정

- 검정통계량(점추정): 지금까지 존재하는 알고리즘들로 나올 수 있는 정확성 (1회성)
- 신뢰구간(구간추정): 정확성을 여러번 반복해서 계산 시 정확성의 히스토그램 또는 분포 (반복성)
- 유의수준: 일반적인 알고리즘 정확성이 최대 80%인데, 알고리즘 정확성이 80% 이상 나올 수 있는 허용오류

(3) 의사결정: (유의수준 5% 기준)

- 유의확률: 일반적인 알고리즘 정확성이 최대 80%인데, 나의 실험에서 정확성이 80% 이상이 관찰될 확률
- 나의주장 참: 허용오류가 5% 인데, 나의 실험에서 80% 이상 정확성이 관찰될 확률은 3%로 더욱 희박하니 90%라는 알고리즘 정확성을 신뢰할만 하고 내가 만든 알고리즘은 뛰어난 알고리즘!
- 대중주장 참: 허용오류가 5% 인데, 나의 실험에서 80% 이상 정확성이 관찰될 확률은 7%로 높은 오류이니 90%라는 알고리즘 정확성을 신뢰하기 어렵고 내가 만든 알고리즘은 일반적인 알고리즘!

3) 심플정리1: 에너XXX 건전지 수명이 듀XX 보다 길다?

(1) 가설 확인

- 대중주장(H0): 에너XXX 수명 = 듀XX 수명
- 나의주장(H1): 에너XXX 수명 > 듀XX 수명

(2) 유의수준 설정 및 유의확률 확인

- 유의수준: 5%
- 유의확률: 1% (H_0 가 참이란 가정에, 건전지 평균 수명(검정통계량) 100개를 실험)

(3) 의사결정

- 유의수준 > 유의확률: 나의주장 참!
-> 에너XXX 수명이 더 김
- **유의수준 < 유의확률:** 대중주장 참!
-> 에너XXX 수명이 더 길지 않음

4) 심플정리2: 숟가락을 잘 구부리는 나는 초능력자다?

(1) 가설 확인

- 대중주장(H0): 내 능력 = 다른 사람의 능력
- 나의주장(H1): 내 능력 > 다른 사람의 능력

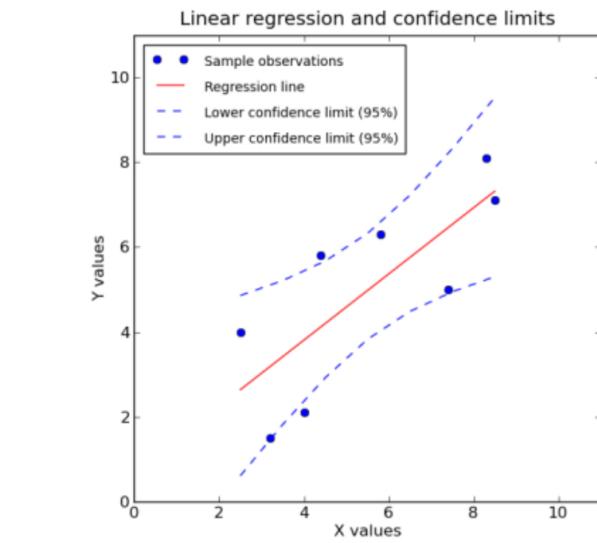
(2) 유의수준 설정 및 유의확률 확인

- 유의수준: 5%
- 유의확률: 8% (H_0 가 참이란 가정에, 숟가락 구부린 횟수(검정통계량) 100명과 비교)

(3) 의사결정

- 유의수준 > 유의확률: 나의주장 참!
-> 나는 초능력자!
- **유의수준 < 유의확률:** 대중주장 참!
-> 나는 일반인!

5) 현실문제1: 회귀분석 (공유자전거의 수요는 어떤 변수가 영향을 미치나?)



OLS Regression Results

Dep. Variable:	count	R-squared:	0.390
Model:	OLS	Adj. R-squared:	0.390
Method:	Least Squares	F-statistic:	593.6
Date:	Sat, 16 Feb 2019	Prob (F-statistic):	0.00
Time:	02:13:49	Log-Likelihood:	-81290.
No. Observations:	13003	AIC:	1.626e+05
Df Residuals:	12988	BIC:	1.627e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
season	12.3801	2.305	5.371	0.000	7.862	16.899
holiday	-21.4220	8.249	-2.597	0.009	-37.592	-5.252
workingday	0.2884	3.991	0.072	0.942	-7.535	8.112

weather	-7.1833	1.938	-3.706	0.000	-10.982	-3.384
temp	1.8075	1.184	1.526	0.127	-0.514	4.129
atemp	4.5848	1.083	4.235	0.000	2.463	6.707
humidity	-1.6066	0.069	-23.305	0.000	-1.742	-1.471
windspeed	0.4438	0.148	3.001	0.003	0.154	0.734
Year	-0.0033	0.004	-0.742	0.458	-0.012	0.005
Quater	-32.9994	4.584	-7.199	0.000	-41.985	-24.014
Quater_ver2	20.8358	0.662	31.487	0.000	19.539	22.133
Month	5.5936	1.376	4.066	0.000	2.897	8.290
Day	-0.2107	0.126	-1.667	0.096	-0.458	0.037
Hour	6.8639	0.169	40.695	0.000	6.533	7.195
DayofWeek	0.6025	0.915	0.658	0.510	-1.191	2.396

Omnibus:	2395.271	Durbin-Watson:	0.552
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4484.367
Skew:	1.146	Prob(JB):	0.00
Kurtosis:	4.739	Cond. No.	1.57e+04

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.57e+04. This might indicate that there are strong multicollinearity or other numerical problems.

6) 현실문제2: 출구조사

- (방송표현) "출구조사 결과 A후보의 지지율은 40%로 추정되며, 95% 신뢰구간에서 +-3%의 오차가 발생할 수 있습니다"
- (통계표현) "샘플링을 통해 A후보의 지지율(검정 통계량)의 평균값은 40%(점 추정)이며, 실제 모집단 확대시 A후보의 지지율이 37%~43%(구간 추정)에 있을 확률이 샘플링 100번 중 95번 정도다"