데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

신뢰구간과 가설검정

1. 중심극한정리

탐색적 자료 분석 vs 확증적 자료 분석

- 데이터 분석의 첫 번째 단계는 탐색적 자료 분석(explorato-ry data analysis)이다.
- 두 번째 단계인 확증적 자료 분석(confirmatory data an-alysis)은 일반적으로 자료에 관한 수치적 요약치를 제시하는 것으로 시작한다.
- 이러한 요약치를 일반적으로 통계량(statistics)이라고 하고 통계량은 표본에 따라서 값이 다르게 나올 수 있기 때문에 이러한 변동성을 알아보기 위해서 표본분포(sampling distrib-ution)을 아는 것이 중요하다.
- 표본분포를 알아내기 위해 (1)붓스트랩을 사용하거나 (2)통계이론을 사용할 수 있다.

표본분포

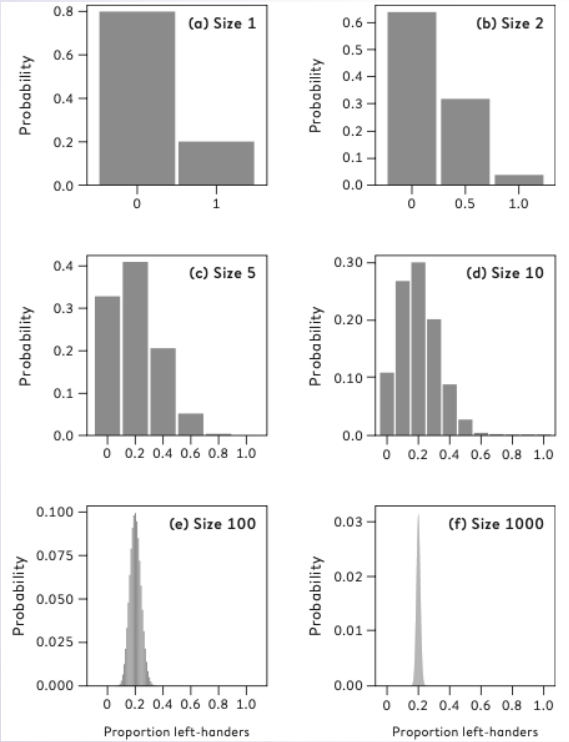
- 표본분포를 이론적으로 알아내는 방법을 설명하기 위해서 다음 예제를 생각해보자.
- 정확히 왼손잡이가 20%이고 오른손잡이가 80%인 모집단에서 크기가 서로 다른 표본을 뽑았을 때 각 표본에서 관측되는 왼손잡이의 비율에 관한 표본분포에 대해 알아보자.

표본분포

- 여기서 통계량은 다음과 같이 정의되는 표본비율이다. 먼저 한 명을 뽑았을 때 왼손잡이일 경우 1, 오른손잡이일 경우 0으로 정의되는 확률변수 X를 정하자.
- 표본비율은 이러한 확률변수들의 평균이다. 즉 n명의 표본을 뽑았을 경우 표본비율은 다음과 같이 정의된다.

$$\hat{p} = \sum_{i=1}^{n} X_i / n$$

이항분포(Binomial Distribution)

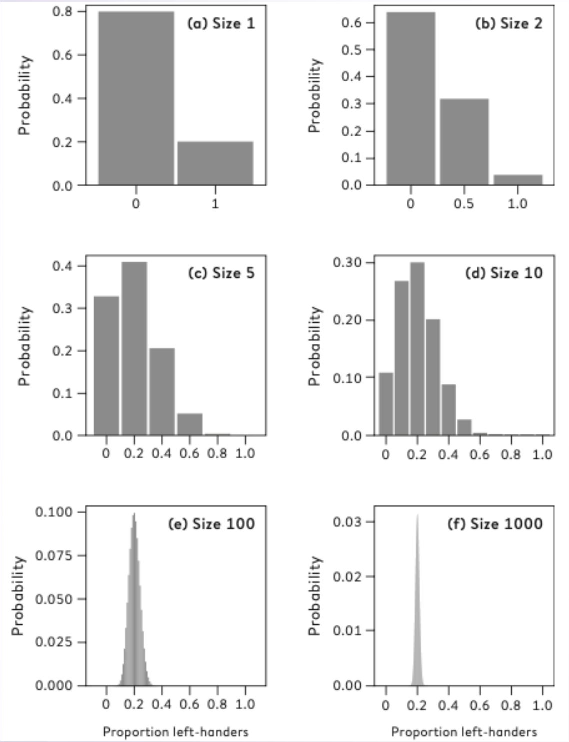


[왼손잡이의 비율이 20%인 모집단에서 크기가 1, 2, 5, 10, 100, 1,000인 표본을 뽑았을 때 표본비율의 분포]
(The Art of Statistics, p232)

기타 출처 #1

- 왼쪽 그림은 표본크기가 1, 2, 5, 10, 100, 1,000인 경우 표본비율의 분포를 보여준다. 예를 들면 표본크기가 1일 경우 표본비율이 0이 될 확률은 0.8이며 1이 될 확률은 0.2이다.

이항분포(Binomial Distribution)



[왼손잡이의 비율이 20%인 모집단에서 크기가 1, 2, 5, 10, 100, 1,000인 표본을 뽑았을 때 표본비율의 분포]
(The Art of Statistics, p232)

기타 출처 #1

- 표본크기가 2일 경우 표본비율의 가능한 값은 총 3가지(0, 0.5, 1)이며 각각의 경우 확률은 0.64, 0.32, 0.04로 주어진다.
- 위와 같은 확률분포를 이항분포라고 한다.

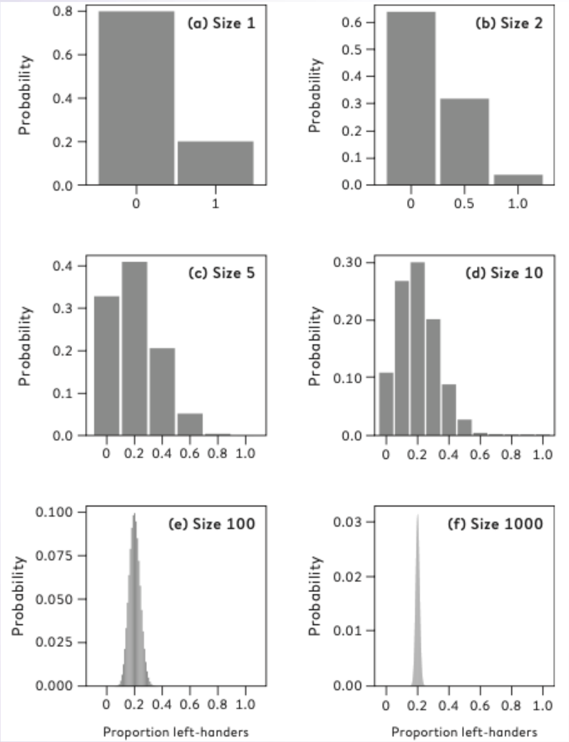
이항분포(Binomial Distribution)

- 이항분포는 앞에서 배운 베르누이 분포를 확장한 경우라고 생각할 수 있다. 즉 베르누이 확률변수 X는 두 가지 값 0(실패), 1(성공)을 가질 수 있으며 이 분포는 $p = Pr(X = 1)$, (성공 확률)이 분포의 모양을 결정한다.
- 베르누이 분포를 n번 시행한다고 가정하자. 이 경우 성공 횟수 ($Y = \sum_{i=1}^{n} X_i$)의 분포가 이항분포를 따른다고 한다.
- 예를 들면 동전던지기에서 앞면이 나오는 경우 1, 뒷면이 나오는 경우 0의 값을 가지는 확률변수를 정의하고 동전던지기를 10번 했을 때 앞면이 나오는 횟수는 이항분포를 따른다.

이항분포(Binomial Distribution)

- 앞의 왼손잡이의 예에서는 표본 크기가 10인 경우는 시행 횟수 $n=10$, 성공확률 $p=0.2$인 이항분포이다.
- 이항분포에서 평균은 $E(Y) = n \cdot p$이다. 즉 표본을 10개 뽑았을 경우 왼손잡이는 평균적으로 2명이 있을 것으로 생각된다.
- 표본비율 $\hat{p}$은 표본에서 전체 왼손잡이의 숫자를 표본크기로 나눈 것으로 생각할 수 있으며 표본비율과 같은 통계량의 표준편차를 표준오차(Standard Error)라고 한다.

이항분포(Binomial Distribution)



[왼손잡이의 비율이 20%인 모집단에서 크기가 1, 2, 5, 10, 100, 1,000인 표본을 뽑았을 때 표본비율의 분포]
(The Art of Statistics, p232)

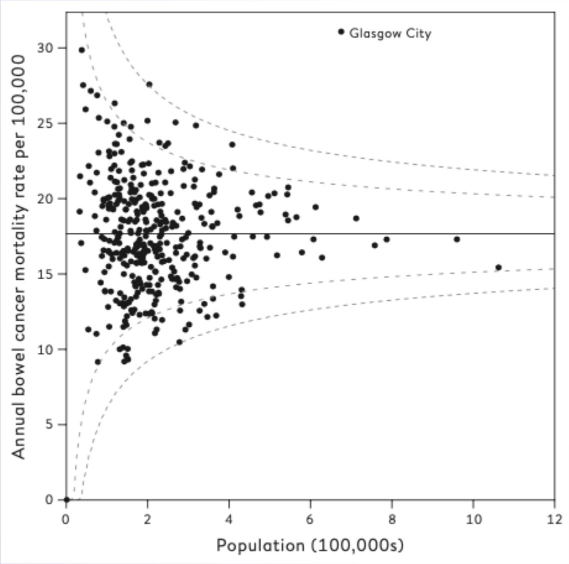
- 왼쪽 그림에서 표본크기가 커질수록 이항분포의 모습이 (1)대칭적인 정규분포의 모습과 가까워지고 (2)분포가 가운데로 집중된다는 것을 알 수 있다.

기타 출처 #1

지역별 대장암 사망률은 왜 차이가 큰가?

- 다음 예제를 통해서 앞의 왼손잡이 예제에서 표본크기 증가에 따른 분포 형태 변화에 관한 설명을 해보자.
- 2011년 9월 영국 BBC 뉴스에 영국의 지역별 대장암 사망률 차이가 최대 3배까지 이른다는 기사가 실렸다. 지역별 대장암 사망률 차이의 원인으로 불균등한 의료서비스가 지목되었다.
- 위의 주장이 사실인지를 알아보기 위해 폴 바든은 영국의 380개 지자체별 인구와 대장암 사망률에 관한 산점도를 그려보았다. 이런 그림은 funnel plot이라고 한다.

지역별 대장암 사망률은 왜 차이가 큰가?

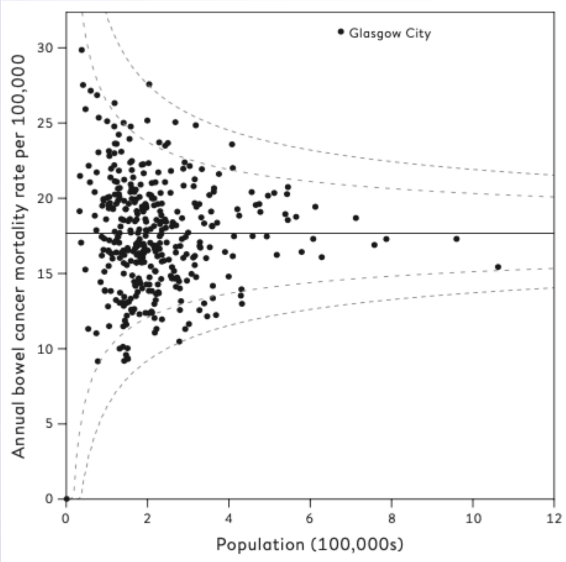


[영국 380개 지자체별 인구 10만 명당 대장암 사망자 숫자]
(The Art of Statistics, p235)

- 왼쪽 그림은 funnel plot을 보여준다. 이 그림에서 몇 가지 특징을 찾아볼 수 있다.
- 먼저 인구 숫자가 작은 지자체의 경우 사망률이 아주 높거나 낮은 경우가 많다. 사실 한국에서도 특정 암의 사망률로 지자체별 순위를 매긴다면 비슷한 현상을 관측할 수 있다.

기타 출처 #2

지역별 대장암 사망률은 왜 차이가 큰가?



[영국 380개 지자체별 인구 10만 명당 대장암 사망자 숫자]
(The Art of Statistics, p235)

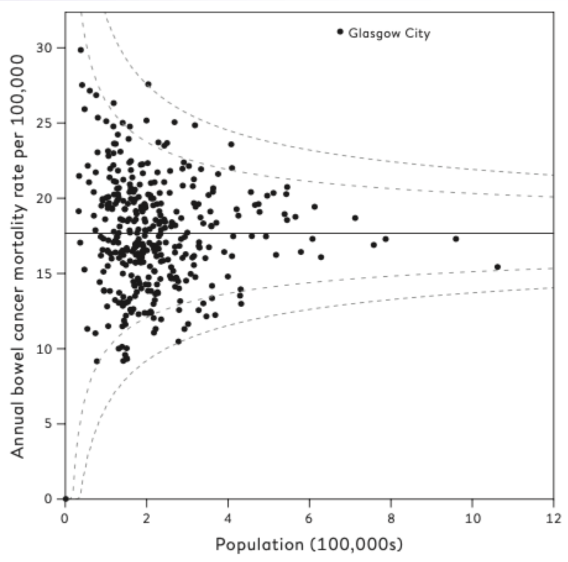
- 즉 인구가 작을 경우 약간의 변동만 있더라도 사망률에 큰 차이가 있을 수 있다. 따라서 그 사실을 고려해서 옆의 그림에서는 두 개의 점선으로 표시된 control limit을 표시하였다.

기타 출처 #2

지역별 대장암 사망률은 왜 차이가 큰가?

- Control limit은 각 지자체에 사망자 숫자가 이항분포를 따른다는 사실에 기반하여 그려졌다.
- 영국에서 대장암 사망률은 0.000176으로 알려져 있다. 사망한 경우를 성공(!)이라고 간주한다면 각 지자체에서 대장암으로 사망한 사람의 숫자는 시행 횟수가 지자체 인구이고 성공률이 대장암 사망률인 이항분포를 따른다고 할 수 있다.

지역별 대장암 사망률은 왜 차이가 큰가?



[영국 380개 지자체별 인구 10만 명당 대장암 사망자 숫자]
(The Art of Statistics, p235)

- Control limit은 대장암 사망률을 중심으로 인구별로 ±2·(표준오차), ±3·(표준오차)를 그린 점선으로 이 안에 각각 95%, 99.8%의 데이터가 포함되어 있으리라 생각한다.
- 왼쪽 그림에서 인구를 고려했을 때 글래스고가 특이하게 대장암 사망률이 높음을 알 수 있다.

기타 출처 #2

대수의 법칙(Law of Large Numbers)

- 왼손잡이 예제와 대장암 사망률 예제에서 알 수 있듯이 표본 크기가 커짐에 따라 표본비율(왼손잡이 비율, 대장암 사망률)이 평균 근처로 점차 좁혀짐을 알 수 있다.
- 이 이유는 18세기 초 스위스 수학자 야코프 베르누이(Jacob Bernoulli)에 의해서 확립된 대수의 법칙으로 설명할 수 있다.

대수의 법칙(Law of Large Numbers)

- 동전을 던져서 앞면이 나오는 비율을 생각해보자. 처음 10번을 던질 경우 비율이 0.7 혹은 0.2와 같은 값이 나오는 경우는 종종 있다. 그렇지만 만 번을 던진다면 그 비율은 0.5에 수렴한다.
- 여기서 중요한 사실은 표본비율(표본평균)은 특정 값으로 수렴하지만, 앞면이 나온 횟수와 뒷면이 나온 횟수의 차이가 줄어들지는 않는다! 위키피디아의 다음 컴퓨터 모의실험을 본다면 이해가 쉽게 된다. (https://bit.ly/2KKdnFG)

기타 출처 #3

도박사의 오류(Gambler's Fallacy)

- 야구 경기에서 3할 타자가 오늘 첫 2타석에 안타를 치지 못했을 경우 해설자가 이번 타석에는 이 선수가 안타를 칠 때가 되었다고 얘기하는 경우가 종종 있다.
- 마찬가지로 동전을 던져서 연속으로 4번 앞면이 나왔다면 이번에는 뒷면이 나올 확률이 높을 것이라고 기대하는 것이 자연스러워 보인다.
- 야구의 타율과 동전의 앞면의 비율이 모두 표본비율이므로 대수의 법칙을 생각해보면 위의 주장이 일리가 있다고 생각할 수 있다.

도박사의 오류(Gambler’s Fallacy)

- 하지만 이항분포에서 각 시행 간은 서로 독립이라고 가정하기 때문에 이전의 결과가 지금의 시행에 영향을 주지 않는다!
- 또한 앞에서 얘기한 시행 횟수가 증가하더라도 성공과 실패 횟수 간의 차이가 줄어들지 않는다는 것으로 위의 주장이 사실이 아님을 알 수 있다.

중심극한정리(Central Limit Theorem)

- 예제들에서 표본비율은 단순히 특정 값에 수렴하는 것만이 아니라 분포가 정규분포 형태를 띠게 되는 것을 관측할 수 있었다.
- 표본크기가 커질수록 이런 현상이 관측되는 것은 표본비율에만 한정된 이야기가 아니다. 표본평균들도 표본크기가 증가하면 분포의 형태가 정규분포 모양을 갖게 된다.
- 위의 현상은 1733년 프랑스 수학자 아브라함 드 무아브르에 의해서 다음과 같은 중심극한정리라는 이름으로 증명되었다.
- 모집단의 평균과 분산을 각각 μ, σ^2라고 하자. 표본의 크기가 증가하면 표본평균은 평균과 분산이 $\mu, \sigma^2/n$인 정규분포 형태를 가진다.

오늘의 강의 요점

- 이항분포
- 대수의 법칙
- 중심극한정리

출처

#1~2 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

#3 Wikipedia https://bit.ly/2KKdnFG

신뢰구간과 가설검정

2. 신뢰구간

통계량의 불확실성

- 앞에서 우리는 확률분포를 통해서 데이터가 어떻게 생성되는지 알아보았다.
- 우리가 통계를 배우는 목적은 관측된 데이터로부터 데이터를 생성하는 확률모형에 관한 추론을 하기 위해서이다. 예를 들면 왼손잡이 예제에서 관측된 자료를 토대로 실제 전체 모집단의 왼손잡이 비율(모수)을 알고 싶다.

통계량의 불확실성

- 만약 통계학 과목 수강생 20명에게 한국에서 왼손잡이 비율을 추정하라는 과제를 주었다고 가정하자. 수강생 각각은 다른 표본을 이용하기 때문에 서로 다른 표본비율을 추정치로 제시할 것이다.

- 이러한 추정치(통계량)의 변동성 또는 불확실성을 추정치와 같이 제시할 필요가 있다. 변동성이 작다면 추정치는 보다 신뢰할 만한 값을 제시한다고 볼 수 있다.

신뢰구간

- 통계량의 불확실성은 일반적으로 표준오차(통계량의 표준편차), 혹은 신뢰구간을 이용하여 제시한다.

신뢰구간

- **신뢰구간의 아이디어는 다음과 같은 절차를 통해 도출되었다.**
  - 중심극한정리를 이용하여 알고자 하는 모집단의 모수(왼손잡이 비율)에 대해 추정치가 그 안에 포함될 확률이 95%인 예측구간을 먼저 구한다. Funnel plot에서 본 control limit이 95%, 99% 예측구간이다. 즉 $Pr(\overline{X} \in (\mu - 2 \cdot SE, \mu + 2 \cdot SE)) = 0.95$, 여기서 SE는 표본평균의 표준오차를 의미한다.
  - 실제 데이터를 이용하여 추정치를 계산한다.(각각의 학생들이 다른 추정치를 계산할 수 있다.)
  - 통계량이 95% 예측구간 안에 놓일 수 있는 모수의 범위를 구한다. 이 범위를 95% 신뢰구간이라고 한다. 위의 경우 95% 신뢰구간은 $(\overline{X} - 2 \cdot SE, \overline{X} + 2 \cdot SE)$

신뢰구간

- 다음 예제를 통해서 신뢰구간을 구하는 방법과 그 의미에 대해서 알아보자.
- 임의로 선정한 50명의 서울대 학생들에게 연애 횟수를 물어보았다. 결과는 평균 3.2회, 표준편차는 1.7회였다. 연애 횟수에 대한 95% 신뢰구간을 구해보자.

신뢰구간

- 95% 신뢰구간은 $(\overline{X}-2\cdot SE, \overline{X}+2\cdot SE)$ 으로 주어지면 여기서 표준오차는 중심극한정리에 의해서 $\sqrt{\sigma^2/n}=\sigma/\sqrt{n}$으로 주어진다. 여기서 모집단의 표준편차 σ는 알지 못하기 때문에 일반적으로 대신 표본 표준편차를 대신 사용한다.
- 따라서 연애 횟수에 대한 95% 신뢰구간은

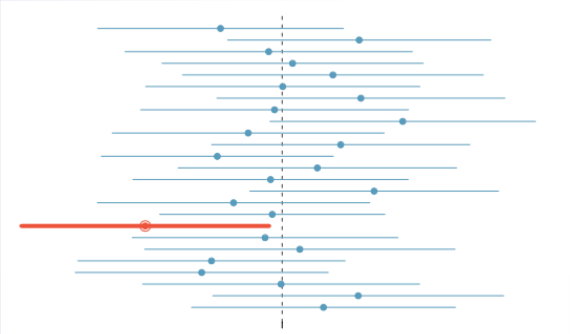
$$\left(3.2-2\cdot\frac{1.7}{\sqrt{50}}, 3.2+2\cdot\frac{1.7}{\sqrt{50}}\right)=(2.7, 3.7)$$

95% 신뢰구간의 의미

- 앞의 결과를 다음과 같이 해석할 수 있다.
- 서울대학생의 평균 연애 횟수가 2.7에서 3.7 사이에 있을 확률이 95%이다. (X)
- 서울대학교 학생을 대상으로 표본을 100개 뽑아 연애 횟수에 대한 신뢰구간을 만든다면 이중 평균적으로 95개의 신뢰구간이 모집단의 평균 연애 횟수를 포함하고 (2.7, 3.7)은 이렇게 구해진 100개의 신뢰구간 중 하나이다. (O)

95% 신뢰구간의 의미

- 서울대학교 학생 전체를 대상으로 연애 횟수 조사를 한 결과 연애 횟수 평균은 3.14이였다고 가정하자. 즉 모평균이 3.14이다.



[연애 횟수에 관한 신뢰구간]
(The Art of Statistics, p182)

- 왼쪽 그림은 25개의 표본을 이용하여 만든 신뢰구간에서 빨간색으로 표시된 1개의 신뢰구간은 실제 모집단의 평균 연애 횟수를 포함하지 않는 경우를 보여준다.

- 여기서 주의할 점은 25개 표본의 크기가 반드시 같지 않아도 된다는 점이다.

기타 출처 #1

중심극한정리와 붓스트랩을 이용한 95% 신뢰구간

	추정값	표준오차	95% 신뢰구간
중심극한정리	0.33	0.05	(0.23, 0.42)
붓스트랩	0.33	0.06	(0.22, 0.44)

[어머니와 딸 키 관계에 관한 회귀계수 추정값, 표준오차, 95% 신뢰구간]
(The Art of Statistics, p243)

기타 출처 #2

여론조사에서 95% 신뢰구간

- 여론조사에서 오차범위는 어떻게 정해지는 것일까?
- 여기서 오차범위도 마찬가지로 $2 \cdot SE$를 바탕으로 한다. 그런데 표본비율의 표준오차는 어떻게 구할까? 우선 표본비율의 분산을 구하면 다음과 같다.

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}(Y/n) = Var(Y)/n^2 = np(1-p)/n^2 = p(1-p)/n$$

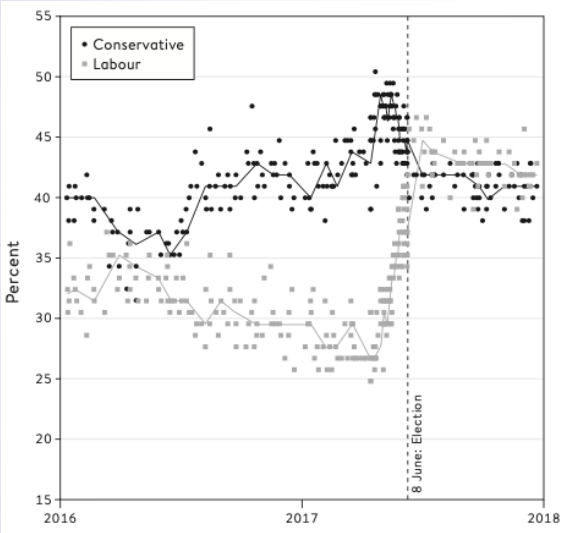
- 앞에서는 표본표준편차를 구하여 모집단의 표준편차 대신 사용하였지만, 이 경우 표준오차의 공식에 p가 들어가 있기 때문에 표본표준오차를 사용하는 것이 사실 적절하지 않을 수 있다.

여론조사에서 95% 신뢰구간

- 그래서 대신 표본비율의 분산이 최대가 되는 경우는 $p=1/2$을 활용하여 표본비율은 분산을 $1/(4\mathrm{n})$으로 대체할 수 있다. 따라서 이경우 오차범위는 $2 \cdot \sqrt{1/(4n)} = 1/\sqrt{n}$으로 주어진다.
- 따라서 표본크기가 1,000명이라면 여론조사의 오차범위는

  $1/\sqrt{1000} \times 100\%=3\%$로 간주할 수 있다.

여론조사에서 95% 신뢰구간



[2017년 영국 BBC의 총선 관련 보수당과 노동당의 지지도 여론조사 자료]
(The Art of Statistics, p246)

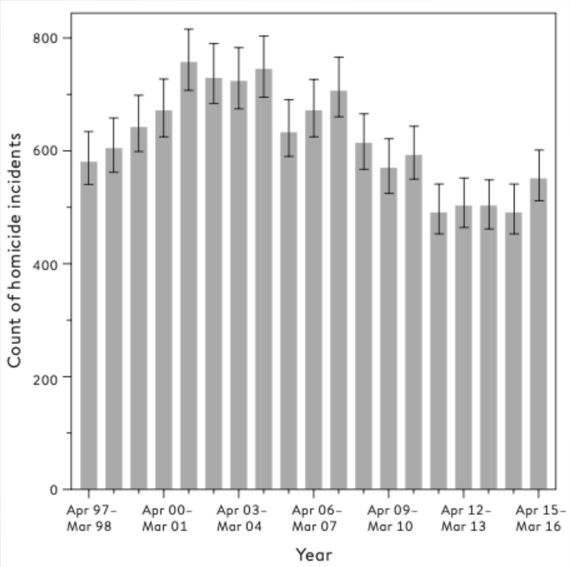
- 왼쪽 그림은 2017년 영국 BBC에서 총선 관련 여론조사 데이터를 보여준다. 가운데 선은 이전 7번 조사의 중앙값을 나타낸다.
- 1,000명의 대상으로 한 여론조사이기 때문에 오차범위는 3%로 생각할 수 있지만, 실제 여론조사의 변동성은 이 범위를 넘는 것으로 볼 수 있기 때문에 이 여론조사 방법에는 문제가 있어 보인다.

기타 출처 #3

영국에서 살인사건은 계속 증가하고 있는가?

- 설문조사와 같이 모집단에서 임의로 추출한 표본을 바탕으로 관심 있는 모집단의 특징(모수)에 대한 추론을 할 경우 오차 범위를 제시하는 것은 자연스럽다.
- 하지만 만약 우리가 모집단 전체의 자료를 가지고 있다면 신뢰 구간을 제시하는 것은 의미가 있을까?

영국에서 살인사건은 계속 증가하고 있는가?

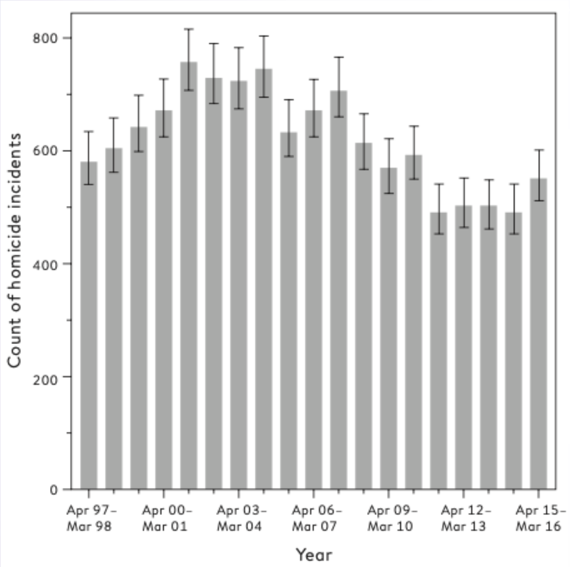


[1998년부터 2016년까지 영국과 웨일스에서 발생한 살인사건 숫자와 95% 신뢰구간]
(The Art of Statistics, p250)

- 예를 들면 영국에서 살인사건이 계속 증가하고 있는지 여부에 대해 알고 싶다면 우선 연간 살인사건의 통계를 살펴볼 것이다.
- 영국통계청에서는 2014년 4월부터 2015년 3월까지는 497건의 살인사건을, 이듬해 같은 기간에는 557건의 살인사건을 보고하였다.

기타 출처 #4

영국에서 살인사건은 계속 증가하고 있는가?

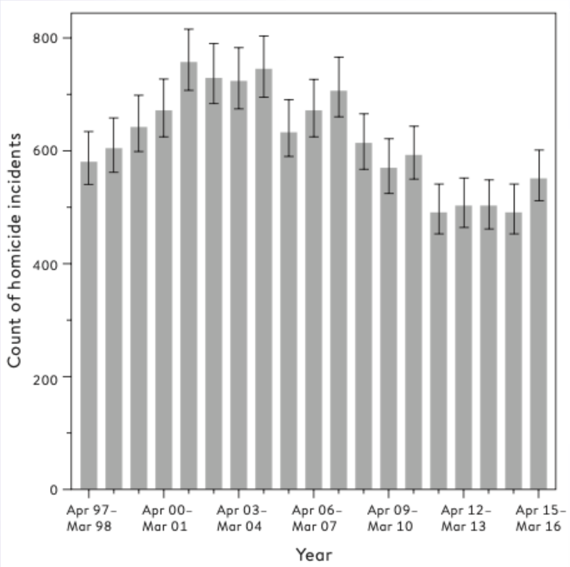


[1998년부터 2016년까지 영국과 웨일스에서 발생한 살인사건 숫자와 95% 신뢰구간]
(The Art of Statistics, p250)

- 이 경우 매년 살인사건의 발생 건수를 특정 확률분포(포아송 분포)에서 생성된 관측치라고 생각하면 여전히 신뢰구간을 구할 수 있다.
- 포아송 분포는 평균과 분산이 모두 같다. 따라서 표준오차는 "표본평균의 제곱근/표본크기의 제곱근"으로 계산할 수 있다.

기타 출처 #4

영국에서 살인사건은 계속 증가하고 있는가?



[1998년부터 2016년까지 영국과 웨일스에서 발생한 살인사건 숫자와 95% 신뢰구간]
(The Art of Statistics, p250)

- 왼쪽 그림은 위의 공식을 이용하여 매년 살인사건에 대한 95% 신뢰구간을 제시하고 있다.
- 여기서 신뢰구간은 실제 살인사건의 건수에 대한 신뢰구간이 아니라 살인사건이 생길 수 있는 기저 건수(즉 모집단의 평균)에 관한 신뢰구간이다.

기타 출처 #4

영국에서 살인사건은 계속 증가하고 있는가?

- 실제 살인사건의 평균이 변화했는지 여부를 알아보기 위해서는 두 해 살인사건 **평균의 차이**의 신뢰구간을 구한 후 신뢰구간이 0을 포함하는지 여부를 알아보는 것이 정확하다.

- 2014년~2015년에는 497건, 2015~2016년에는 557건의 살인사건이 발생했으므로 총 60건이 증가하였다. 이 경우 두 해 살인사건 평균의 차이의 95% 신뢰구간을 구한 결과(-4,124)이며 신뢰구간이 0을 포함하고 있으므로 변화가 있었다고 확신할 수는 없다. 다만 0이 신뢰구간의 끝자락에 걸쳐 있으므로 변화가 전혀 없다고 주장하는 것보다 좀 더 자료를 추가하여 경향을 파악하는 것을 고려할 수 있다.

오늘의 강의 요점

- 신뢰구간의 공식
- 신뢰구간의 의미

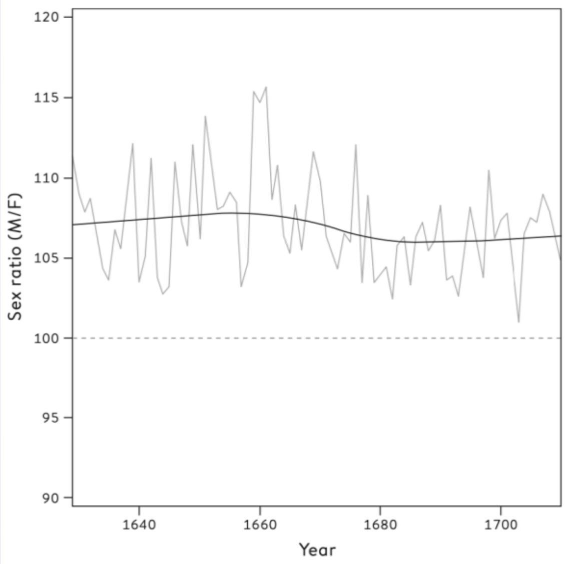
출처

#1~4 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

신뢰구간과 가설 검정

3. “무죄”가 아니라 “유죄라고 할 수 없다”가 맞다

남아 출생률이 여아 출생률보다 높은가?



[1629에서 1701년 사이 런던에서 태어난 유아의 남녀 성비]
(The Art of Statistics, p254)

- 1705년 앤 여왕의 주치의 존 아버스넛은 남아 출생률이 여아 출생률보다 높은지 여부가 궁금하였다.
- 왼쪽 그림은 그가 1629년에서 1701년 사이 런던에서 치러진 유아 영세를 기준으로 남녀성별 출생 비율을 조사한 결과이다.
- 82년 동안 전체 성비는 107이었고 매년 성비는 101에서 116 사이에서 변동하고 있었다.

기타 출처 #1

남아 출생률이 여아 출생률보다 높은가?

- 아버스넛은 유아의 남녀 성 비율이 1이라고 가정한다면 이런 데이터를 관측한 확률은 굉장히 작을 것이라고 생각했다(정확히 $1/2^{82}$이다).
- 아버스넛은 상대적으로 높은 남성 사망률을 극복하기 위해 창조주가 남녀 성비를 1 이상이 되도록 조정한다고 결론을 내렸다.
- 오늘날 자연스러운 성비는 대략 105이다. 즉 여자아이 20명당 남자아이가 21명씩 태어난다. 참고로 한국의 신생아 남녀 성비는 1999년 109.5였으나 2019년에는 105.5였다.

PPDAC에서 Analysis

- **데이터 분석 문제해결 방식 5단계에서 4단계에 해당하는 분석에 대해서 자세히 더 논의해 보자.**
- **다음과 같은 과학 가설에 대해서 어떻게 분석을 할 것인가?**

1. 영국에서 실업률이 지난 4분기에 변했는가?
2. 스타틴 복용이 기저질환이 있는 중년 남자의 심장마비나 뇌졸중 위험을 감소시키는가?
3. 아버지의 키를 통제하면 어머니의 키와 아들의 키 사이에 연관성이 있는가?
4. 힉스 입자는 존재하는가?

PPDAC에서 Analysis

- **앞의 질문은 통계학을 통해 다양한 분야의 질문에 답변을 할 수 있음을 알 수 있다.**

1. 실업률의 변화: 주어진 시간과 장소에서 벌어지는 특정 사건에 관한 질문
2. 스타틴: 특정 그룹에 국한된 의학적 명제
3. 어머니의 키: 일반적인 과학적 관심사
4. 힉스 입자: 우주의 물리법칙에 관한 근본적인 고민

- **이러한 질문에 답변하기 위해 우리는 (통계적)가설 검정을 사용할 것이다.**

가설 검정

- 가설이란 어떤 현상에 관한 설명으로 잠정적인 가정으로 생각할 수 있다.
- 앞에서 배운 회귀모형의 예를 들어 설명해보자.

$$y = \beta_0 + \beta_1 x + \epsilon$$

- 여기서 우리는 반응변수와 예측변수 사이에 선형관계(결정론적 모형)에 대한 가정을 하거나 혹은 오차항(확률론적 요소)에 대하여 정규성, 등분산성, 독립성 등에 대한 가정을 한다.
- 우리는 이러한 가정을 가설로 간주 할 수 있다. 예를 들면 반응변수와 예측변수는 정말 선형관계인지 여부는 알지 못하지만, 데이터를 통해서 이 가설이 맞는지 여부를 확인할 수 있다.

가설 검정의 절차

- **가설 검정은 다음과 같은 절차를 거쳐서 진행한다.**

1. 가설 설정
   1) 귀무가설: 현 상태에 대한 잠정적 가정
   2) 대립가설: 우리가 알고 싶은 것
2. 검정통계량
3. 검정통계량의 표본분포
4. 결론

가설 검정을 통한 분석

- **앞의 질문들을 가설 검정의 프레임에 맞춰서 귀무가설을 제시하면 다음과 같다.**

1. 영국에서 실업률은 지난 4분기 동안 변하지 않고 그대로였다.
2. 스타틴은 기저질환이 있는 중년 남자의 심장마비나 뇌졸중 위험을 감소시키지 않는다.
3. 아버지의 키를 통제하면 어머니의 키는 아들의 키에 영향을 주지 않는다.
4. 힉스 입자는 존재하지 않는다.

법정 시스템 vs 통계적 가설 검정

- 법정 시스템은 통계적 가설 검정과 매우 흡사하다.
- 무죄 추정의 원칙은 귀무가설과 동일하다고 볼 수 있다.
- 법정에서 검사는 여러 가지 증거를 통해서 피고가 무죄라면 이러한 증거를 확보하기 힘들 것이라는 점을 강조하여 피고가 유죄라는 판결을 도출하는 것을 목표로 한다.
- 여기서 중요한 것은 법정에서의 결론은 다음 두 가지라는 것이다.
  - 유죄가 아니다(not guilty)
  - 유죄(guilty)

법정 시스템 vs 통계적 가설 검정

- 법정에서는 피고가 무죄라는 결론은 내리지 않는다. 즉 피고가 유죄(guilty)이거나 유죄라고 할만한 충분한 증거가 없다는 것이다(not guilty).
- 우리말과는 달리 영어로 표현 시 판결이 “innocent”가 아닌 “not guilty”임을 주목하자.
- 통계적 가설 검정에서도 귀무가설이 참이라는 결론은 내리지 않는다. 즉 대립가설이 참이거나(귀무가설 기각) 또는 대립가설이 참이라고 할 만한 충분한 증거가 없다(귀무가설을 기각할 수 없다)는 것이 결론이다.

오늘의 강의 요점

- 가설 검정의 절차
- 법정 시스템과 통계적 가설 검정

출처

#1 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

신뢰구간과 가설검정

Lab 9. 사례연구

이항분포

- 이항분포에서는 다음과 같은 명령어를 사용한다.
  - 확률밀도함수(probability density function)
    : dbinom(x, n, prob)
  - 누적밀도함수(cumulative distribution function)
    : pbinom(q, n, prob)
  - n개의 이항분포를 따르는 확률변수 생성: rbinom(n, prob)
  - Quantile function(누적밀도함수의 역함수)
    : qbinom(p, n, prob)

이항분포

동전을 10번 던졌을때 앞면이 5번 나올 확률과 100번 던졌을 때 앞면이 50번 나올 확률을 계산해보자.

```
dbinom(5,10,0.5)
```

```
## [1] 0.2460938
```

```
dbinom(50,100,0.5)
```

```
## [1] 0.07958924
```

동전을 10번 던졌을때 앞면이 4번 이하로 나올 확률과 100번 던졌을때 앞면이 40번 이하 나올 확률을 계산해 보자.

```
pbinom(4,10,0.5)
```

```
## [1] 0.3769531
```

```
pbinom(40,100,0.5)
```

```
## [1] 0.02844397
```

룰렛게임

[프렌치 룰렛]
(Wikipedia)

- 룰렛은 돌아가는 바퀴에 하나의 알을 놓고 빠른속도로 돌리다가 정지할 경우 알의 위치에 따라 상금을 받는 도박게임이다.

사진 출처 #1

룰렛게임

[프렌치 룰렛]
(Wikipedia)

- 왼쪽 그림은 프렌치 룰렛으로 0부터 36까지 총 37개의 숫자로 표시된 눈금이 있으며 미국 룰렛의 경우 0앞에 00을 추가한 총 38개의 숫자로 표시된 눈금이 있다.
- 미국식 룰렛의 경우 빨간색 18, 검은색 18, 녹색 2으로 색상이 구성되어 있다.

사진 출처 #1

룰렛게임

- 강원랜드에서 룰렛게임을 설치하는 것을 고려하고 있다고 가정하자. 룰렛게임의 수익이 어느정도 인지 예측하기 위해 여러분들에게 컨설팅을 의뢰한 경우를 생각해 보자.
- 룰렛게임에는 여러가지 종류가 있지만 단순화하기 위해 고객은 구슬이 빨간색, 혹은 검은색에 놓여있는지 여부만 베팅을 하고 고객이 이길 경우 1,000원을 주고 카지노가 이길 경우 1,000원을 받는다고 가정하자.
- 고객이 1000명일 경우 강원랜드의 수익이 얼마가 될까?

룰렛게임

먼저 s를 강원랜드의 수입이라고 가정하자. 1000명에 대해서 나오는 수입의 분포를 총 B=10000번 시뮬레이션을 통해 알아보자

```
n <- 1000
B <- 10000
roulette_winnings <- function(n){
  X <- sample(c(-1,1), n, replace = TRUE, prob=c(9/19, 10/19))
  sum(X)
}
S <- replicate(B, roulette_winnings(n))
```

강원랜드가 돈을 잃을 확률은?

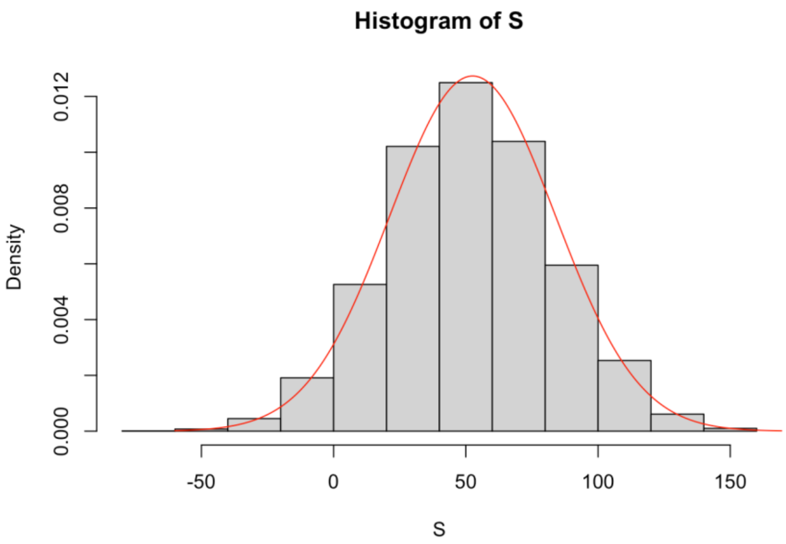
```
mean(S<0)
```

```
## [1] 0.0429
```

이항분포의 정규근사

이항분포의 정규분포 근사

```
hist(S, freq=FALSE)
x<-seq(-60, 200, by=1)
lines(x,dnorm(x, mean=mean(S),sd=sd(S)),col="red")
```



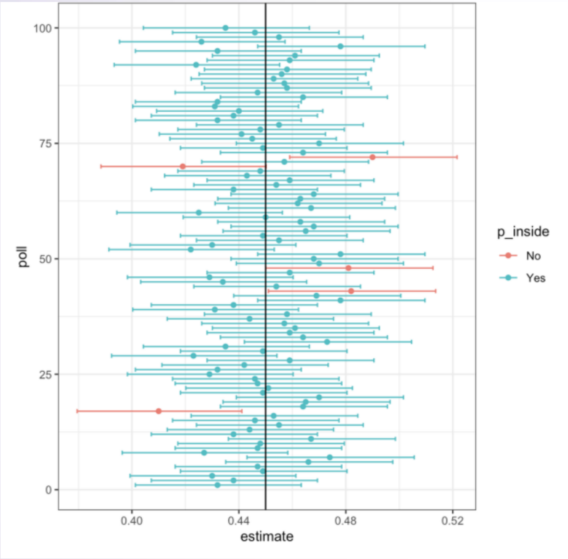
신뢰구간

- 시뮬레이션을 통해 신뢰구간의 개념을 설명해보자. 아래 코드를 실행하기 위해 R library dplyr을 사용하였다.

```
N <- 1000
B <- 10000
p <-0.45
inside <- replicate(B, {
  x <- sample(c(0,1), size = N, replace = TRUE, prob = c(1-p, p))
  x_hat <- mean(x)
  se_hat <- sqrt(x_hat * (1 - x_hat) / N)
  between(p, x_hat - 1.96 * se_hat, x_hat + 1.96 * se_hat)
})
mean(inside)
```

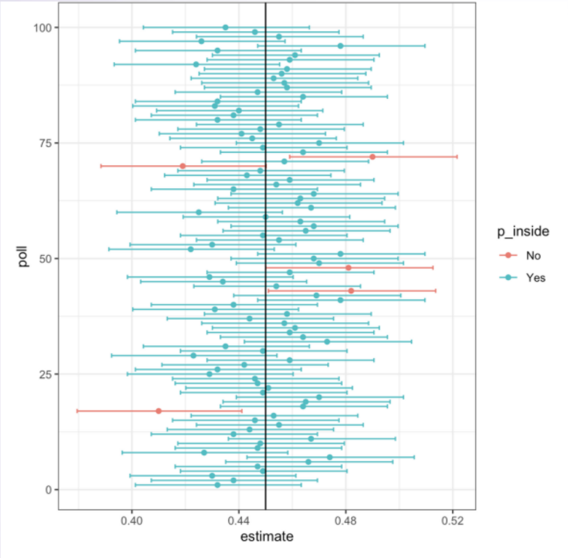
```
## [1] 0.9493
```

신뢰구간



- 왼쪽 그림은 시뮬레이션의 결과를 시각화한 것이다.
- 빨간색으로 표시된 신뢰구간이 실제 참값을 포함하지 않은 경우이다.

신뢰구간



- 신뢰구간을 생성하는데 사용된 표본들의 크기가 같기때문에 신뢰구간의 길이도 똑같다. 하지만 시뮬레이션에서 표본의 크기를 다르게 할 수 있으며 그렇게 하더라도 신뢰구간 중 참값을 포함하는 비율은 95% 정도로 수렴한다.

오늘의 강의 요점

- 이항분포
- 신뢰구간
  - 강원랜드 예제와 신뢰구간 예제의 코드는 Irizarry (2020). Introduction to Data Science(https://rafalab.github.io/dsbook/)를 참조하였다.

출처

#1 Wikipedia https://ko.wikipedia.org/wiki/%EB%A3%B0%EB%A0%9B