Chapter 02. 데이터와 표본분포 - 상지

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

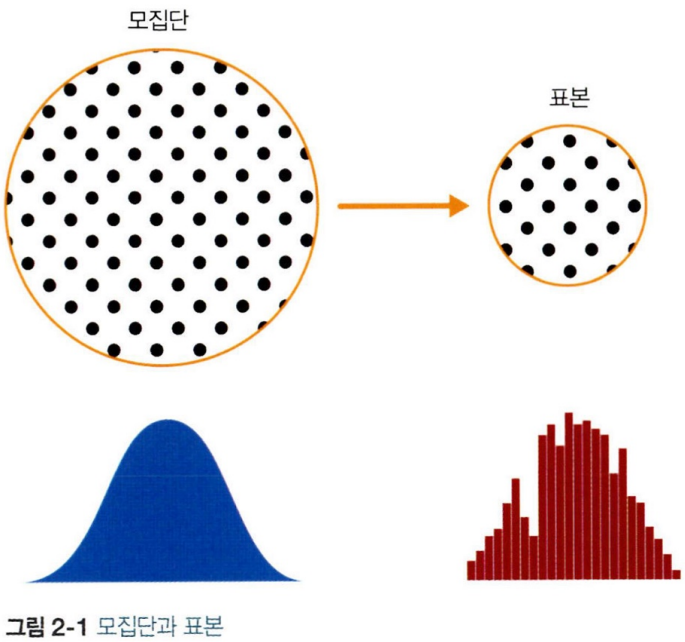

그림 2-1 모집단과 표본

- 모집단에서 **표본**을 얻어내는 것이 **표본추출**이다.
- 전통적인 통계학에서는 왼쪽의 모집단을 밝혀내는데 초점을 맞춰왔다면, 현대의 통계학에서는 오른쪽의 **표본에 대한 연구**로 방향이 옮겨지기 시작했다.

▼ 2.1 임의표본추출과 표본편향

> 통계학에서의 모집단은 생물학에서의 모집단과는 조금 차이가 있다. 생물학에서처럼 전체 집합의 크기가 크면서도 유한한 값으로 명

> 확히 정의될 때도 있지만 이론적인 가상의 집합을 의미하기도 한다.

✔️ **용어 정리**
표본 : 더 큰 데이터 집합으로부터 얻은 부분집합
모집단 : 어떤 데이터 집합을 구성하는 전체 대상 혹은 전체 집합
N(n) : 모집단(표본)의 크기
임의표본추출(임의표집, 랜덤표본추출) : 무작위로 표본을 추출하는 것
층화표본추출(층화표집) : 모집단을 층으로 나눈 뒤, 각 층에서 무작위로 표본을 추출하는 것
계층 : 공통된 특징을 가진 모집단의 동종 하위 그룹(복수형은 strata로 쓴다.)
단순임의표본(단순랜덤표본) : 모집단 층화 없이 임의표본추출로 얻은 표본
편향 : 계통상의 오류
표본편향 : 모집단을 잘못 대표하는 표본

임의표본추출

- 대상이 되는 모집단 내의 선택 가능한 원소들을 무작위로 추출하는 과정
- 각 추첨에서 모든 원소는 동일한 확률로 뽑힌다.
- 그 결과 얻은 샘플을 **단순임의표본**이라고 한다.

복원추출

- 다음번에도 중복추출이 가능하도록, 뽑은 샘플을 다시 모집단에 포함시키는 표본추출 방법

비복원추출

- 한번 뽑힌 원소는 추후 추첨에 사용하지 않는 표본추출 방법입니다.

통계에서의 데이터 품질

- 완결성
- 형식의 일관성
- 무결성(깨끗함)
- 정확성

- 대표성

비임의와 표본 편향

- 비임의란, 아무리 랜덤표본이라고 해도 어떤 표본도 모집단을 정확하게 대표할 수 없다는 것을 의미한다.
- 모집단과 표본 사이의 차이가 유의미할 만큼 크고, 첫 번째 표본과 동일한 방식으로 추출된 다른 샘플들에서도 이 차이가 계속될 것으로 예상될 때 표본편향이 발생했다고 볼 수 있다.

2.1.1 편향

- 통계적 편향은 측정 과정 혹은 표본추출 과정에서 발생하는 계통적인 오차
- 임의표본추출로 인한 오류와 편향에 따른 오류는 신중하게 구분해서 봐야 한다.

- 사격한 총알의 산점도로 보는 예시

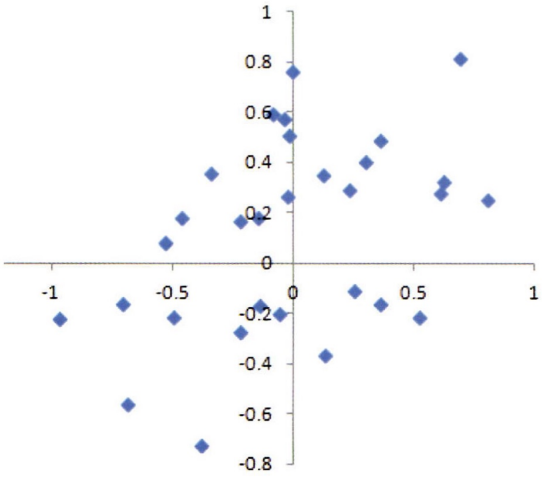

그림 2-2 정확한 조준 사격에 의한 총알의 산점도

편향되지 않은 프로세스에도 오차가 있긴 하지만, 그것은 랜덤하며 어느 쪽으로 강하게 치우치는 경향이 없다.

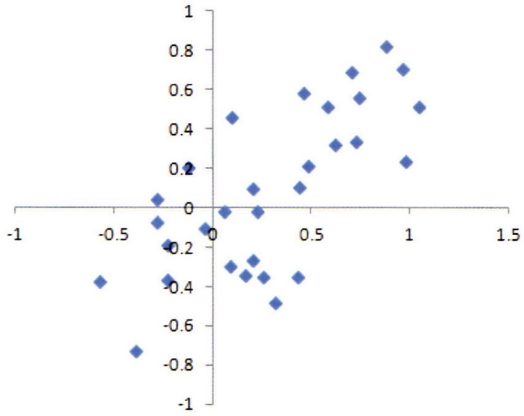

그림 2-3 편향된 조준 사격에 의한 총알의 산점도

x 방향과 y 방향 모두에서 랜덤한 오차가 있고 편향도 있다. 탄착점이 오른쪽 제1사분면에 떨어지는 경향을 볼 수 있다.

2.1.2 임의 선택

- 임의표본추출(동일한 확률로 무작위 추출)은 데이터의 대표성을 담보하는 여러 방법 중 하나이다.
- 임의표본추출이을 위해서는 접근 가능한 모집단을 적절하게 정의하는게 매우 중요하다.
    - 예) 고객의 정의
        - 구매 금액이 0보다 큰 고객
        - 모든 과거 고객
        - 제품을 환불한 고객
        - 내부의 테스트 구매자
        - 사업자
        - 대금 청구 대행사와 고객
- 다음으로 표본추출 절차를 정해야 한다.

2.1.3 크기와 품질 : 크기는 언제 중요해질까

- 빅데이터 시대라고 해도 의외로 데이터 개수가 적을수록 더 유리한 경우가 있다.

- 임의표본 추출에 시간과 노력을 기울일수록 편향이 줄 뿐만 아니라 데이터 탐색 및 데이터 품질에 더 집중할 수 있다.
    - 예) 결측값이나 특잇값으로부터 유용한 정보를 얻는 경우

2.1.4 표본평균과 모평균

- 기호 $\overline{x}$ 는 모집단의 표본평균을 나타내는데 사용되는 반면, m은 모집단으 ㅣ평균을 나타내는 데 사용된다.
- 이 둘을 왜 따로 구분할까?
- 표본에 대한 정보는 관찰을 통해 얻어지고, 모집단에 대한 정보는 주로 작은 표본들로부터 추론하기 때문이다.

▼ 2.2 선택편향

> ✔️ **용어 정리**
> **선택편향** : 관측 데이터를 선택하는 방식 때문에 생기는 편향
> **데이터 스투핑** : 뭔가 흥미로은 것을 찾아 광범위하게 데이터를 살피는 것
> **방대한 검색 효과** : 중복 데이터 모델링이나 너무 많은 예측변수를 고려하는 모델링에서 비롯되는 편향 혹은 비재현성

- 선택 편향은 데이터를 의식적이든 무의식적이든 선택적으로 고르는 관행을 의힘한다.
- 선택 편향은 결국 오해의 소지가 있거나 단편적인 결론을 얻게 된다.
- 보통은 데이터를 먼저 확인한 후 그 안에서 패턴을 찾고자 한다.
- 하지만 이것이 참된 패턴인지 흥미로운 것이 나올 때까지 데이터를 너무 샅샅이 뒤진 결과가 아닌지 확실히 알 수 없다.
- 빅데이터를 반복적으로 조사하는 것이 데이터 과학의 중요한 가치 명제이기 때문에, 선택 편향에 대해 조심할 필요가 있다.
- 데이터 과학자들이 특별히 걱정하는 선택편향의 한 형태는 존 엘더가 **방대한 검색 효과**라고 부르는 것이다.
    - 큰 데이터 집합을 가지고 반복적으로 다른 모델을 만들고 다른 질문을 하다보면 언젠가 흥미로운 것을 발견하기 마련이다.

    - 하지만 그것이 정말로 의미가 있는 것인지, 아니면 우연히 얻은 예의 경우인지 아는 것이 중요하다.
- 성능을 검증하기 위해서 둘 이상의 홀드아웃(holdout) 세트를 사용하면 이를 방지할 수 있다.
- 방대한 검색 효과 외에도, 통계에서 일반적으로 나타나는 선택편향으로는 비임의표본추출, 데이터 체리 피킹 (선별), 특정한 통계적 효과를 강조하는 시간 구간 선택, '흥미로운' 결과가 나올 때 실험을 중단하는 것 등이 여기에 포함된다.

2.2.1 평균으로의 회귀

- 평균으로의 회귀란, 주어진 어떤 변수를 연속적으로 측정했을 때 나타나는 현상이다.
- 예외적인 경우가 관찰되면 그 다음에는 중간 정도의 경우가 관찰되는 경우가 있다.
- 따라서 예외의 경우를 너무 특별히 생각하고 의미를 부여하면 선택편향으로 이어질 수 있다.

> ✔️ **주요 개념**
> 1. 가설을 구체적으로 명시하고 임의표본추출 원칙에 따라 데이터를 수집하면 편향을 피할 수 있다.
> 2. 모든 형태의 데이터 분석은 데이터 수집/분석 프로세스에서 생기는 편향의 위험성을 늘 갖고 있다.

▼ 2.3 통계학에서의 표본분포

> ✔️ **용어 정리**
> **표본통계량** : 더 큰 모집단에서 추출된 표본 데이터들로부터 얻은 측정 지표
> **데이터 분포** : 어떤 데이터 집합에서의 각 개별 값의 도수분포
> **표본분포** : 여러 표본들 혹은 재표본들로부터 얻은 표본통계량의 도수분포
> **중심극한정리** : 표본크기가 커질수록 표본분포가 정규분포를 따르는 경향
> **표준오차** : 여러 표본들로부터 얻은 표본통계량의 변량

- 통계의 표본분포라는 용어는 하나의 동일한 모집단에서 얻은 여러 샘플에 대한 표본통계량의 분포를 나타낸다.

- 일반적으로 우리는 표본을 통해 추정이나 모델링을 하기 때문에 오류가 있을 수 있다.
- 주요 관심사는 **표본의 변동성**이다. 표본에 따라 결과가 얼마나 달라질지에 관심이 있다.

CAUTION
흔히 데이터 분포라고 알려진 개별 데이터 포인트의 분포와 표본분포라고 알려진 표본통계량의 분포를 구별하는 것이 중요하다.

- 평균과 같은 표본통계량의 분포는 데이터 자체의 분포보다 규칙적이고 종 모양일 가능성이 높다.
- 통계의 기반이 되는 표본이 클수록 종모양일 가능성이 더 높아지고 표본통계량의 분포는 좁아진다.

2.3.1 중심극한정리

- 방금 설명한 이러한 현상을 **중심극한정리**라고 한다.
- 모집단이 정규분포가 아니더라도, 표본크기가 충분하고 데이터가 정규성을 크게 이탈하지 않는 경우, 여러 표본에서 추출한 평균은 종모양의 정규곡선을 따른다.
- 중심극한정리 덕분에, 추론을 위한 표본분포에 즉, 신뢰구간이나 가설검정을 계산하는 데에 t분포와 같은 정규근사 공식을 사용할 수 있다.
- 하지만 대부분의 경우 **부트스트랩**(2.4절 참고)을 사용할 수 있기 때문에, 데이터 과학의 관점에서는 중심극한정리가 그렇게 중요하지는 않다.

2.3.2 표준오차

- **표준오차**는 통계에 대한 표본분포의 변동성을 한마디로 말해주는 단일 측정 지표이다.

  표준오차 $= SE = \frac{s}{\sqrt{n}}$ (s : 표준편차, n : 표본크기)
- 표본크기가 커지면 표준오차가 줄어든다.
- 표준오차와 표본크기 사이의 관계를 **n제곱근의 법칙**이라고 한다.

- 실질적으로 표준오차를 추정하기 위해서는 새 샘플을 수집해야 하는데 **부트스트랩** 재표본을 사용하면 새로운 샘플을 뽑을 필요가 없다.
- 현대 통계에서는 부트스트랩은 표준오차를 추정하는 표준 방법이 되었다.
- 부트스트랩은 사실상 모든 통계에서 사용할 수 있으며 중심극한정리 또는 기타 분포 가정에 의존하지 않는다.

▼ 2.4 부트스트랩

> ✔ **용어 정리**
> **부트스트랩 표본** : 관측 데이터 집합으로부터 얻은 복원추출 표본
> **재표본추출** : 관측 데이터로부터 반복해서 표본추출하는 과정, 부트스트램과 순열(셔플링) 과정을 포함한다.

- **부트스트랩** : 현재 있는 표본에서 추가적으로 표본을 복원추출하고 각 표본에 대한 통계량과 모델을 다시 계산하는 것
- 부트스트랩에는 데이터나 표본통계량이 정규분포를 따라야 한다는 가정은 꼭 필요하지 않다.

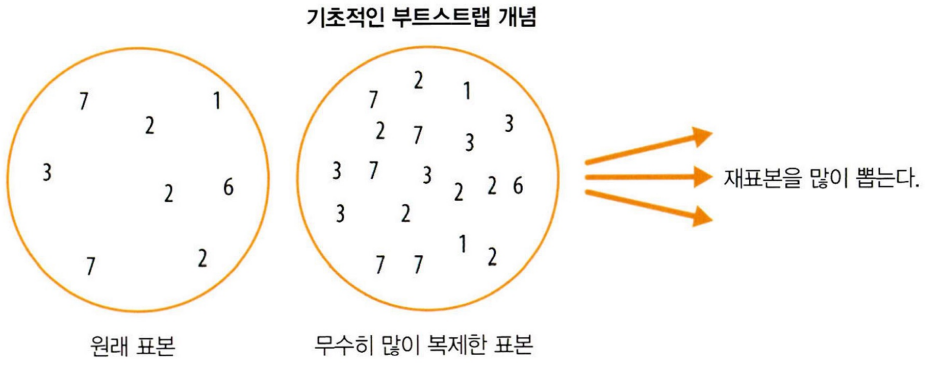

그림 2-7 부트스트랩의 아이디어

- 부트스트랩은 원래 표본을 여러번 복제하여 이로부터 원래 표본으로부터 얻어지는 모든 정보를 포함하는 가상 모집단을 얻게 된다.
- 그런 다음 이 가상 모집다능로부터 표본분포를 추정할 목적으로 표본을 수집할 수 있다.
- 대신 각각의 표본을 뽑은 후 각 관측치를 원래 자리에 돌려놓는 **복원추출**을 한다.

- 이런 식의 방법으로 뽑을 때마다 각 원소가 뽑힐 확률은 그대로 유지하면서 무한한 크기의 모집단을 만들어낼 수 있다.
- 크기 n의 샘플의 평균을 구하는 부트스트랩 재표본추출 알고리즘
  1. 샘플 값을 하나 뽑아서 기록하고 다시 제자리에 놓는다.
  2. n번 반복한다.
  3. 재표본추출된 값의 평균을 기록한다.
  4. 1~3단계를 R번 반복한다. (R은 임의로 설정한다.)
  5. R개의 결과를 사용하여
     a. 표준편차(표본평균의 표준오차)를 계산한다.
     b. 히스토그램 또는 상자그림을 그린다.
     c. 신뢰구간을 찾는다.
- 반복 횟수가 많을수록 표준오차나 신뢰구간에 대한 추정이 더 정확해진다.
- 부트스트랩의 활용
  - 모델 파라미터의 안정성(변동성)을 추정하거나 예측력을 높이기 위해, 부트스트랩 데이터를 가지고 모델을 돌려볼 수 있다.
  - 분류 및 회귀 트리를 사용할 때, 여러 부트스트랩 샘플을 가지고 트리를 여러 개 만든 다음 각 트리에서 나온 예측값을 평균 내는 것 → **배깅**

CAUTION
부트스트랩은 표본 크기가 작은 것을 보완하기 위한 것이 아니다. 새 데이터를 만드는 것도 아니며 기존 데이터 집합의 빈 곳을 채우는 것도 아니다. 모집단에서 추가적으로 표본을 뽑는다고 할 때, 그 표본이 얼마나 원래 표본과 비슷할지를 알려줄 뿐이다.

2.4.1 재표본추출 대 부트스트래핑

- 종종 재표본추출이라는 용어가 부트스트랩과 비슷한 의미로 사용된다.
- **재표본추출**은 여러 표본이 결합되어 비복원추출을 수행할 수 있는 순열 과정을 포함한다.
- **부트스트랩**은 항상 관측된 데이터로부터 복원추출한다는 것을 의미한다.

▼ 2.5 신뢰구간

- 신뢰구간은 통계적 샘플링 원칙에 근거한다.
- 신뢰국나은 항상 90% 또는 95%와 같이 (높은) 백분율로 표현되는 포함 수준과 함께 나온다.
- 90% 신뢰구간이란, 표본통계량의 부트스트랩 표본분포의 90%를 포함하는 구간을 말한다.
- 더 일반적으로, 표본추정치 주위의 x% 신뢰구간이란, 평균적으로 유사한 표본추정치 x% ㅈ정도가 포함되어야 한다.
- 부트스트랩 신뢰구간을 구하는 방법
    1. 데이터에서 복원추출 방식으로 크기 n인 표본을 뽑는다.
    2. 재표본추출한 표본에 대한 원하는 통계량을 기록한다.
    3. 1~2 단계를 R번 반복한다.
    4. x% 신뢰구간을 구하기 위해, R개의 재표본 결과의 분포 양쪽 끝에서 [(100-x)/2]%만큼 잘라낸다.
    5. 절단한 점들은 x% 부트스트랩 신뢰구간의 양 끝점이다.
- 신뢰구간과 관련된 백분율을 **신뢰수준**이라고 부른다.
- 신뢰수준이 높을수록 구간이 더 넓어지며, 표본이 작을수록 구간이 넓어진다. (즉, 불확실성이 커진다.)
- 데이터가 적을수록, 확실히 참값을 얻기에 충분한 신뢰구간을 확보해야 한다.

▼ 2.6 정규분포

- 종 모양의 종규분포는 전통적인 통계의 상징이다.
- 표본통계량 분포가 보통 어떤 일정한 모양이 있다는 사실은 이 분포를 근사화하는 수학 공식을 개발하는 데 강력한 도구가 되었다.

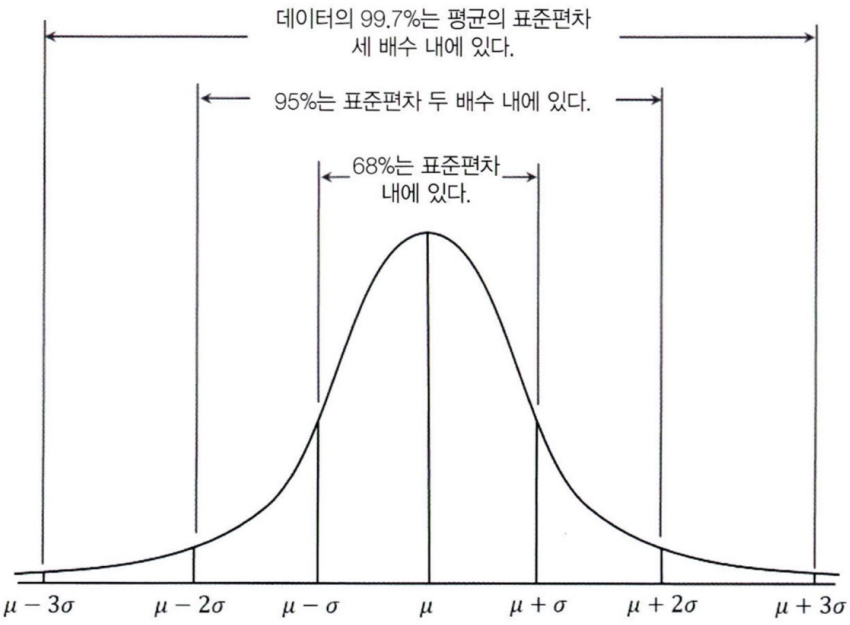

그림 2-10 정규곡선

- 위 그림의 정규분포에서 데이터의 68%는 평균의 표준편차 내에 속하며 95%는 표준펴차 두 배수 내에 있다.

2.6.1 표준정규분포와 QQ그림

- 표준정규분포는 x축의 단위가 평균의 표준편차로 표현되는 정규분포를 말한다.
- 데이터를 표준정규분포와 비교하려면 데이터에서 평균을 뺀 다음 표준편차로 나누면 된다. 이를 **정규화** 또는 **표준화** 라고 한다.
- 이렇게 변환한 값을 **z 점수**라고 하며, 정규분포를 **z 분포**라고도 한다.
- QQ그림은 표본이 특정 분포에 얼마나 가까운지를 시각적으로 판별하는데 사용된다.
- QQ그림은 z 점수를 오름차순으로 정렬하고 각 값의 z점수를 y축에 표시한다.
- x 축은 정규분포에서의 해당 분위수를 나타내며, 단위는 평균으로부터 떨어진 데이터의 표준편차 수에 해당한다.

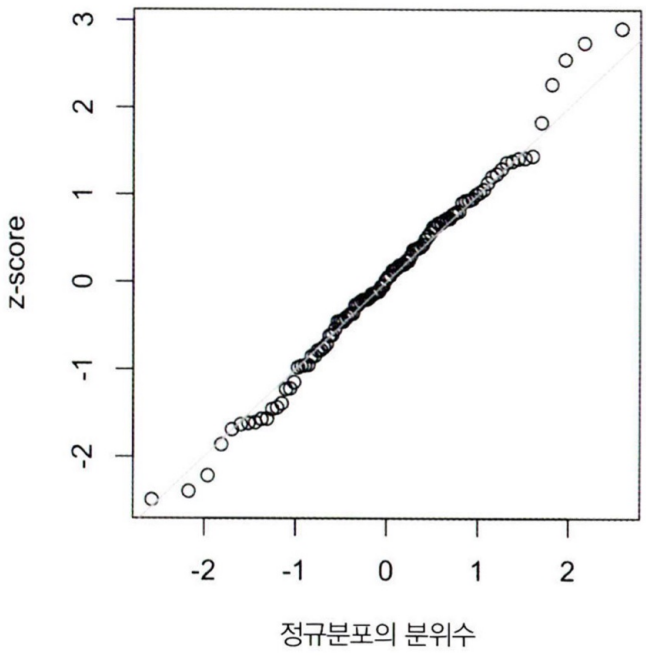

그림 2-11 표준정규분포로부터 추출한 100개 표본의 QQ 그림

▼ 2.7 긴 꼬리 분포

- 정규분포의 중요성에도 불구하고 데이터는 일반적으로 정규분포를 따르지 않는다.
- 정규분포가 일반적으로 원시 데이터 분포의 특징을 나타내지는 않는다.
- 대칭 및 비대칭 분포 모두 **긴 꼬리**를 가질 수 있다.
- 분포의 꼬리를 양 극한값에 해당한다.
- 실무에서는 긴 꼬리와 긴 꼬리를 잘 들여다보는 것을 중요하게 여긴다.
- 예) 넷플릭스의 주가수익률

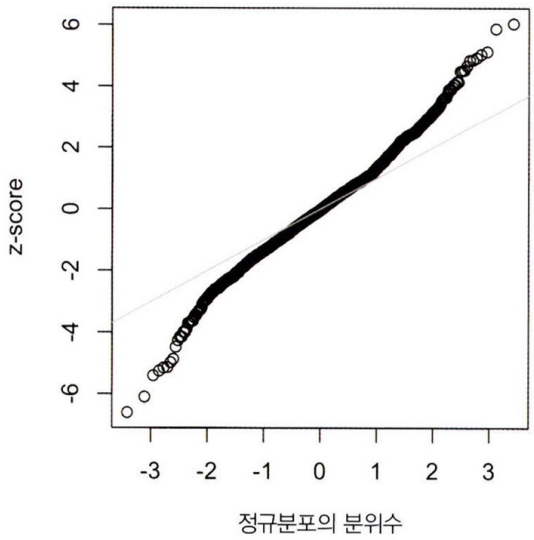

그림 2-12 넷플릭스(NFLX)의 일일 주식 수익률에 대한 QQ 그림

- 위 그림은 [그림 2-11]과 달리 낮은 값의 점들은 대각선보다 훨씬 낮고 높은 값은 선보다 훨씬 위에 위치한다.
- 이는 데이터가 정규분포를 따르지 않는다는 것을 의미한다.
- 또한 이는 데이터가 정규분포를 따른다고 할 때 예상되는 것보다 훨씬 더 많은 극단값을 관찰할 가능성이 있음을 의미한다.

▼ 2.8 스튜던트의 t 분포

- t-분포는 정규분포와 생김새가 비슷하지만, 꼬리 부분이 약간 더 두껍고 길다.
- t-분포는 표본통계량의 분포를 설명하는 데 광범위하게 사용된다.
- 일반적으로 표본평균의 분포는 t 분포와 같은 모양이며, 표본크기에 따라 다른 계열의 t 분포가 있다.
- 표본통계량의 상태를 묘사할 때 t 분포의 정확도는 표본에 대한 통계량의 분포가 정규분포를 따른다는 조건을 필요로 한다.
- 원래 모집단이 정규분포를 따르지 않을 때 조차도, 표본통계량은 보통 정규분포를 따르는 것으로 나타났다. 이는 이미 앞에서 봤던 현상으로 **중심극한정리**라고 부른다.

▼ 2.9 이항분포

- 이항식(예/아니오)의 결론은 구매/구매하지 않음, 클릭/클릭하지 않음, 생존/사망 등과 같은 의사 결정 과정에서 아주 중요하기 때문에 분석에서 핵심이라고 할 수 있다.
- 이항분포를 이해할 때 핵심 개념은 일련의 **시행**들인데, 각 시행은 정해진 확률로 두 가지 결과를 갖는다.
- 예/아니오 또는 0/1 결과를 이진 결과라고 한다.꼭 50대 50의 확률을 가질 필요는 없다. 확률의 합이 1.0이 되면 된다.
- **이항분포**란, 각 시행마다 그 성공확률(p)이 정해져 있을 때, 주어진 시행 횟수(n) 중에서 성공한 횟수(x)의 도수분포를 의미한다.
- 이항분포의
  - 평균 = nXp
  - 분산 = $nXp(1-p)$
- 시행 횟수가 충분할 경우(특히, p가 0.50에 가까울 때) 이항분포는 사실상 정규분포와 구별이 어렵다.
    - 대부분의 통계 절차에서는 평균과 분산으로 근사화한 정규분포를 사용한다.

▼ 2.10 카이제곱분포

- 통계학에서 중요한 개념에는 범주의 수에 대해 기댓값에서 이탈하는 것이 있다.
- 기댓값이란 '데이터에서 특이하거나 주목할 만한 것이 없다'는 의미로 대략 정의할 수 있다.
- 이를 '귀무가설' 또는 '귀무 모델'이라고도 한다.
- 카이제곱통계량은 검정 결과가 독립성에 대한 귀무 기댓값에서 벗어난 정도를 측정하는 통계량이다.
- 더 일반적으로 카이제곱통계량은 관측 데이터가 특정 분포에 '적합'한 정도를 나타낸다(적합도검정).
- **카이제곱분포**는 귀무 모델에서 반복적으로 재표본추출한 통계량 분포다.
- 개수 집합에 대해 카이제곱 값이 낮다는 것은 기대 분포를 거의 따르고 있음을 나타낸다.
- 카이제곱 값이 높다는 것은 기대한 것과 현저하게 다르다는 것을 나타낸다.

▼ 2.11 F 분포

- 과학 실험의 일반적인 절차는 여러 그룹에 걸쳐 서로 다른 처리를 테스트하는 것이다.
- F 통계량은 각 그룹 내 변동성에 대한 그룹 평균 간 변동성의 비율을 의미한다.
- 이러한 비교를 **분산분석**(ANOVA)이라고 한다.
- F 통계량의 분포는 모든 그룹의 평균이 동일한 경우(즉, 귀무모델) 무작위 순열 데이터에 의해 생성되는 모든 값의 빈도 분포다.

▼ 2.12 푸아송 분포와 그 외 관련 분포들

2.12.1 푸아송 분포

- 푸아송분포는 시간 단위 혹은 공간 단위로 표본들을 수지밯ㄹ 때, 그 사건들의 분포를 알려준다.
- 푸아송 분포의 핵심 파라미터는 λ(람다)이다. 람다는 어떤 일정 시간/공간의 구간 안에서 발생한 평균 사건 수를 의미한다. 푸아송 분포의 분산 역시 λ(람다)이다.

2.12.2 지수 분포

- 지수분포는 푸아송 분포에 사용된 것과 동일한 변수 λ(람다)를 사용하여 사건과 사건 간의 시간 분포를 모델링할 수 있다.
- 푸아송이나 지수분포에 대한 시뮬레이션 연구에서 핵심은 λ(람다)가 해당 기간 동안 일정하게 유지된다는 가정이다.