데이터 과학을 위한 통계

Chapter2. 데이터와 표본분포

여는 말

빅데이터 시대가 되면서 더는 표본추출(표집, 샘플링)이 필요 없을 거라고 오해하는 사람들이 많다.

데이터의 질과 적합성을 일정 수준 이상으로 담보할 수 없으면서 데이터 크기만 늘어나는 것이 오늘날 상황이다.

오히려 다양한 데이터를 효과적으로 다루고 데이터 편향을 최소화하기 위한 방법으로 표본추출의 필요성이 더 커지고 있다.

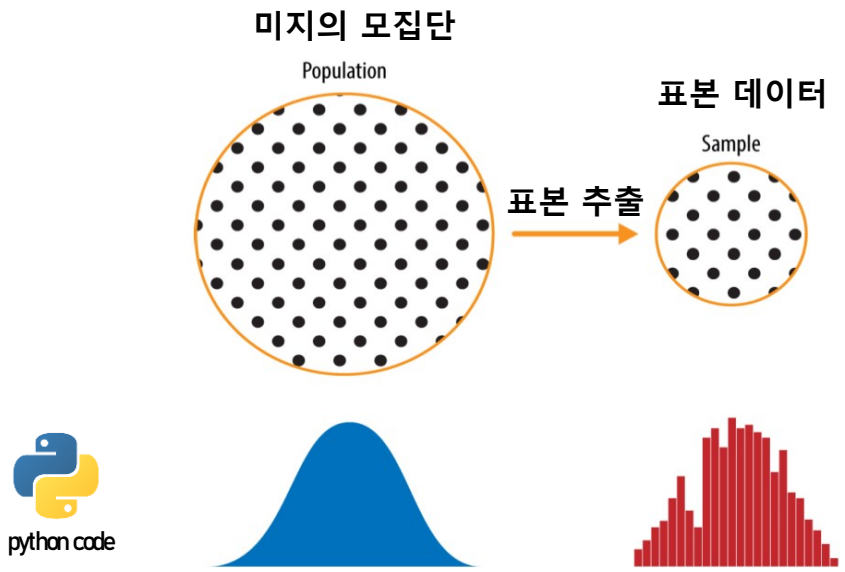


Figure 2-1. Population versus sample

전통적인 통계학

- 강력한 가정(ex.정규분포)에 기초한 이론을 통해 왼쪽의 모집단을 밝혀내는 데 초점

현대 통계학

- 가정이 필요하지 않은 오른쪽의 표본에 대한 연구로 방향 전환

Data Scientist

- 일반적으로 모집단보다는 표본추출 과정과 데이터 집중할 필요가 있음

- 하지만 때로는 모델링이 가능한 물리적 과정을 통해 데이터가 생성

ex) 이항분포(동전뒤집기), 포아송분포 등

2.1 임의표본추출과 표본 편향

샘플 기반의 추정이나 모델링에서 데이터 품질은 데이터 양보다 더욱 중요하다.

데이터 품질

– **완결성, 형식의 일관성, 깨끗함, 정확성** @데이터 과학 + **대표성** @통계

표본편향 예, 1936년 미국 대통령 당선 조사

1) 리터러리 다이제스트 : 1000만명 조사
2) 조지 갤럽 2000명

기관	조사방식	표본크기	예측	표본추출
리터러리 다이제스트	경제적지위 구독자 설문조사	10M	랜던 승리	비임의 non-random 표본편향
조지 갤럽	모집단 격주 여론조사	2000	루스벨트 승리	임의 random

모집단	population	어떤 데이터 집합을 구성하는 전체 대상 혹은 전체 집합
표본	sample	더 큰 데이터 집합으로부터 얻은 부분집합
N(n)	-	모집단(표본)의 크기
임의표본추출 (임의표집, 랜덤표본추출)	random sampling	무작위로 표본을 추출하는 것, 복원추출 vs 비복원추출
층화표본추출 (층화표집)	stratified sampling	모집단을 층으로 나눈 뒤, 각 층에서 무작위로 표본을 추출하는 것
계층	stratum	공통된 특징을 가진 모집단의 동종 하위 그룹(복수형 strata)
단순임의표본 (단순랜덤표본)	simple random sample	모집단 층화 없이 임의표본추출로 얻은 표본
편향	bias	계통상의 오류
표본편향	sample bias	모집단을 잘못 대표하는 표본

2.1.1 편향, bias

통계적 편향은 측정 과정 혹은 표본추출 과정에서 발생하는 계통적인 ,systematic, 오차를 의미한다.

정확한 조준 사격

오차 random
경향이 없음

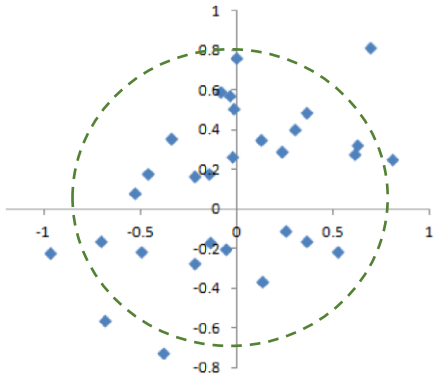


Figure 2-2. Scatterplot of shots from a gun with true aim

편향된 조준 사격

오차
경향이 존재

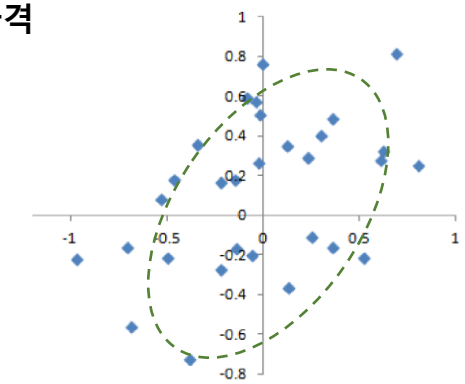


Figure 2-2. Scatterplot of shots from a gun with true aim

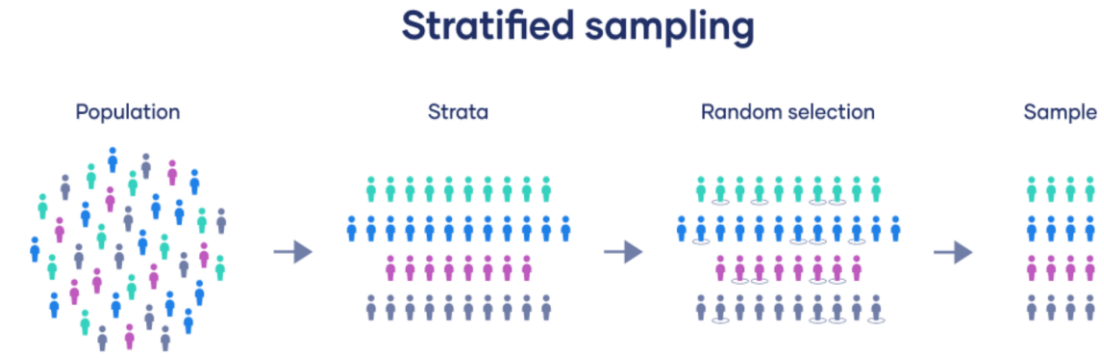
2.1.2 임의 선택, random selection

조지 갤럽 – 표본편향을 피하고 미국 유권자를 **대표하는 표본**을 얻기 위해 과학적으로 조사자를 선정 ➔ 핵심은 **임의표본추출**

임의표본추출 예, 고객의 대표 프로필 생성을 위한 설문 조사

1) 고객이 누구인지 정의
   - 과거/환불 고객 포함?, 사업자는? 등
2) 표본추출 절차
   - 무작위 100명
   - 시기를 고려한 실시간 거래 고객이나 웹 방문자 등

층화표본추출 – 모집단을 여러 층으로 나누고 각 층에서 무작위로 샘플 추출



참조: https://www.scribbr.com/methodology/stratified-sampling/

2.1.3 크기와 품질: 크기는 언제 중요해질까?

의외로 데이터 개수가 적을수록 더 유리한 경우가 있다. 임의표본추출에 시간과 노력을 기울일수록 편향이 줄 뿐만 아니라 데이터 탐색 및 데이터 품질에 더 집중할 수 있다.

그렇다면 대량의 데이터는 언제 필요할까?

빅데이터가 필요한 예시

- **구글 검색 정확성 향상** : 시간 정보를 포함한 방대한 양의 데이터 누적, 다양한 검색 쿼리 필요
- **실제 연관된 ,pertinent, 레코드** : 클릭한 사용자 정보를 포함한 검색쿼리 레코드, 막대한 데이터 포인트 필요

2.1.4 표본평균과 모평균

$\bar{x}$ **: 표본평균** - **관찰을 통해 획득**

μ : **모집단의 평균** - **표본들로 부터 추론**

※ 주요 개념

- 빅데이터 시대에도 **임의표본추출**은 중요하다.
- **편향**은 측정이나 **관측**에 **계통적 오차**가 있어 전체 모집단을 대표하지 못할 경우 발생한다.
- **데이터 품질이 데이터 양보다 중요할 때가 자주 있다. 임의표본추출**은 편향을 줄이고, 품질향상을 용이하게 한다.

2.2 선택편향

선택편향은 데이터를 선택적으로 고르는 관행을 의미

결국 오해의 소지가 있거나 단편적인 결론을 얻는다.

보통 가지고 있는 데이터를 먼저 확인한 후 그 안에서 패턴을 찾고자 한다.

실험을 통해 가설검정으로 확인한 현상 vs 사용 가능한 데이터를 통해 발견한 현상

방대한 검색효과, 비임의 표본추출, 데이터 체리 피킹(선별), 특정한 통계적 효과를 강조하는 시간 구간 선택 등

선택편향	**selection bias**	**관측데이터를 선택하는 방식 때문에 생기는 편향**
데이터 스누핑	**data snooping**	**뭔가 흥미로운 것을 찾아 광범위하게 데이터를 살피는 것 의미 있는 것이 나올 때까지 데이터를 너무 샅샅이 뒤지는 것**
방대한 검색효과	**vast search effect**	**중복 데이터 모델링이나 너무 많은 예측변수를 고려하는 모델링에서 비롯되는 편향 혹은 비재현성 큰 데이터 집합을 가지고 반복적으로 다른 모델을 만들고 질문하다 보면 흥미로운 것을 발견한다 그 결과는 의미 있는 것인가? 아니면 우연히 얻은 예외 경우인가?**
목표값 섞기	**target shuffling**	**성능 검증을 위해 둘 이상의 홀드아웃 세트를 이용하면 방대한 검색효과 방지 엘더는 데이터 마이닝 모델에서 제시하는 예측들을 검증하기 위해 목표값 섞기(순열검정) 추천 참조 https://www.elderresearch.com/resource/innovations/target-shuffling-process/**

2.2.1 평균으로의 회귀, regression to the mean

주어진 어떤 변수를 연속적으로 측정했을 때, 예외적인 경우가 관찰되면 그 다음에는 중간 정도의 경우가 관찰되는 경향

→ 예외 경우에 너무 많은 의미를 부여하면 선택편향으로 이어질 수 있다.

예를 들어 키가 엄청나게 큰 남성의 자식들도 아버지처럼 키가 큰 것은 아니었다.

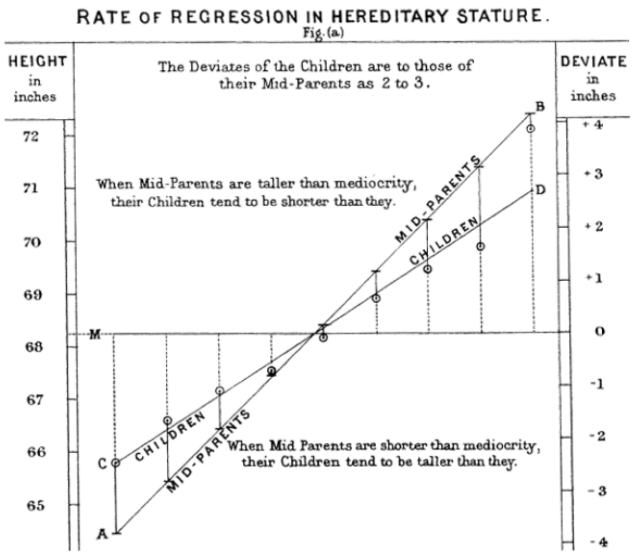


Figure 2-5. Galton's study that identified the phenomenon of regression to the mean

※ 주요 개념

- **가설**을 구체적으로 명시하고 **임의표본추출 원칙**에 따라 데이터를 수집하면 편향을 피할 수 있다.
- 모든 형태의 데이터 분석은 **데이터 수집/분석 프로세스**에서 생기는 **편향의 위험성**을 늘 갖고 있다.
- 데이터마이닝 모델 반복 실행, 데이터 스누핑, 흥미로운 사건의 사후 선택 등

2.3 통계학에서의 표본분포, Sampling Distribution of a Statistic

통계의 **표본분포**라는 용어는 하나의 동일한 모집단에서 얻은 여러 샘플에 대한 **표본통계량의 분포**를 나타낸다.

표본통계량	sample statistic	더 큰 모집단에서 추출된 표본 데이터로 얻은 측정 지표, 예) 평균, 분산, 표준편차 등
데이터분포	data distribution	어떤 데이터 집합에서의 각 개별 값의 도수분포
표본분포	sampling distribution	여러 표본들 혹은 재표본들로부터 얻은 표본통계량(평균 등)의 도수분포
중심극한정리	central limit theorem	표본크기가 커질수록 표본분포가 정규분포를 따르는 경향
표준오차	standard error	여러 표본들로부터 얻은 표본 통계량의 변량 (개별 데이터 값들의 변량을 뜻하는 표준편차와 혼동X)

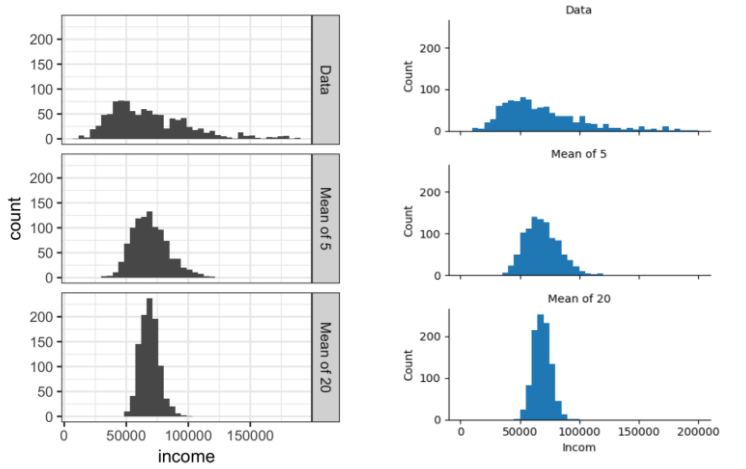


Figure 2-6. Histogram of annual incomes of 1,000 loan applicants (top), then 1,000 means of n=5 applicants (middle), and finally 1,000 means of n=20 applicants (bottom)

대출신청자 1000명의 연간소득 데이터분포

5개의 값의 평균 1000개 표본분포

20개의 값의 평균 1000개 표본분포

python code

■ 중심극한정리 및 표준오차

모집단이 정규분포가 아니더라도, 표본크기가 충분하고 데이터가 정규성을 크게 이탈하지 않는 경우, **여러 표본에서 추출한 평균은 종모양의 정규곡선을 따른다**.

중심극한정리 덕분에 신뢰구간이나 가설검정을 계산하는데 **t분포**와 같은 정규근사공식을 사용할 수 있다. 데이터과학에서는 **부트스트랩**을 사용할 수 있어서, 중심극한정리의 중요성이 떨어진다.

표준오차는 **표본분포**의 **변동성**을 말해주는 단일 측정 지표

표준오차 = SE = $\frac{s}{\sqrt{n}}$

s = 표본 값들의 표준편차
n = 표본 크기

통계량의 표준편차

표준오차 측정

1. 모집단에서 새로운 샘플 수집
2. 새 샘플에 대한 통계량(평균 등) 계산
3. 통계량의 표준편차 계산

현대 통계에서는 부트스트랩을 이용하여 표준오차 추정

※ 주요 개념

- **표본통계량**의 도수**분포**는 표본마다 다르게 나타날 수 있음
- **부트스트랩** 혹은 **중심극한정리**에 의해 **표본분포 추정** 가능
- **표준오차**는 **표본통계량의 변동성**을 요약하는 주요 지표

2.4 부트스트랩, bootstrap

부트스트랩 - 통계량이나 모델 파라미터(모수)의 **표본분포**를 **추정** 시 효과적인 방법은, 현재 있는 표본에서 추가적으로 **표본을 복원추출**하고 각 표본에 대한 통계량과 모델을 다시 계산하는 것이다. 데이터가 정규분포 아니어도 된다.

표본을 수천, 수백만번 **복원추출**하여 **가상 모집단 획득**

부트스트랩 표본	bootstrap sample	관측 데이터 집합으로부터 얻은 복원추출 표본
재표본추출 (재표집, 리샘플링)	resampling	관측 데이터로부터 반복해서 표본추출하는 과정 부트스트랩과 순열(셔플링)과정을 포함

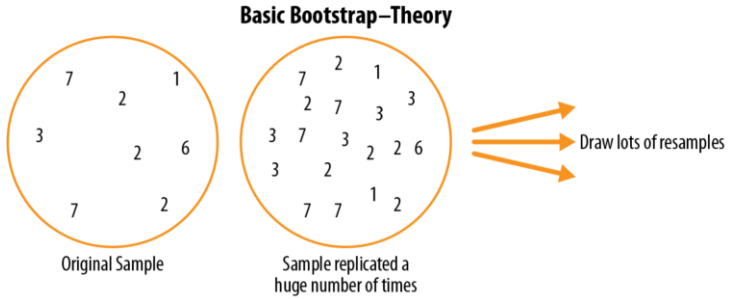


Figure 2-7. The idea of the bootstrap

부트스트랩 재표본추출 알고리즘

1. 1개 샘플링하여 기록하고 복원
2. n번 반복
3. 재표본추출된 값의 평균을 기록
4. 1~3단계를 R번 반복
5. R개의 결과 사용
   - 표본평균의 표준오차, 히스토그램, 신뢰구간 계산

배깅(bagging)

- Decision Tree에서 여러 부트스트랩 샘플을 가지고 트리를 여러 개 만든 다음 각 트리에서 나온 예측 값을 평균내는 것

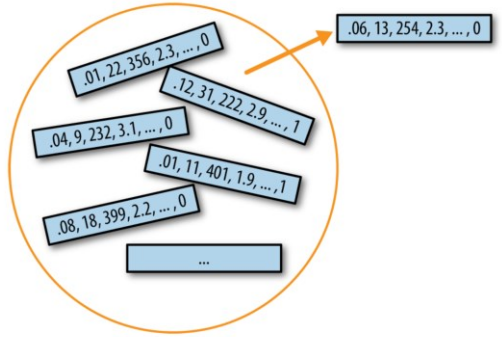
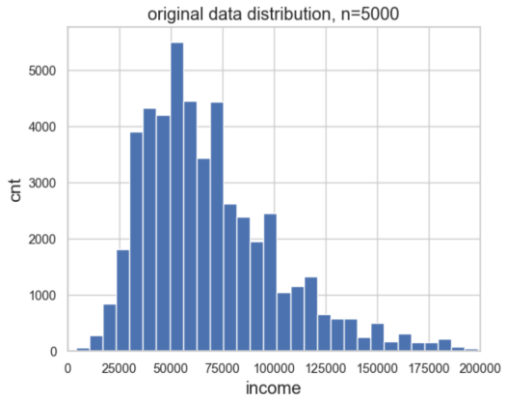
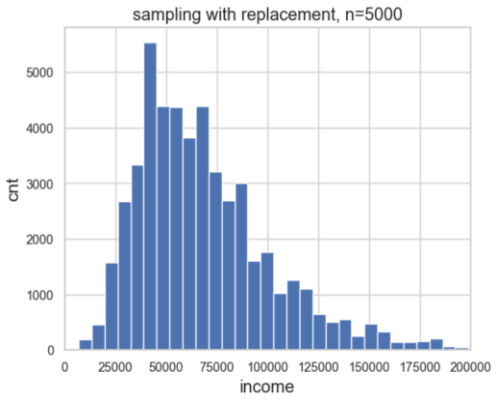


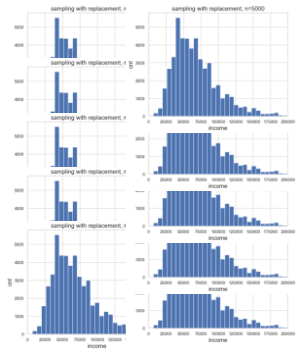
Figure 2-8. Multivariate bootstrap sampling



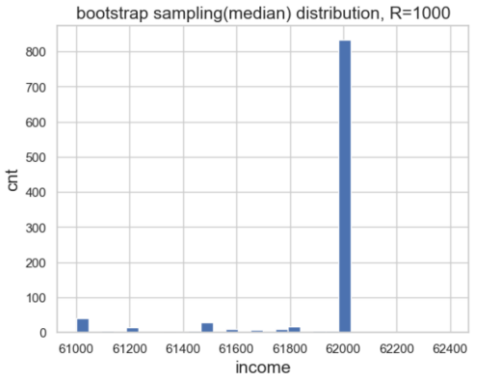
n번
복원추출



R번 반복



평균 표준오차
계산



2.4.1 재표본추출 대 부트스트래핑

python code

```
Bootstrap Statistics:
original mean: 62000.0
bootstrap mean: 61927.1135
bias: -72.88650000000052
stadard error: 214.16648115286253
```

재표본추출 – 여러 표본이 결합되어 비복원추출을 수행할 수 있는 순열(셔플) 과정을 포함한다.

부트스트랩 – 항상 관측된 데이터로부터 복원추출한다.

※ 주요 개념

- **부트스트랩(데이터로부터 복원추출)은 표본통계량의 변동성을 평가하는 강력한 도구이다.**
- **부트스트랩은 표본분포의 수학적 근사치에 대한 엄청난 연구 없이도 다양한 환경에서 유사한 방식으로 적용될 수 있다.**
- **수학적 근사가 어려운 통계량에 대해서도 샘플링 분포를 추정할 수 있다.**
- **예측 모델을 적용할 때, 여러 부트스트랩 표본들로부터 얻은 예측값을 모아서 결론을 만드는 것(배깅)이 단일 모델을 사용하는 것보다 좋다.**

2.5 신뢰구간, confidence intervals

90% 신뢰구간이란, 표본통계량의 부트스트랩 표본분포의 90%를 포함하는 구간을 말한다.

표본크기: n, 관심있는 표본통계량이 주어졌을 때

부트스트랩 신뢰구간 구하는 법

1. n개 복원추출 (재표본추출)
2. 원하는 통계량 기록
3. 1~3단계를 R번 반복
4. xx% 만큼 자르기
5. 절단한 점이 xx% 부트스트랩 신뢰구간의 양 끝점

신뢰수준	**confidence level**	**같은 모집단으로부터 같은 방식으로 얻은, 관심 통계량을 포함할 것으로 예상되는 신뢰구간의 백분율**
구간끝점	**interval endpoint**	**신뢰구간의 최상위, 최하위 끝점**

※ 주요 개념

- **신뢰구간은 구간 범위로 추정값을 표시하는 일반적인 방법이다.**
- **더 많은 데이터를 보유할수록 표본추정치의 변이가 줄어든다.**
- **허용할 수 있는 신뢰수준이 낮을수록 신뢰구간은 좁아진다.**
- **부트스트랩은 신뢰구간을 구성하는 효과적인 방법이다.**

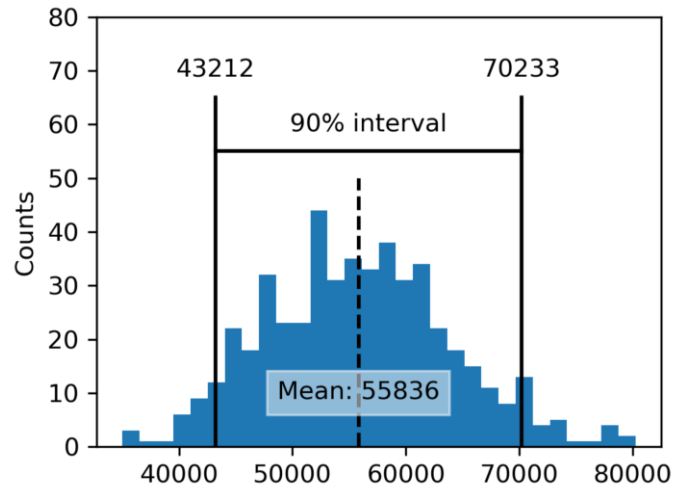


Figure 2-9. Bootstrap confidence interval for the annual income of loan applicants, based on a sample of 20

2.6 정규분포

정규분포(Normal Distribution)

- 통계분석에서 가장 중요한 확률분포 : 실무적으로 목표값/참값이 명확하고 우연변동만 존재하는 경우 대부분 정규분포로 가정
- 구간/비율 척도로 측정되는 프로세스의 관심 특성에 대한 특징을 파악하고자 할 때 사용
- 특징 : 평균을 기준으로 좌우 대칭이고, 종(Bell) 모양을 가짐
- **분포의 모수 : 평균(μ), 분산(σ^2)**

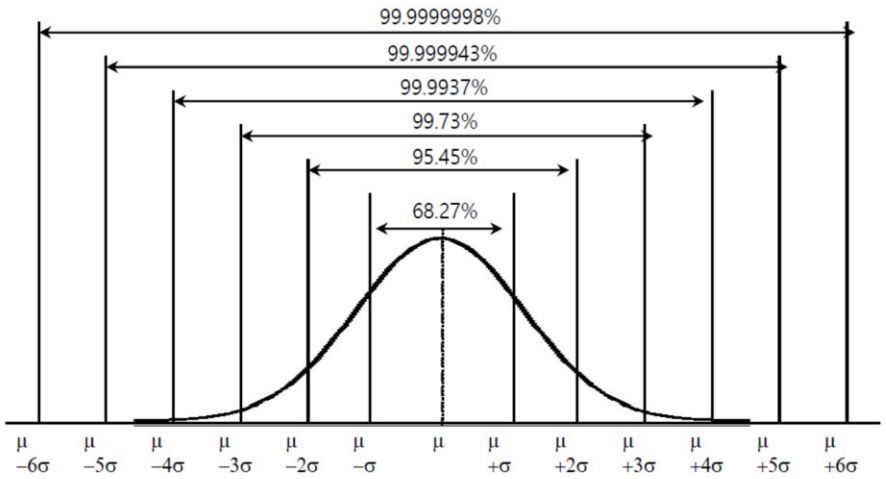
$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ : 모집단의 평균치

σ : 모집단의 표준편차

π : 3.14159

e : 지수 = 2.71828

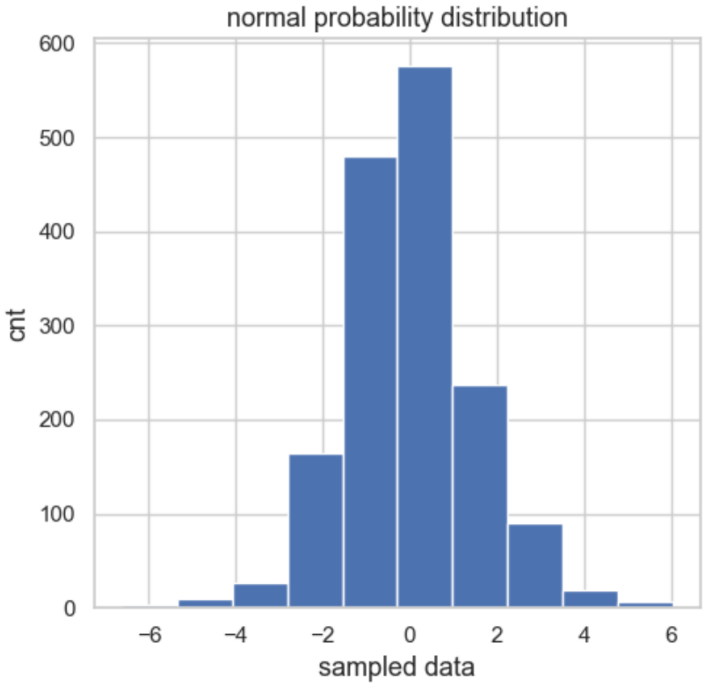


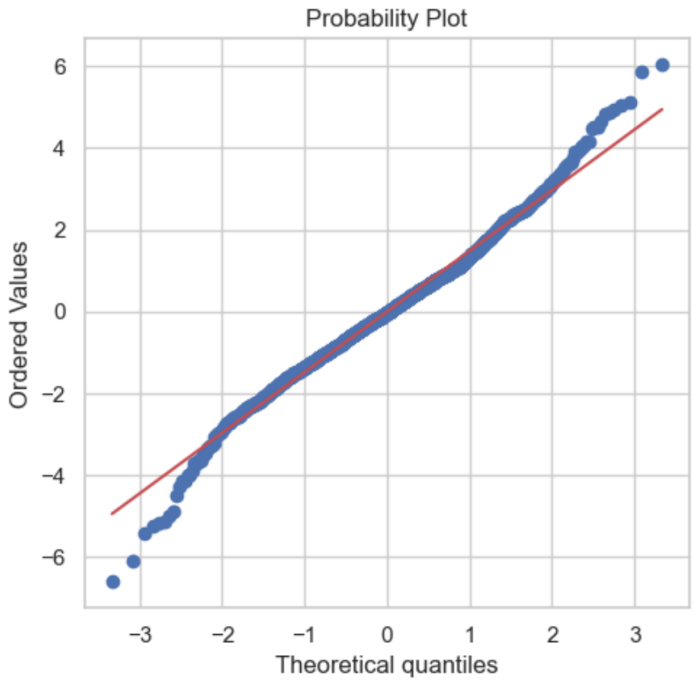
(Carl Friedrich Gauss – 독일)

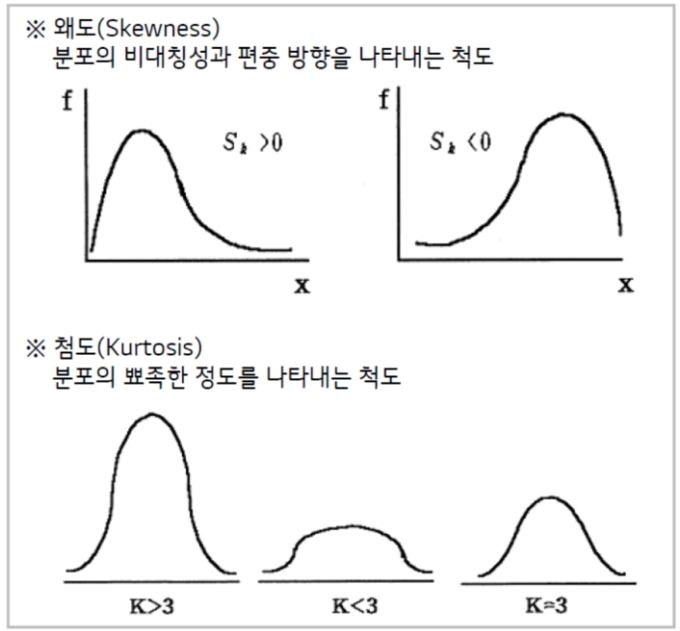
오차	error	데이터 포인트와 예측값 혹은 평균 사이의 차이
표준화(정규화)	interval endpoint	평균을 빼고 표준편차로 나눈다
Z점수	z-score	개별 데이터 포인트를 정규화
표준정규분포	standard normal distribution	평균=0, 표준편차=1인 정규분포
QQ그림	QQ-plot	표본분포가 정규분포에 얼마나 가까운지 확인

2.7 긴꼬리분포

데이터는 일반적으로 정규분포를 따르지 않는다.







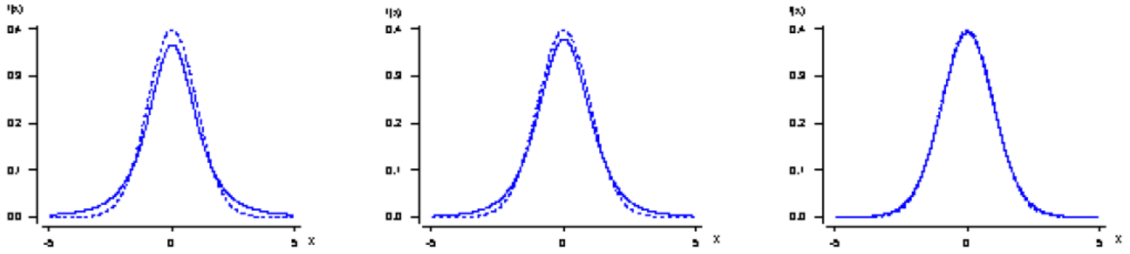
2.8 t-분포

t - 분포

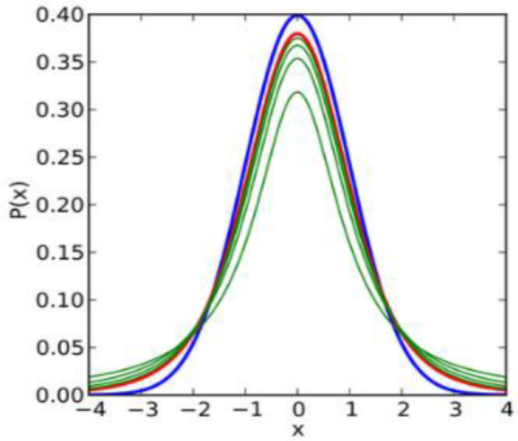
- 표본의 평균이 가지는 분포
- n ≤ 30 일 때 $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ ~ t(n-1)
- n > 30 일 때 $Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ ~ Z(0, 1) < -- 표본의 크기가 충분히 큰 경우

t 분포는 영국의 수리통계학자인 W. S. Gosset가 필명 Student로 발표한 소표본에 대한 확률분포이다. 그래서, Student의 t분포라고도 함

예시) df = 3, 5, 20에 따른 t-분포(실선)와 표준정규분포(점선)



일반 정규분포를 평균과 σ를 이용하여 표준화시킨 것을 표준정규분포라하고, 표본평균 분포를 총평균과 표준오차를 이용해 표준화시킨 분포를 t분포라 함



- t-분포는 자유도(df, degree of freedom, n-1)에 따라 분포의 모양이 결정됨
- 표준정규분포와 같이 '0'을 기준으로 좌우대칭이지만, 표준정규분포보다 더 긴 꼬리를 갖는다.
- t-분포의 자유도가 커질수록 표준정규분포에 근사한다.

◆ 모집단의 분산을 알지 못하고, 샘플의 크기(n)가 크지 않을 때 모평균의 검/추정에 사용 됨

2.9 이항분포

이항 분포(binomial distribution)

이항 분포는 무한모집단에서 추출한 n 개의 표본 중에서 불량품의 개수 x에 대한 분포이다.
불량률을 P라 하고, n개의 표본은 복원추출(with replacement) 하는 확률의 분포다.
각 표본을 관측 한 후 모집단으로 되돌려 넣고 다음 표본을 추출하는 방식을 취한다.

이항 분포의 확률 분포와 평균과 분산

- $f(x) = \binom{n}{x} p^x (1-P)^{n-x}$, x = 0, 1, 2,...., n
  여기서 n은 시행 횟수(=표본 크기), P 는 모불량률

- $X \sim B(n, P)$이면
  $E(X) = nP,\ Var(X) = nP(1-P)$

① P = 0.5 일때 분포의 형태는 기대값 nP에 대하여 좌우 대칭이 된다

② P = 0.1 이고 n ≥ 50일 때에는 포아송 분포에 근사한다.

③ nP ≥ 5 이고 n(1-P) ≥ 5 일 때에는 정규분포에 근사한다

※ **초기하 분포**는 이항분포의 전제조건인 매시행마다 발생확률이 일정하다는 가정을 만족하지 않는 경우에 사용하는 분포이며, 확률 분포 함수는

- $f(x1, x2) = \frac{{}_{N1}C_{X1} \cdot {}_{N2}C_{X1}}{{}_{(N1+N2)}C_{(x1+x2)}}$

※ 이항분포는 일어날 수 있는 경우의 수가 오직 2가지만 있을 때의 확률 분포를 말함.
예를 들어 동전을 던져 나올 수 있는 경우의 수는 앞면 또는 뒷면의 두 가지 뿐이며, 제품을 검사하는 경우에도 양품 또는 불량품의 두 가지만 있을 수 있음.

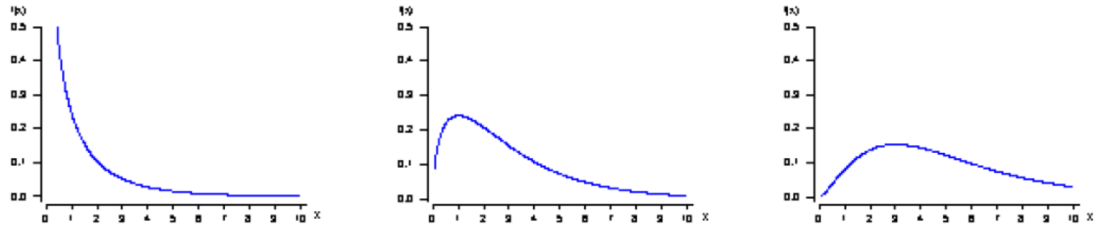
※ 주머니에 흰색 구슬 5개와 파란색 구슬 5개를 집어 넣고 임의로 1개를 꺼내 색깔을 본 후, 꺼낸 공을 다시 주머니에 넣고 1개를 새로 꺼내는 경우는 **"복원 추출"**, 1개를 꺼낸 후 다시 집어 넣지 않고 남은 9개에서 1개를 꺼내는 경우는 **"비복원 추출"**이라 함.
복원 추출, 즉 매시행마다 성공확률이 일정한 경우에는 이항 분포를 따르고, 비복원 추출, 즉 매시행마다 성공확률이 달라지는 경우에는 초기하 분포를 따른다.

2.10 카이분포, F분포

Chi-Square 분포

- 양의 정수 k개의 독립적이고 표준정규분포를 따르는 확률변수 $x_1, x_2, \ldots x_k$를 정의하면
- $Z^2 = \left(\frac{(x-\mu)}{\sigma}\right)^2$ 은 자유도 k의 Chi square 분포를 따른다
- $\frac{(n-1)S^2}{\sigma^2} = \chi^2(n-1)$ (분자의 S는 제곱합임)

예시) 자유도 = 1, 2, 3에 따른 카이제곱분포



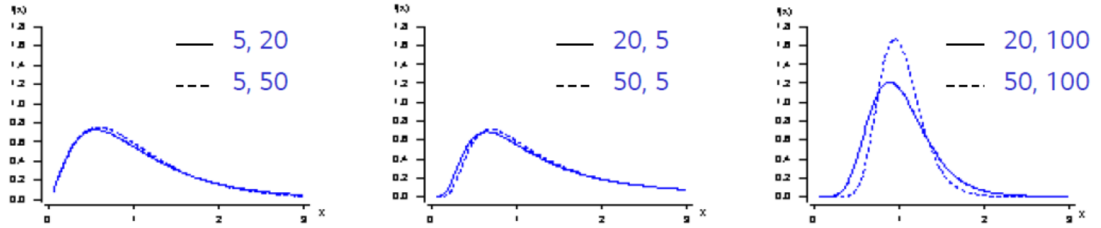
◆ 모분산의 검/추정 또는 이산형 변수의 독립, 동질성, 분포의 적합도 검정에 사용

χ^2분포는 영국의 수리통계학자인 Karl Pearson에 의해 고안되었다. 확률변수 u가 μ~ N(0, 1)인 표준정규분포를 따를 때 μ2은 자유도가 1인 χ^2 분포를 따른다. 즉, μ2 ~ $\chi^2(1)$이다. 만약 $\mu_1, \mu_2, \ldots, \mu_n$이 N(0,1)인 표준정규분포로부터 얻어진 확률표본이라면 $\chi^2 = \mu_1^2 + \mu_2^2 + \ldots + \mu_n^2$은 자유도가 n인 χ^2 분포를 하는 확률 변수가 된다.

F-분포

- 분산이 같은 2개의 정규모집단에서 추출된 불편분의 비(V_2/V_1)가 가지는 분포
- $F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$ *(k_1, k_2는 각 샘플의 자유도)*

예시) 두개의 자유도에 따른 F-분포



◆ 두 모분산 비(Ratio)의 검/추정에 사용

F분포는 R.A. Fisher에 의해 발견된 분포이다. 두 확률변수 V_1, V_2가 각각 자유도가 k_1, k_2이고 서로 독립인 카이제곱분포를 따른다고 할 때, 다음과 같이 정의되는 확률변수 F는 자유도가 (k_1, k_2)인 F-분포를 따른다고 한다.

2.12 포아송분포

- **일정비율로 발생하는 사건의 경우, 시간 단위 또는 공간 단위당 발생하는 사건의 수를 포아송 분포로 모델링할 수 있다.**

포아송 분포(Poisson distribution)

포아송 분포는 일반적으로 일정한 단위 공간(길이, 면적 등)이나 사건 내에 어떤 사건의 출현 횟수 또는 생산라인에서 특정 시간에 발생하는 결점의 수 x에 관심이 있는 경우에 사용된다.
이는 이항분포에서 nP=m 이라는 크지 않은 일정한 값으로 하고,
P→0 그리고 n→∞ 일 경우에 발생하는 극한 분포로 유도되는 분포임.

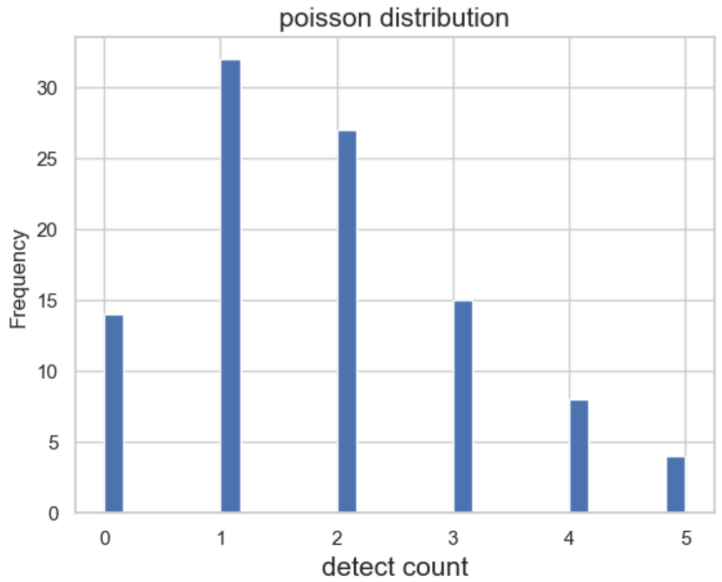
포아송 분포의 확률분포와 평균과 분산

- $f(x) = \frac{e^{-m} * m^x}{x!}$, $x = 0, 1, 2, \cdots, e \cong 2.7182 \quad m > 0$
- $X \sim P(m)$이면
  $E(X) = m$, 이고 $V(X) = m$

m = np(결점개수)

- 일정 시간에 드물게 발생하는 교통사고의 수,
- 계약 과정에서 가끔 발생하는 오타의 수,
- 단위시간에 많은 제품을 생산하는 자동 생산라인에서 간헐적으로 발생하는 불량품의 수 등

• 이항분포의 포아송 근사
: n이 크고, p가 매우 작은 경우에 포아송 분포는 이항분포의 근사 확률을 제공한다.
이때, np = m을 가정



- 지수분포 : 고장발생 시간, 개별 고객 상담 소요 시간 모델링
- 고장률 추정 : 데이터가 충분하지 않을 때, 시뮬레이션 또는 확률의 직접 계산을 통해 다른 가상 사건 발생률을 평가하고, 그 이하로 떨어지지 않은 임계값을 추산.
- 베이불 분포 :시간에 따라 변화하는 사건발생률은 베이불 분포로 모델링할 수 있다.