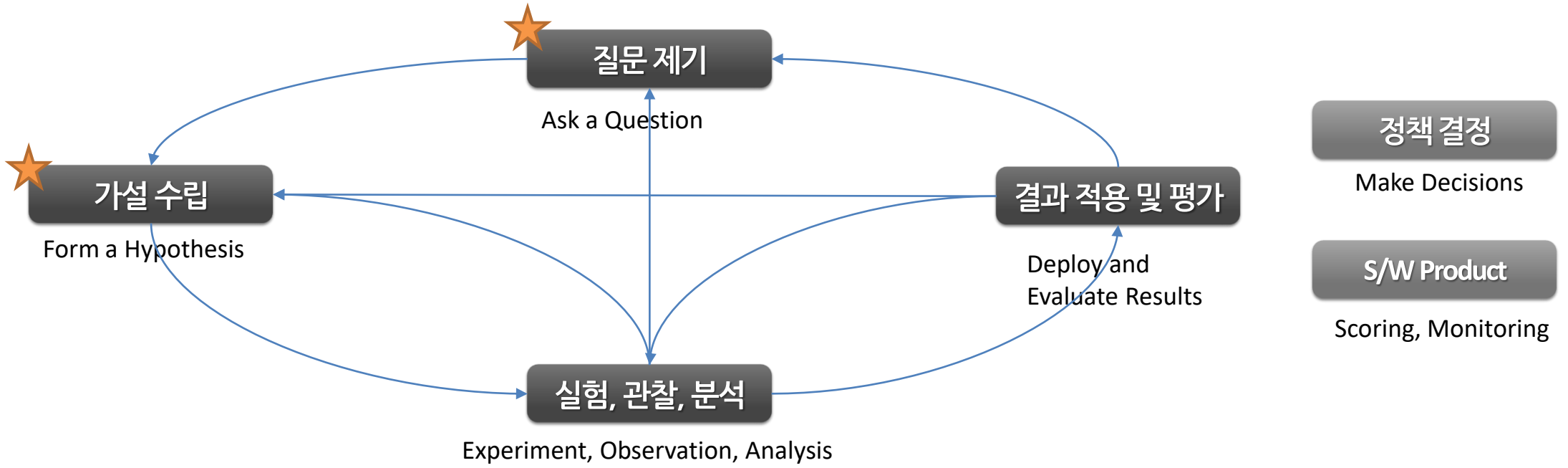
# 통계적 추론
# (Introduction to Statistical Inference)

부산대학교 정보·의생명공학대학
정보컴퓨터공학부

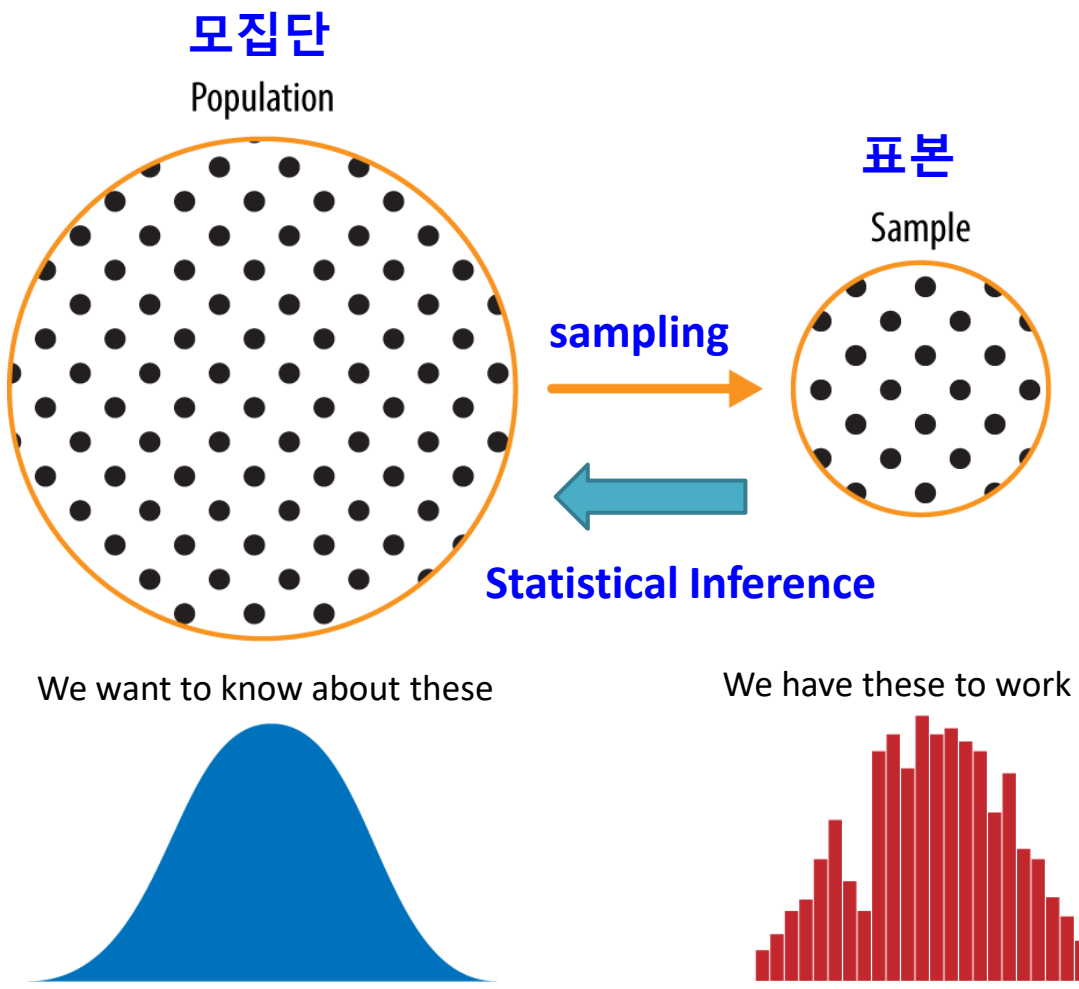
# Data Science Process

❖ 데이터 과학은 과학적 방법으로 데이터를 탐색하여 의미나 통찰을 발견하는 과정

- 일반 과학 분야의 탐색 및 발견 과정 (Process of Scientific Exploration and Discovery)을 따름 : ask-hypothesize-implement/test-evaluate




# Data and Sampling Distribution




❖ The end of a need for sampling ?

- Big Data 시대가 되면서 더는 표본 추출(Sampling)이 필요 없을 거라고 오해하는 사람들이 많다. 하지만 데이터의 질과 적합성을 일정 수준 이상으로 담보할 수도 없으면서 데이터의 크기만 늘어나는 것이 오늘날의 상황이다. 이런 상황에서 오히려 다양한 데이터를 효과적으로 다루고 데이터 편향을 최소화하기 위한 방법으로 표본 추출의 필요성이 더 커지고 있다. 아무리 Big Data 프로젝트라고 해도, 결국 작은 표본(Sample) 데이터를 가지고 예측 모델을 개발하고 테스트한다. 표본은 다양한 종류 (웹 페이지 디자인 클릭 수에 미치는 효과 비교 같은)의 테스트에 쓰인다

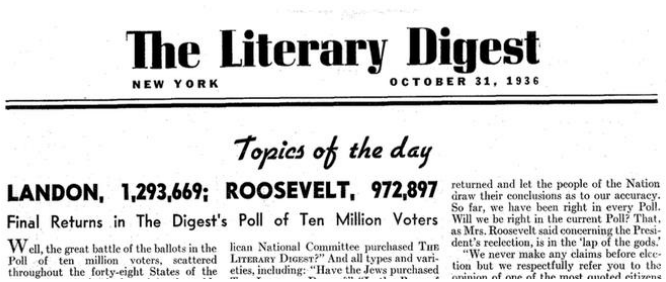
  - A popular misconception holds that the era of big data means the end of a need for sampling. In fact, the proliferation of data of varying quality and relevance reinforces the need for sampling as a tool to work efficiently with a variety of data and to minimize bias. *Even in a big data project, predictive models are typically developed and piloted with samples*. Samples are also used in tests of various sorts (e.g., comparing the effect of web page designs on clicks).

# Random Sampling and Sample Bias

## ❖ 랜덤 표본추출(Random Sampling)

- 모집단 내의 선택 가능한 원소들을 무작위로 추출하는 과정으로 각 추출에서 모든 원소는 동일한 확률로 뽑힌다. 그 결과로 얻은 표본을 **단순랜덤표본** (Simple Random Sample)이라 한다.
  - Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a simple random sample.
- 표본추출은 추출된 원소를 다시 모집단에 포함시켜 재추출을 허용하는 **복원추출**(with replacement)로 수행되거나 한번 뽑힌 원소는 추후에 사용하지 않는 **비복원추출**(without replacement)로 수행될 수 있다.
  - Sampling can be done with replacement, in which observations are put back in the population after each draw for possible future reselection. Or it can be done without replacement, in which case observations, once selected, are unavailable for future draws.

- 표본 기반 추정이나 모델링에서 표본 데이터의 품질이 데이터의 양보다 대체로 더 중요하다. 데이터 과학에서 데이터 품질이란 완결성, 형식의 일관성, 깨끗함 및 각 데이터 값의 정확성을 뜻한다. 통계학은 여기에 **대표성**(representativeness)이라는 개념을 추가한다.
  - Data quality often matters more than data quantity when making an estimate or a model based on a sample. Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. Statistics adds the notion of representativeness.
- 표본 대표성 관련 대표 사례
  - 1936년 미국 대통령 선거 설문 조사
  - Literary Digest : 1,000만명이 넘는 사람을 대상으로 설문 조사 – LANDON의 압도적 승리 예측 :
    - 구독자 대상 설문, **표본 편향 (Sample Bias)**
  - Gallup : 2000명 대상 여론 조사 – Roosevelt의 승리 예측



The Literary Digest

NEW YORK — OCTOBER 31, 1936

*Topics of the day*

**LANDON, 1,293,669; ROOSEVELT, 972,897**

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

*George Gallup, catapulted to fame by the Literary Digest's "big data" failure*

# Size versus Quality : When Does Size Matter ?

## ❖ 작은 데이터가 더 유리하다 ?

- Big Data의 시대라고 해도 의외로 데이터 개수가 적을수록 더 유리한 경우가 있다. 랜덤표본추출에 시간과 노력을 기울일수록 편향이 줄 뿐만 아니라 데이터 탐색 및 데이터 품질에 더 집중할 수 있다. 예를 들어 결측값이나 특잇값으로부터 유용한 정보를 얻을수 있다. 몇백만 개 데이터 중에서 결측치를 추적하거나 특잇값을 평가하는 것은 어려울 수 있지만, 수천개의 sample에서는 가능할 수 있다. 데이터가 너무 많을 경우, 데이터를 일일이 손으로 조사하고 검사하기는 매우 어렵다.
    - In the era of big data, it is sometimes surprising that smaller is better. Time and effort spent on random sampling not only reduces bias but also allows greater attention to data exploration and data quality. For example, missing data and outliers may contain useful information. It might be prohibitively expensive to track down missing values or evaluate outliers in millions of records, but doing so in a sample of several thousand records may be feasible. Data plotting and manual inspection bog down if there is too much data.

## ❖ 언제 대량의 데이터가 필요할까?

- Big Data가 유용한 전형적 상황은 데이터가 크고 동시에 희박할 때이다. 구글이 입력 받은 검색 쿼리를 처리하는 상황을 생각해보자. 행렬을 만들어 열은 용어를, 행은 개별 검색 쿼리를 의미하고 쿼리에 해당 용어가 포함되는지 여부에 따라 원소의 값이 0 또는 1이 된다고 하자. 목표는 주어진 쿼리에 대해 가장 잘 예측된 검색 대상을 결정하는 것이다. 영어 단어는 150,000개가 넘으며 구글은 연간 1조 이상의 검색어를 처리한다. 따라서 대부분 원소가 0인 거대한 행렬이다.
    - The classic scenario for the value of big data is when the data is not only big but sparse as well. Consider the search queries received by Google, where columns are terms, rows are individual search queries, and cell values are either 0 or 1, depending on whether a query contains a term. The goal is to determine the best predicted search destination for a given query. There are over 150,000 words in the English language, and Google processes over one trillion queries per year. This yields a huge matrix, the vast majority of whose entries are “0.”
- 이는 방대한 양의 데이터가 누적될 때만 대부분의 쿼리에 대해 효과적인 검색 결과를 반환할 수 있는, 진정한 의미의 빅데이터 문제이다. 더 많은 데이터가 축적될 수록 결과가 더 좋을 수 밖에 없다. 인기 검색어의 경우 전혀 문제가 되지 않는다. 현대 검색 기술의 진정한 가치는 백만 번에 한 번 정도 발생하는 검색 쿼리까지도 포함하여 다양한 검색 쿼리에 대해 상세하고 유용한 결과를 얻을 수 있다는 데 있다.
    - This is a true big data problem—only when such enormous quantities of data are accumulated can effective search results be returned for most queries. And the more data accumulates, the better the results. For popular search terms this is not such a problemThe real value of modern search technology lies in the ability to return detailed and useful results for a huge variety of search queries, including those that occur with a frequency, say, of only one in a million.

# Big Data ?

## ❖ "Big"은 움직이는 표적이다.

- 절대적 기준(예 : 1 petabyte)으로 Big Data를 정의하는 것은 무의미해 보인다. 그렇게 정의하면 데이터 크기가 문제될 때만 'Big'이라고 부를 가치가 있을 텐데, 그 경우 'Big'은 데이터 양이 최첨단 컴퓨터의 계산 능력(메모리, 저장 용량, 복잡성, 처리 속도)을 넘어서는 경우를 지칭하는 상대적인 용어가 된다. 그러면 1970년대의 'Big'은 오늘날의 'Big'과 다른 의미를 지니게 될 것이다.
  - **"Big" is a moving target**. Constructing a threshold for Big Data such as 1 petabyte is meaningless because it makes it sound absolute. Only when the size becomes a challenge is it worth referring to it as "Big." So it's a relative term referring to when the size of the data outstrips the state-of-the-art current computational solutions (in terms of memory, storage, complexity, and processing speed) available to handle it. So in the 1970s this meant something different than it does today.

## ❖ "Big"은 하나의 기계로 처리할 수 없을 때 적용한다.

- 각양각색의 사람과 기업은 다양한 수준의 컴퓨팅 자원을 가지고 있다. 만약 데이터가 너무 방대해서 한 대의 컴퓨터로는 처리할 수 없고 추가적으로 새로운 도구와 방법이 적용되어야 한다면 그 데이터는 "Big"이라고 간주할 수 있다.
  - **"Big" is when you can't fit it on one machine**. Different individuals and companies have different computational resources available to them, so for a single scientist data is big if she can't fit it on one machine because she has to learn a whole new host of tools and methods once that happens.

## ❖ Big Data는 문화적 현상이다.

- Big Data는, 얼마나 많은 데이터가 생활의 일부를 이루고 있으며 그것이 가속된 기술 발전으로 인해 촉발되었는지 보여 준다.
  - **Big Data is a cultural phenomenon**. It describes how much data is part of our lives, precipitated by accelerated advances in technology.

## ❖ 4V

- 용량(Volume), 다양성(Variety), 속도(Velocity), 가치(Value). 많은 사람은 Big Data를 규정하는 하나의 방식으로 이 용어를 유포하고 있다. 여러분이 하고자 하는 것을 이 개념에서 구하라.
  - **The 4 Vs:** Volume, variety, velocity, and value. Many people are circulating this as a way to characterize Big Data. Take from it what you will.

# 통계적 추론의 주요 개념


Silicon Valley
실리콘밸리
데이터과학자가
알려주는
Data Scientist
따라 하며 배우는
데이터 과학
Data Science
Jpub

# 학습 목표

- 학습 목표 – 올바른 분석을 위한 통계 기본 개념의 이해

  Objectives – understanding basic concepts for statistical analysis

- p 값이란 ? (p value)
- 신뢰구간이란 ? (confidence interval)
- 표본분포란 ? (sampling distribution)

# 예제 : 수면제 효과 연구

❖ 문제 : 숙면에 도움을 주기위해 개발한 약의 효과

❖ 표본 데이터 : 각각 10명의 실험 지원자로 구성된 2개의 그룹

- 실험 지원자가 약을 먹었을 때 증가한 수면 시간
- 데이터 : https://github.com/inetguru/IDS-CB35533/blob/main/sleep.csv
- Extra : 증가한 수면 시간(hour)

| | no | extra | group | ID |
|---|---|---|---|---|
| 1 | 1 | 0.7 | 1 | 1 |
| 2 | 2 | -1.6 | 1 | 2 |
| 3 | 3 | -0.2 | 1 | 3 |
| 4 | 4 | -1.2 | 1 | 4 |
| 5 | 5 | -0.1 | 1 | 5 |
| 6 | 6 | 3.4 | 1 | 6 |
| 7 | 7 | 3.7 | 1 | 7 |
| 8 | 8 | 0.8 | 1 | 8 |
| 9 | 9 | 0 | 1 | 9 |
| 10 | 10 | 2 | 1 | 10 |

❖ Pandas 라이브러리 이용 데이터 읽어 오기

```
import pandas as pd

sleeps = pd.read_csv('https://raw.githubusercontent.com/inetguru/IDS-CB35533/main/sleep.csv',index_col='no')
sleeps.head()
```

- 읽어 들인 sleeps는 DataFrame이라는 구조를 가짐
  - DataFrame은 Spreadsheet와 같은 Table화된 데이터를 저장, 처리하기 위한 Class
  - Record에 해당하는 항목 구분/식별을 위해 index 가 필요
    - Index column을 지정하지 않으면 index가 자동 생성
    - 이 예에서는 "no"라는 속성을 index로 사용
- 읽어 들인 DataFrame의 구조를 확인 → head()

| no | extra | group | ID |
|---|---|---|---|
| 1 | 0.7 | 1 | 1 |
| 2 | -1.6 | 1 | 2 |
| 3 | -0.2 | 1 | 3 |
| 4 | -1.2 | 1 | 4 |
| 5 | -0.1 | 1 | 5 |

# 탐색적 데이터 분석 (EDA) (1)

## ❖ 값의 확인

```
print(sleeps['extra'][sleeps['group']==1])
```

- 1 그룹 지원자의 수면 시간 증가 값만 출력
  - sleeps['extra'] : extra column만 선택
  - sleeps['group']==1 : group column의 값이 1인 항목만 선택
  - Tutorial : How do I select a subset of a DataFrame ?
- 수량형 데이터 (Numerical Data)

```
no
1  0.7
2 -1.6
3 -0.2
4 -1.2
5 -0.1
6  3.4
7  3.7
8  0.8
9  0.0
10 2.0
```

## ❖ 기본 통계량(Statistics)의 확인

```
print(sleeps['extra'][sleeps['group']==1].describe())
```

- (Filtered) DataFrame에 대해 count(), mean(), std(), max(), min() 등의 함수를 적용하면 각각 데이터 항목의 수, 평균, 표준편차, 최대값, 최소값 등을 구할 수 있다.
- describe()는 이러한 기본 통계량을 출력
  - R의 summary와 유사 (similar to R's summary)

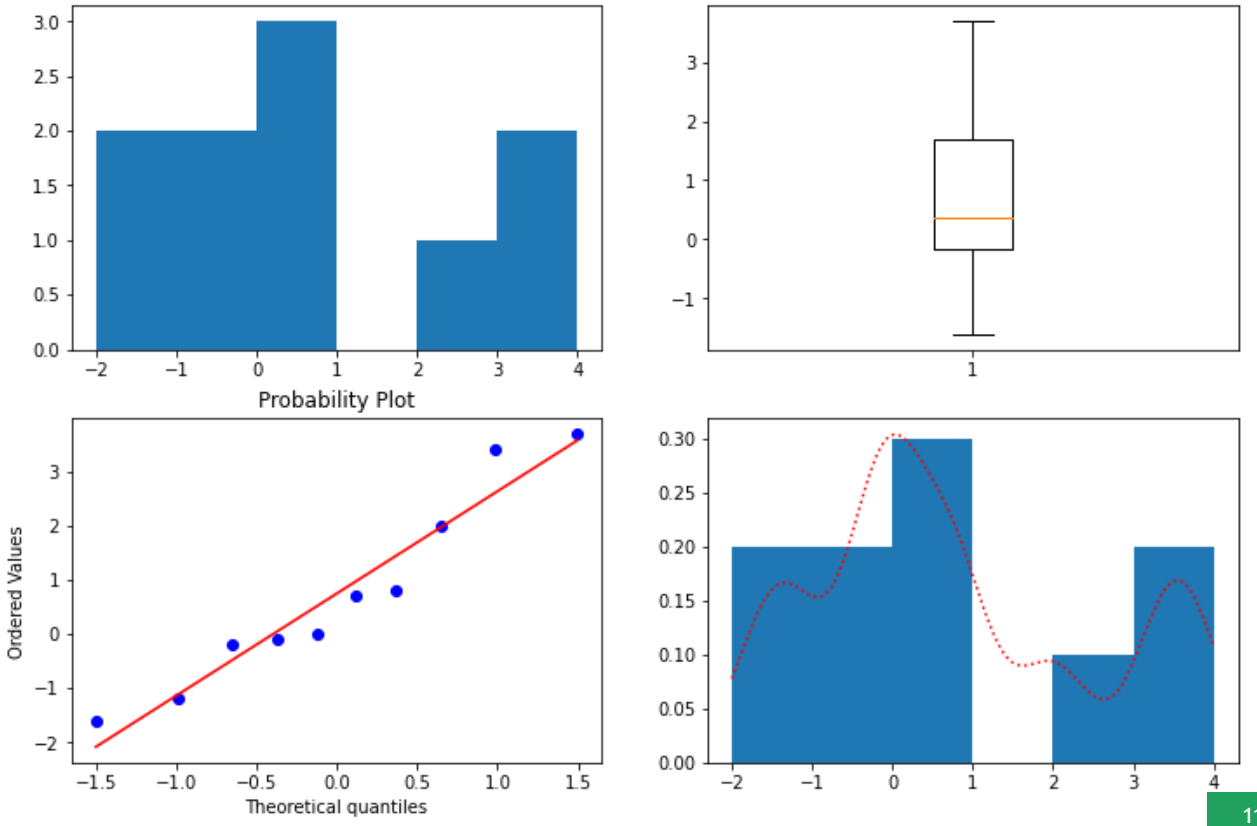
```
count 10.00000
mean  0.75000
std   1.78901
min  -1.60000
25%  -0.17500
50%   0.35000
75%   1.70000
max   3.70000
```

# 탐색적 데이터 분석 (EDA) (2)

```
import matplotlib.pyplot as plt
from scipy.stats import probplot
Group1 = sleeps['extra'][sleeps['group']==1]

plt.rcParams['figure.figsize'] = [12,8]
fig = plt.figure()
ax1 = fig.add_subplot(2,2,1)
ax1.hist(Group1,bins=[-2,-1,0,1,2,3,4])

ax2 = fig.add_subplot(2,2,2)
ax2.boxplot(Group1)

ax3 = fig.add_subplot(2,2,3)
probplot(Group1, plot=ax3)

ax4 = fig.add_subplot(2,2,4)
from scipy.stats import gaussian_kde
import numpy as np
density = gaussian_kde(Group1)
density.covariance_factor = lambda : .25
density._compute_covariance()
xs = np.linspace(-2,4,200)
ax4.plot(xs,density(xs),'r:')
ax4.hist(Group1,bins=[-2,-1,0,1,2,3,4],
density=True)
```

❖ 기초 시각화 (Basic visualization)

- Histogram
- Boxplot
- Q-Q Plot
- 커널밀도추정치 (Kernel Density Estimate)



# 통계 작업의 목표/결론 ?

❖ EDA의 주요 결과 Main findings of EDA

- 10명 중 4명은 수면 시간 감소, 6명은 증가
- 평균 수면 시간 증가는 0.75시간
- 표본표준편차(sample standard deviation)는 1.8 시간
- 데이터 분포에서 약간의 Bimodality (봉우리가 두 개 보이는 것)가 보이는 것 외에 그리 특이할 만한 사항 없음

➔ 수면제의 효과에 대한 의미 있는 설명(Narrative)으로 이어지지 않음

❖ 이 수면제는 효과가 있는가?

- 가설 검정 (Hypothesis Test)
  - 검정(檢定) : 검사하여 정함
    - 규칙에 따라 자격이나 조건을 검사, 예) 교과서 검정
  - 검증(檢證) : 검사하여 증명함
  - 가설검정(假說檢定), ~~가설검증(假說檢證)~~

❖ 이 수면제의 효과는 얼마인가?

- 신뢰 구간 (Confidence Interval)

❖ 누군가 다른 사람이 이 수면제를 복용하면 어떤 효과가 있을 것인가?

- 예측 (Prediction)

❖ 예와 같은 수량형 데이터에 대한 통계학에서의 교과서적 방법 → one-sample t-test

# One sample t-test

❖ One sample t-test

- Scipy.stats에는 여러 statistical tests 함수 지원

```
import scipy.stats as stats

Group1 = sleeps['extra'][sleeps['group']==1]

result1 = stats.ttest_1samp(Group1,0)
print(result1)
result2 = stats.ttest_1samp(Group1,0, alternative='greater')
print(result2)
```

```
Ttest_1sampResult(
statistic=1.3257101407138212, pvalue=0.2175977800684489)
Ttest_1sampResult(
statistic=1.3257101407138212, pvalue=0.10879889003422438)
```

❖ ttest_1samp(a, popmean, *alternative='two-sided'*)

- a : 1 sample t-test 대상 데이터
- popmean : Population Mean in null hypothesis
- alternative : {'two-sided', 'less', 'greater'}
  - Alternative hypothesis
  - Colab의 scipy Library version : 1.4.1 : alternative 미지원

❖ 결과 값

- statistic : t-통계량
- pvalue : p-value

❖ test 함수 호출 자체는 단순 & 쉬움, 문제는 해석

- [참고] R의 경우 t.test()라는 함수 제공
  - R은 결과 값으로 t-통계량, pvalue 외에 confidence interval 도 함께 반환

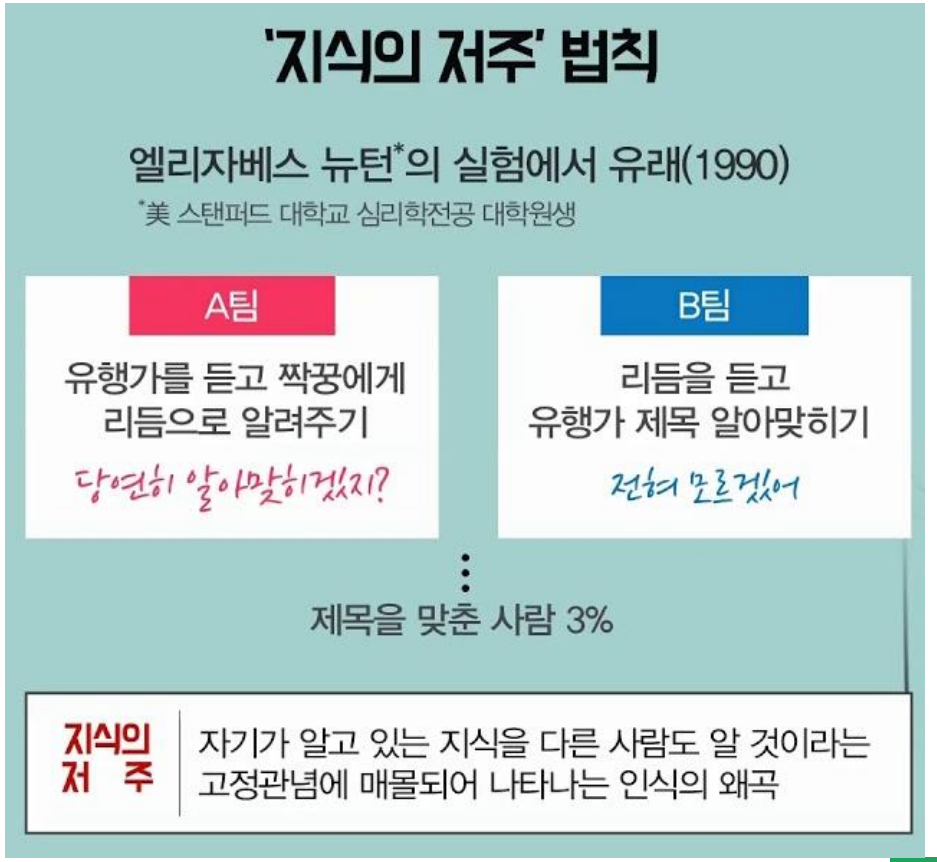
# P-value ?

❖ 예제 t-test의 결과 P-value의 의미

- Two-sided : 약이 수면 시간 변화에 효과가 없는데 이러한 표본 평균 수면 시간 변화가 관측될 확률은 21.75%이다.
- Greater : 약이 수면 시간 증가에 효과가 없는데 이러한 표본평균 수면 시간 증가가 관측될 확률은 10.87%이다.

❖ 데이터 과학 통계 분야 인터뷰의 주요 질문

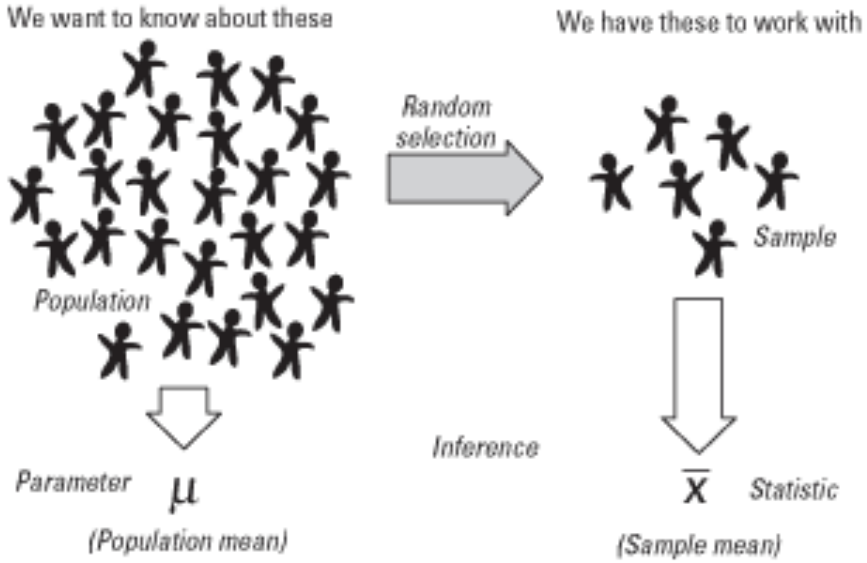
- P-Value의 정의
- 다양한 가설 검정 상황에서 비전문가들이 이해하기 쉽게 P-Value를 설명하라.
- 데이터 과학자는 P-Value를 정확히 이해해야 하고
- '지식의 저주(The Curse of Knowledge)'를 피하여 비전문가에게 전달할 수 있어야 함



# 첫째, 통계학은 숨겨진 진실을 추구한다

❖ 알려지지 않은 참값이 있음을 가정한다.

- (모집단/Population의) 알려지지 않은 참값/진실 (모수/Population parameter)이 있음을 가정한다.
- Ex) Population Mean : $\mu$ (또는 $\mu_X$)
- 통계적 추정 : 불완전한, 잡음이 섞여 있는 데이터(표본/Sample)로부터, 숨겨진 진실, 즉 모수 값을 찾는 작업




❖ 가설검정 : 수면제는 효과가 있는가?

- 모수(진실)에 대한 두 가지 가설
- $\mu = 0$ : 수면제는 효과가 없다 → 귀무가설(Null Hypothesis)
- $\mu > 0$ : 수면제는 수면시간을 늘리는 데 효과가 있다 → 대립 가설 (Alternative Hypothesis)
  - 보이고/입증하고 싶은 가설

❖ 가설 검정과 무죄추정의 원칙

- Innocent until proven guilty
- Null Hypothesis : Not Guilty,
  - 증거불충분 (Insufficient Evidence), Does not mean “Innocent”
- Alternative Hypothesis : Guilty
  - 대립되는 증거가 많으면 많을수록, 즉 귀무가설에 반하는 혹은 귀무가설 하에서는 관측되기 어려운 데이터 값이 많아지면 많아질수록 유죄 인 것을 선고하는 것이 쉬워진다

# 둘째, 통계학은 불확실성을 인정한다.

❖ 통계학적이지 않은 결론:

- 이 수면제는 수면시간을 늘리는 효과가 있다.
- 평균 수면시간의 증가는 0.75시간이다.

❖ 통계학의 "겸손한" 대답

- 이 수면제가 효과가 없는데 이렇게 큰 표본평균 수면 시간 증가 값이 관측될 확률은 11%다(P-Value).
- 평균 수면시간의 증가에 대한 95% 신뢰구간 (Confidence Interval)은 [- 0.53, 2.03]이다.

❖ 통계학은 어렵다.

- "사람은 오랜 진화의 과정으로 인해 빠르고 직관적인 시스템 1(Fast Thinking)이 발달했지만 통계학 은 시스템 2(Slow Thinking)로 우리의 평이한 본성과는 거리가 있다. 이것이 바로 통계학이 어려운 이유다."
  - Fast thinking : 자동적, 본능적, 감정적, 선입견이 많고 무의식적
  - Slow thinking : 자주 할 수 없고, 논리적, 계산적, 의식적
  - 확률적 생각은 어렵다. 사람의 생각은 100%를 지향한다. 전형적인 수학 문제와 같이 맞거나 틀리거나 100%인 것을 지향한다.
- "통계학은 어려운데 쉽다고 생 각하는 사람(의사)들이 많다 (Statistics—A subject which most statisticians find difficult but in which nearly all physicians are expert)" - 스티븐 센(Stephen Senn)
- 통계학의 답은 언뜻 보면 불필요할 정도로 복잡하다. 하지만 이해하면 할수록 가장 겸손하게(불확실성을 인정한다), 하지만 정확하게 (불확실성을 수량화 한다) 추론 하는 학문이다.

# 셋째, 통계학은 관측된 데이터가 가능한 여러 값 중 하나라고 생각한다

❖ 현재 관측한 데이터는 모집단에서 관측될 수 있는 여러 가능한 데이터 중 하나다

- 자명하다

❖ Simulation으로 아주 많은 Sample을 생성하자

- Simulation
  - 주어진 가정 하에 여러 현실을 만들어 보는 작업
- 모집단 모델링 – Gaussian/Normal
  - 수면제가 효과가 없다 ($\mu = 0$)
  - 표준편차(std)가 1.8시간 ($\sigma = 1.8$)
  - $N(\mu, \sigma^2) = N(0, 1.8^2)$
- 크기가 $n = 10$인 sample을 10,000개 생성

❖ Sample들의 statistic을 구하자

- t-statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

  - $\bar{x}$ : $sample\ mean$
    - $\mu_0$ : assumed population mean value for null hypothesis
  - $s$ : sample standard deviation
    - $\sigma$ : $population\ standard\ deviation$

❖ $Prob(T \geq t)$ ?

- 앞서 t-test에서 구한 결과
  - t-statistic = 1.3257101407138212
  - p-value = 0.10879889003422438 (alternative='greater')
- t-statistic 값이 1.3257101407138212 보다 큰 비율?
  - 참고 교재의 시뮬레이션 결과에서는 11.1%
- 이 비율은 t-test의 결과인 p-value 값에 부합
- p-value = $Prob(T \geq t)$

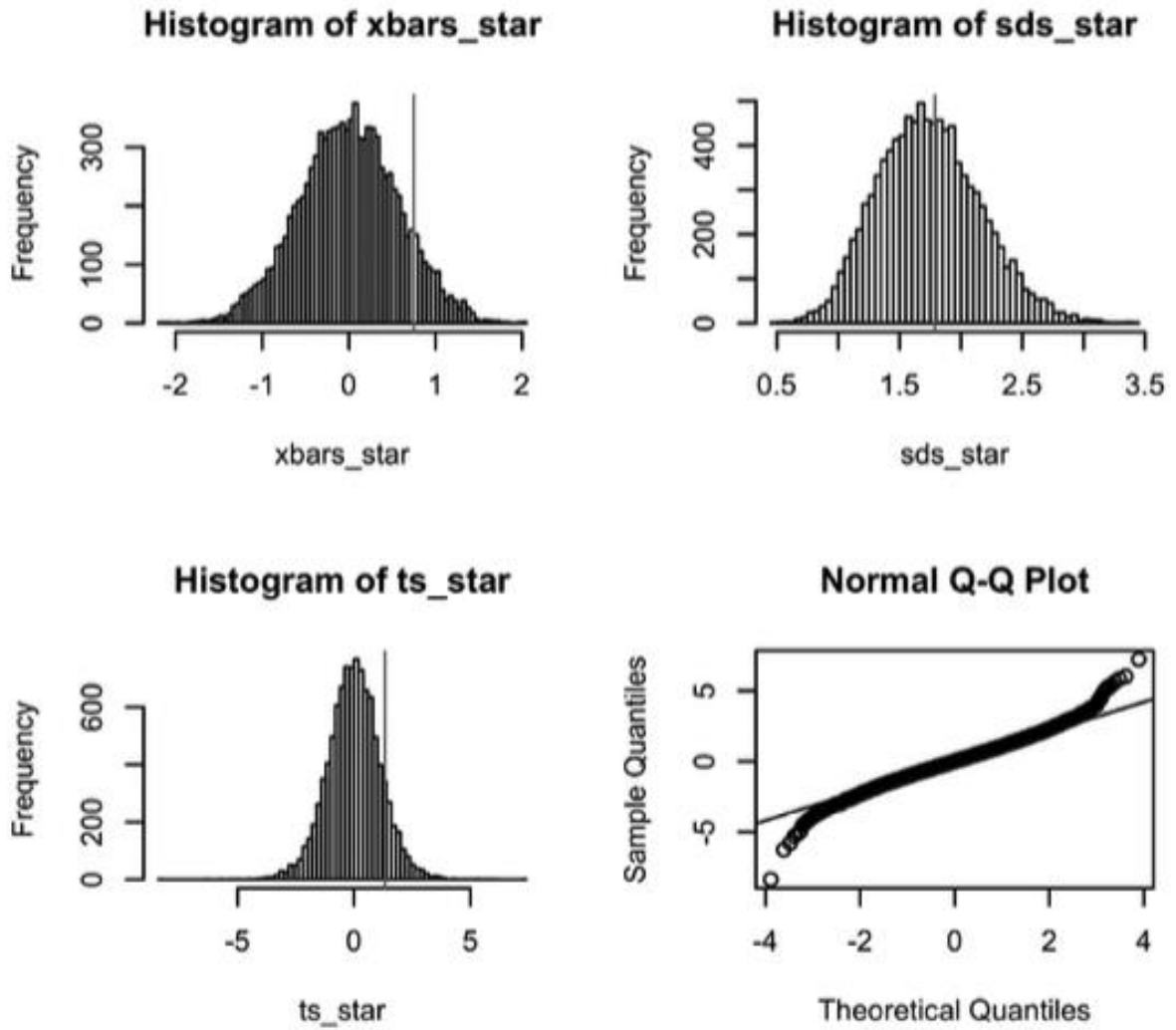



그림 6-3 실제로 수면제가 효과가 없다고 했을 때 크기가 10명일 때 표본의 추가 수면시간의 평균의 분포(왼쪽 위), 표본표준편차의 분포(오른쪽 위), t-통계량의 분포(왼쪽 아래), t-통계량의 정규분포 Q-Q 플롯(오른쪽 아래). B=10,000번의 시뮬레이션/평행우주를 생성하였다. 처음 세 플롯에서 빨간 선은 현재 관측된 표본의 표본평균, 표본표준편차, t-통계량 값을 나타낸다.

# 용어 (1)

## ❖ 모집단 (Population)

- A statistical population is a set of similar items or events (sometimes theoretical or imaginary) which is of interest for some question or experiment.

## ❖ 모수 (Population Parameter)

- Any measured quantity of a statistical population that summarises or describes an aspect of the population, such as a mean or a standard deviation

## ❖ 표본 (Sample)

- A set of individuals or objects collected or selected from a statistical population by a defined (random sampling) procedure

## ❖ 통계량 (Statistics)

- A statistic (singular) or sample statistic is any quantity computed from values in a sample that is used for a statistical purpose

## ❖ 귀무가설(Null Hypothesis)

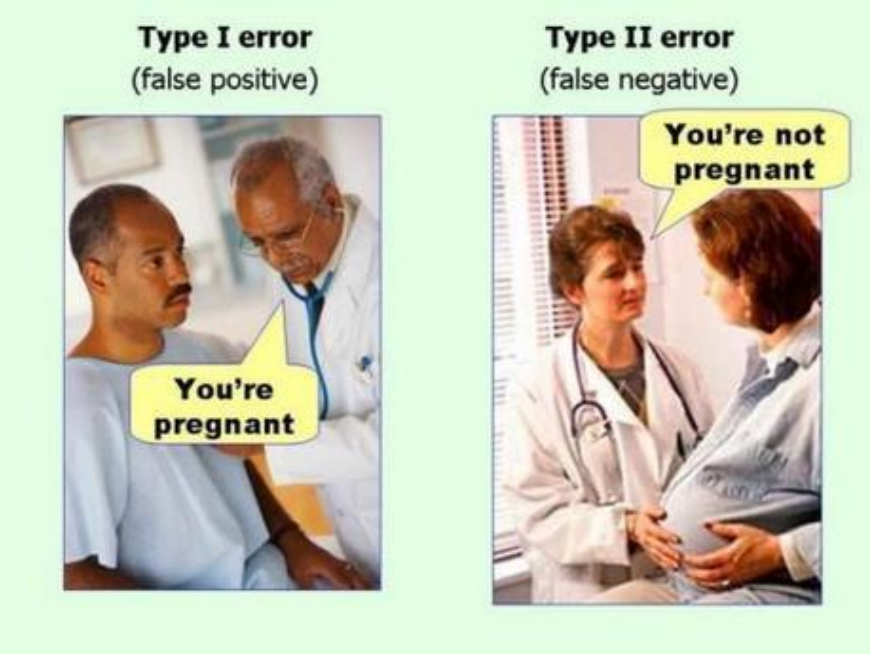
- A default (or status quo) hypothesis that a quantity (population parameter) to be measured is zero (null)
- Often denoted $H_0$
- Ex) $H_0 : \mu = 0$

## ❖ 대립가설(Alternative Hypothesis)

- Counterpoint to the null (what you hope to prove)
- Often denoted $H_1$
- Ex) two-sided alternative
  $H_1 : \mu \neq 0$ *: two-sided*
- Ex) one-sided alternative
  $H_1 : \mu > 0, : greater$, $H_1 : \mu < 0 : less$

# 용어 (2)

- Type-1 Error (False Positive) : the rejection of a true null hypothesis as the result of a test procedure
- Type-2 Error (False Negative) : the failure to reject a false null hypothesis as the result of a test procedure
- 유의 수준(Significance level) : the probability of rejecting the null hypothesis given that it is true.
  - Denoted by the Greek letter $\alpha$ and is also called the alpha level
- P-Value : the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed.




| Table of error types | | Null hypothesis ($H_0$) is | |
|---|---|---|---|
| | | True | False |
| Decision about null hypothesis ($H_0$) | Don't reject | Correct inference (true negative) (probability = $1-\alpha$) | **Type II error (false negative)** |
| | Reject | **Type I error (false positive)** (probability = $\alpha$) | Correct inference (true positive) |

# 예) 수면 시간 증가 문제

- 모집단 (Population) : 무수히 많은 수면 환자들
- 모수 (Parameter) : 무수히 많은 수면 환자들의 평균 수면 시간 증가
- 표본(Sample) : 10명의 랜덤하게 추출한 사람들
- 통계량(Statistic) : 표본의 평균 수면시간 증가
- 귀무가설(Null Hypothesis) : 주어진 수면제는 수면시간 증가에 효과가 없다
- 대립가설(Alternative Hypothesis) : 주어진 수면제는 수면시간 증가에 효과가 있다.
- Type 1 error : 실제로 수면제 효과가 없을 때 효과가 있다고 결론짓는 오류
- Type 2 error : 실제로 수면제 효과가 있을 때 효과가 없다고 결론짓는 오류
- P Value – 실제로 수면제의 효과가 없을 때  평균 수면시간 증가가 표본의 통계량이 보여준 것 만큼 클 확률

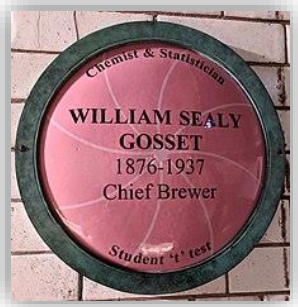
# Student's $t$-distribution

## ❖ $z$-statistic

- 데이터 : sequence of IID gaussian
- $X_i \sim_{iid} N(\mu, \sigma^2)$
- If $H_0$ : $\mu = \mu_0$ → true, then

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

- 문제점 : population std인 $\sigma$를 알기 어려움



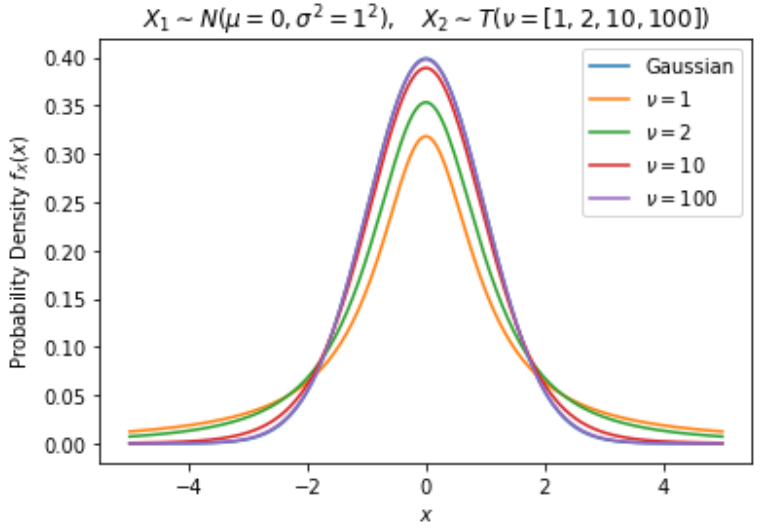

## ❖ $t$-statistic

- 1908년 기네스(Guinnes) 맥주회사에서 일하던 화학자 윌리엄 고셋(William Gosset)이 맥주의 품질을 모니터하기 위한 방법으로 개발

$$t = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} \sim t(\nu)$$

- $s$ : $sample\ std.$
- $\nu$ : $degree\ of\ freedom$ (df) = $n - 1$

- df가 커질 수록 정규 분포에 근사
  - 자유도가 작아질수록 양쪽 꼬리(tail)이 두터워짐



# $t$-test

❖ The $t$-test is any statistical hypothesis test in which the test statistic follows a Student's $t$-distribution under the null hypothesis.

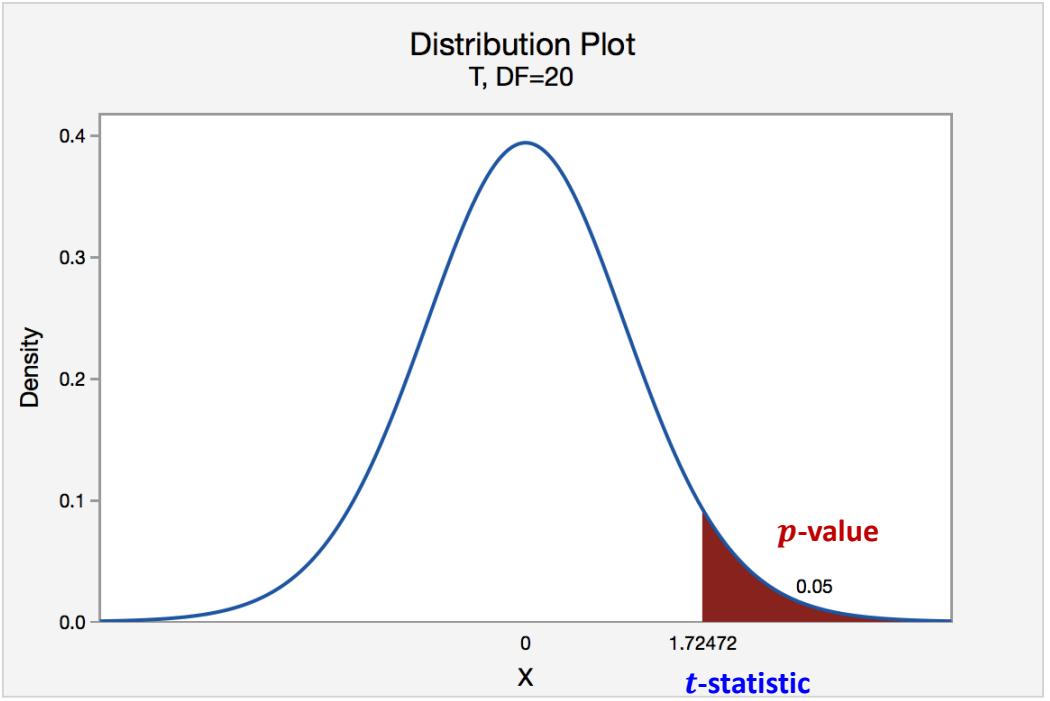
❖ Assumptions

- 조건 1 : 각 관측치가 독립이고 (Independent)
- 조건 2 : 동일한 분포를 따르며 (Identical)
- 조건 3 : 그 분포는 정규 분포, $(\sim N(\mu, \sigma^2))$

- 조건 1과 조건 2가 중요하며 조건 3은 실용적 관점에서 그다지 중요하지 않다고 알려져 있음.

❖ Table을 이용하여 계산

- 컴퓨터의 보급과 활용이 보편화되기 이전




| Degrees of freedom ( $\nu$ ) | Amount of area in one tail ( $\alpha$ ) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.0005 | 0.001 | 0.005 | 0.010 | 0.025 | 0.050 | 0.100 | 0.200 |
| 1 | 636.6192 | 318.3088 | 63.65674 | 31.82052 | 12.70620 | 6.313752 | 3.077684 | 1.376382 |
| 2 | 31.59905 | 22.32712 | 9.924843 | 6.964557 | 4.302653 | 2.919986 | 1.885618 | 1.060660 |
| 3 | 12.92398 | 10.21453 | 5.840909 | 4.540703 | 3.182446 | 2.353363 | 1.637744 | 0.978472 |
| 4 | 8.610302 | 7.173182 | 4.604095 | 3.746947 | 2.776445 | 2.131847 | 1.533206 | 0.940965 |
| 5 | 6.868827 | 5.893430 | 4.032143 | 3.364930 | 2.570582 | 2.015048 | 1.475884 | 0.919544 |
| 6 | 5.958816 | 5.207626 | 3.707428 | 3.142668 | 2.446912 | 1.943180 | 1.439756 | 0.905703 |
| 7 | 5.407883 | 4.785290 | 3.499483 | 2.997952 | 2.364624 | 1.894579 | 1.414924 | 0.896030 |
| 8 | 5.041305 | 4.500791 | 3.355387 | 2.896459 | 2.306004 | 1.859548 | 1.396815 | 0.888890 |
| 9 | 4.780913 | 4.296806 | 3.249836 | 2.821438 | 2.262157 | 1.833113 | 1.383029 | 0.883404 |
| 10 | 4.586894 | 4.143700 | 3.169273 | 2.763769 | 2.228139 | 1.812461 | 1.372184 | 0.879058 |
| 11 | 4.436979 | 4.024701 | 3.105807 | 2.718079 | 2.200985 | 1.795885 | 1.363430 | 0.875530 |
| 12 | 4.317791 | 3.929633 | 3.054540 | 2.680998 | 2.178813 | 1.782288 | 1.356217 | 0.872609 |
| 13 | 4.220832 | 3.851982 | 3.012276 | 2.650309 | 2.160369 | 1.770933 | 1.350171 | 0.870152 |
| 14 | 4.140454 | 3.787390 | 2.976843 | 2.624494 | 2.144787 | 1.761310 | 1.345030 | 0.868055 |
| 15 | 4.072765 | 3.732834 | 2.946713 | 2.602480 | 2.131450 | 1.753050 | 1.340606 | 0.866245 |
| 16 | 4.014996 | 3.686155 | 2.920782 | 2.583487 | 2.119905 | 1.745884 | 1.336757 | 0.864667 |
| 17 | 3.965126 | 3.645767 | 2.898231 | 2.566934 | 2.109816 | 1.739607 | 1.333379 | 0.863279 |

# $t$-test 계산 수행

## ❖ $t$-statistic 계산

```
import scipy.stats as stats
import math

sample = sleeps['extra'][sleeps['group']==1]
t_stat = (sample.mean()-0) / math.sqrt(sample.var()/sample.count())

print(t_statistic)
```

sample.mean()

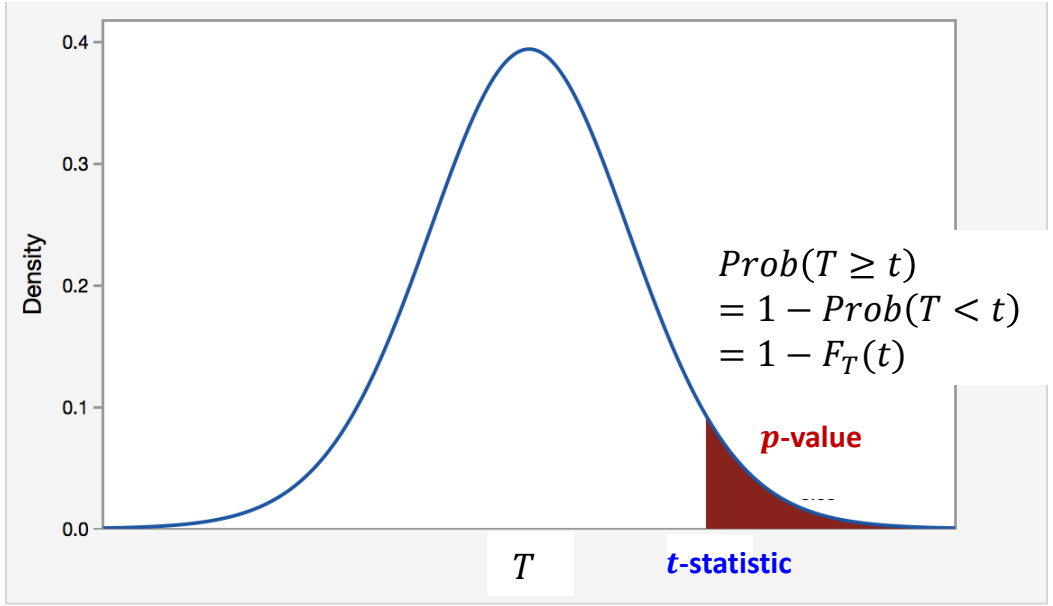
$$t = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

0

sample.var()

sample.count()

- t : 1.3257101407138212




```
df = sample.count()-1
rv_t = stats.t(df)
print(f'pvalue(> 0) = {1-rv_t.cdf(t_stat)}')
```

## ❖ p-value

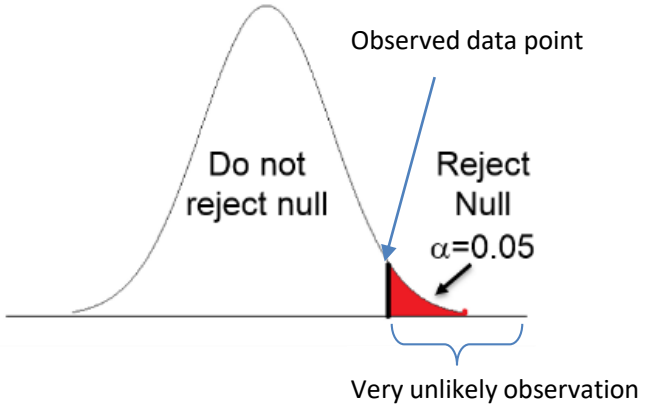
- Alternative : “greater”
  - 어떤 sample의 t-statistic이 관찰된 t 값보다 클 확률
- $t$ –distribution의 CDF 활용

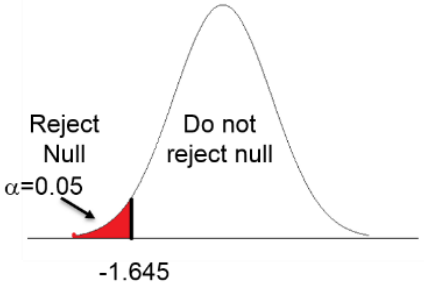
- pvalue(> 0) = 0.10879889003422438

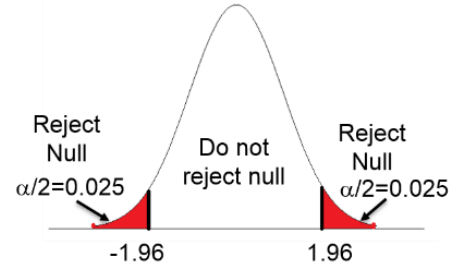
# $t$-test 해석

## ❖ 당신이 판사라면

- Null Hypothesis을 Reject하고 Alternative Hypothesis를 택하겠는가?
- Null Hypothesis - $H_0 : \mu = 0$
  - 주어진 수면제는 수면시간 증가에 효과가 없다
- Alternative Hypothesis - $H_1 : \mu > 0$
  - 주어진 수면제는 수면시간 증가에 효과가 있다 (수면 시간이 증가한다)
- 수집된 증거
  - 크기가 10인 표본에서 표본 평균이 0.75시간, 즉 45분의 평균 수면 시간 증가가 관찰되었음
  - Null Hypothesis가 참이라는 전제하에서
    - 우리가 크기가 10인 표본들을 매우 많이 구할 수 있다면 이러한 수준 이상의 평균 수면 시간 증가를 발생할 가능성(p-value로 표현)은 10.87%이다.
    - 우연히 이런 결과가 나올 가능성이 10.87%이다.
- 증거가 충분한가?




## ❖ 판단의 기준

- 유의 수준 (Significance Level) : $\alpha$
  - $\alpha$ : 10%, 5%, 1%, 0.5%, 0.1%....
- One-side t-test 기준

  If p-value < $\alpha$ : Reject Null,

  else : Do not reject Null
- 통계적 유의성이 있는 결과
  - a result has **statistical significance** when it is very unlikely to have occurred given the null hypothesis.

# P-value의 오해와 남용

❖ P-value보다 유의성만 보고하는 오류

- Ex) "그 수면제가 효과가 있다"는 것이 통계적으로 유의(statistically significant)하다.
  - Ex) $\alpha = 0.05\ (5\%), \text{pvalue} = 0.049, \text{pvalue} = 0.051$
  - Reject 여부 만을 보고하는 것보다 p-value 자체를 보고하는 것이 필요

❖ P-value만을 고려하고, 신뢰구간을 사용하지 않는 오류

❖ P-value를 모수에 대한 확률로 이해하는 오류

- Ex) 그 수면제가 효과가 있을 확률이 95% 이다.
- Ex) 그 수면제가 효과가 없다는 귀무가설이 맞을 확률이 95%이다.
- P-Value : the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed.

❖ 높은 P-value를 "귀무가설이 옳다"는 증거로 이해하는 오류

- 증거의 불충분을 의미할 뿐

❖ 낮은 P-value가 항상 의미 있다고 이해하는 오류

- 낮은 P 값을 얻었는데 평균 수면 시간 증가가 1초라면 ?
- 데이터의 크기가 커지면 P 값은 일반적으로 작아진다
  - 가설 검정의 민감도(Sensitivity)가 커진다

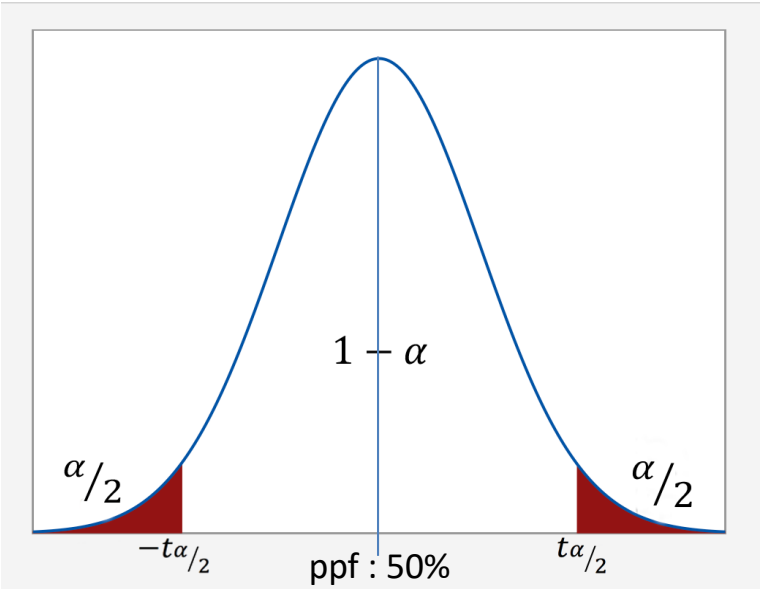
❖ 미국통계학회의 P- value에 대한 설명서

- The ASA's statement on p-values : context, process and purpose

# $t$-test confidence interval

❖ Formula

$$\left[\bar{X} - t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}, \quad \bar{X} + t_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}\right]$$




- $\alpha$ : Significance level - 5%, 1%, …
- $t_{\alpha/2}$ : ppf (percent point function) 활용하여 획득
  - Ex) $\alpha = 5\%$, rv.ppf $(1 - \alpha/2)$ → rv.ppf(0.975)

```
alpha = 0.05
print(f'Lower : {sample.mean() - rv_t.ppf(1-alpha/2) * math.sqrt(sample.var()/sample.count())}')
print(f'Upper : {sample.mean() + rv_t.ppf(1-alpha/2) * math.sqrt(sample.var()/sample.count())}')
```

- `Lower : -0.5297804134938646 Upper : 2.0297804134938646`

- `.interval()` 함수를 사용할 수도 있음 → $[-t_{\alpha/2}, t_{\alpha/2}]$

```
print(sample.mean() + np.array(rv_t.interval(0.95)) * math.sqrt(sample.var()/sample.count()))
```

# Confidence Interval

- **Confidence interval** (**CI**) is a type of estimate computed from the statistics of the observed data. This gives a range of values for an unknown parameter (for example, a population mean).

- Confidence Level
  - Ex) 90%, 95%, 99%, …
  - Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend to 90/95/99%.

- S1 == S2 ?
  - S1 : Sample들 중 95%가 그 신뢰구간 내에 true population parameter를 포함하고 있다
  - S2 : 이 Sample로 구한 신뢰구간 내에 true population parameter가 포함될 확률이 95%이다

# Sampling Distribution and Central Limit Theorem

❖ Parameter Estimation - Confidence Interval

- 모표준편차 $\sigma$ 값이 알려진 경우
- $\left[\bar{X} - z^*_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}},\ \bar{X} + z^*_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}\right]$
- $\sigma$ 값을 모를 경우 표본표준편차 $s$ 사용
- $\left[\bar{X} - t^*_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}},\ \bar{X} + t^*_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}}\right]$

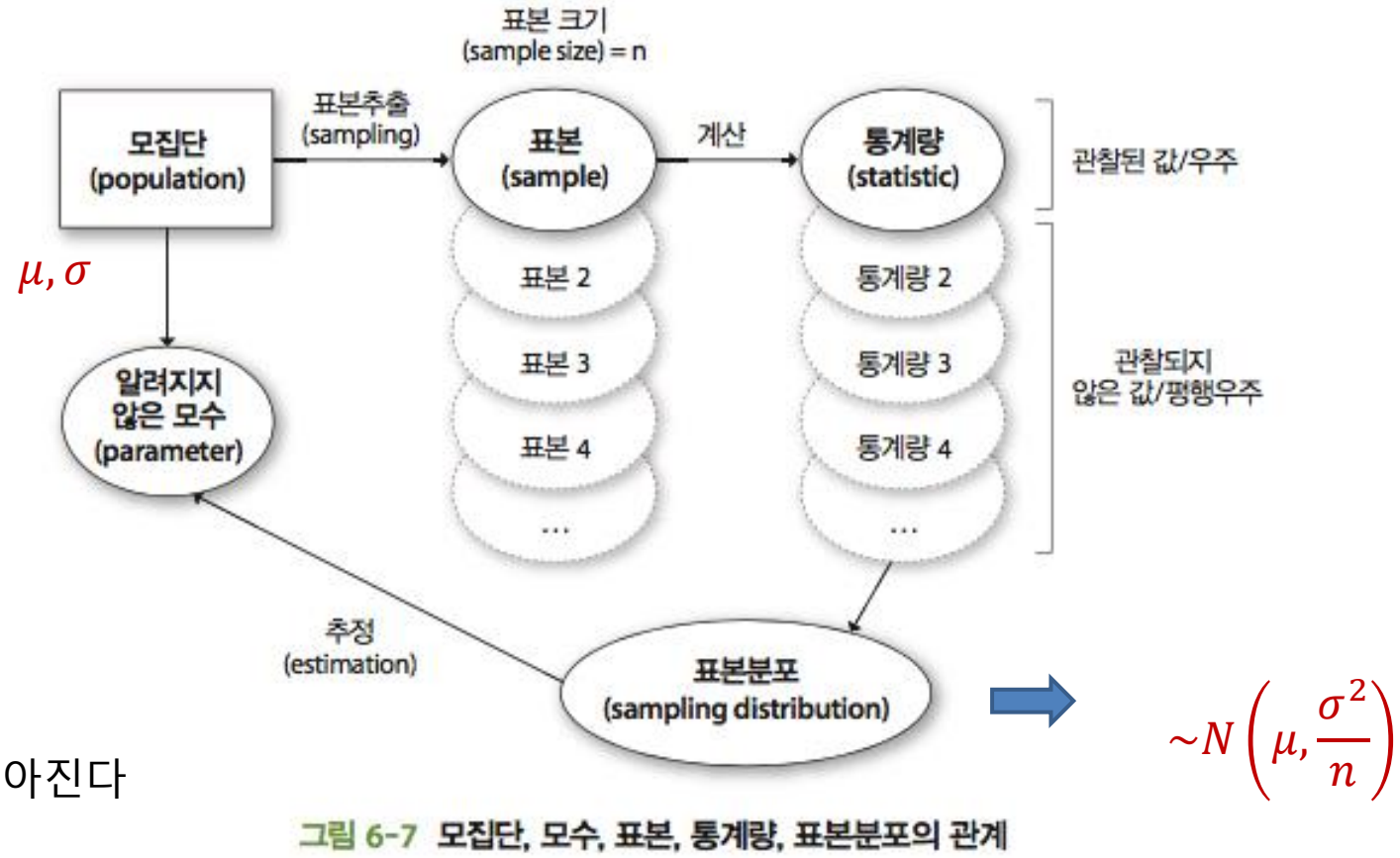



그림 6-7 모집단, 모수, 표본, 통계량, 표본분포의 관계

❖ 정확도는 $\sqrt{n}$ 에 비례한다

- 정확도가 높아진다 ~ 신뢰구간의 폭이 좁아진다

# 데이터 종류에 따른 분석 기법

❖ 80:20 법칙

- 80%의 실제 문제는 20% 정도의 통계 기법으로 처리할 수 있다

| 데이터 형 | 분석기법 | 모형 |
|---|---|---|
| 모든 데이터 | 데이터 내용, 구조 파악, 요약 통계량, 단순 시각화 | |
| 수량형 데이터 | 분포 시각화 (histogram, boxplot), 요약 통계량 T-test | $X_i \sim \ _{iid}N(\mu, \sigma^2)$ |
| 범주형 데이터 (성공 - 실패) | 도수 분포, Bar Graph, Binom-Test | $X \sim Binom(n, p)$ |
| 수량형 x, 수량형 y | Scatter Plot, Correlation, 단순회귀, 로버스트 회귀, 비모수회귀 | $(Y\|X = x) \sim \ _{iid}N(\mu(x), \sigma^2)$ |
| 범주형 x, 수량형 y | 병렬 Boxplot, ANOVA | $Y_{ij} \sim \ _{iid}N(\mu, \sigma^2)$ |
| 수량형 x, 범주형(성공-실패) y | Scatter Plot, 병렬 Boxplot, 로지스틱 회귀분석 | $Y = 1\|X = x \sim Binom\big(1, p(x)\big)$ |