데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

확률로 풀어보는 불확실성

1. 붓스트랩(Bootstrap)

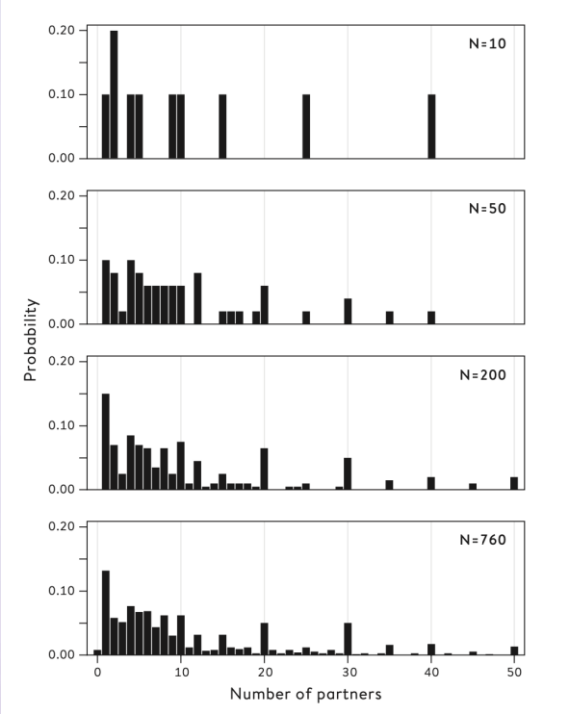
성관계 상대 수

- 앞에서 나온 영국의 성생활 실태조사에 대해서 다시 살펴보자.
- 이 조사의 대상자는 35~44세의 여성 1,215명과 남성 806명이었으며, 성관계 상대 수에 대한 설문 결과에 대한 요약은 다음과 같았다.
- 평균: 남성 14.3명, 여성 8.5명
- 중앙값: 남성 8명, 여성 5명

성관계 상대 수

- 이 조사는 임의추출을 바탕으로 해서 연구모집단(35~44세의 여성 1,215명과 남성 806명)이 목표모집단(영국 성인남녀)에 부합한다는 추론은 합리적이다.
- 그렇다면 앞의 요약 통계량은 실제 영국 성인 남녀를 대상으로 조사한 평균과 중앙값과 비교한다면 얼마나 차이가 있을까?

표본의 분포

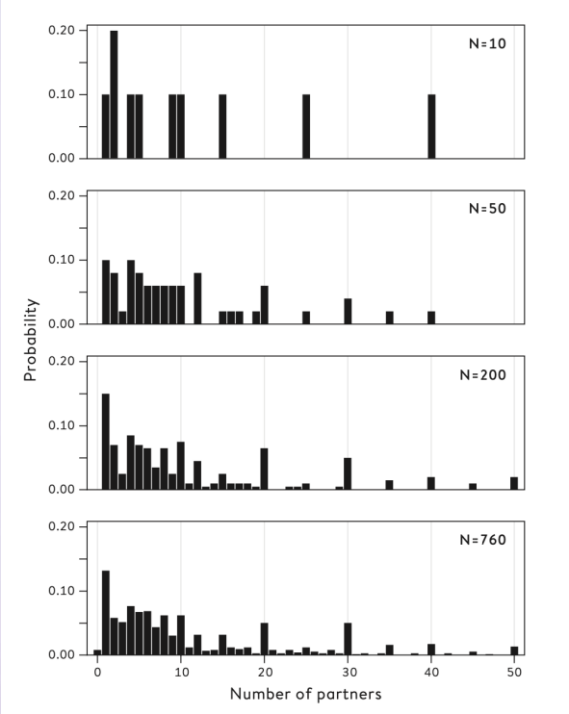


[Resampling Distribution]
(The Art of Statistics, p193)

- 이 질문에 답변하기 위해 먼저 최대 50명의 파트너를 보고한 남성 760명을 모집단으로 가정하자. 왼쪽 그림 맨 아래에 모집단에 대한 히스토그램이 있다.

기타 출처 #1

표본의 분포



[Resampling Distribution]
(The Art of Statistics, p193)

- 이 모집단으로부터 임의로 10, 50, 200명을 뽑은 후 히스토그램을 각각 그린 결과가 왼쪽 그림에 있다.
- 표본의 크기가 커질수록 표본의 히스토그램이 모집단의 히스토그램과 비슷해짐을 알 수 있다.

기타 출처 #1

표본의 분포

- 각 표본의 요약치를 구해서 모평균의 요약치와 비교하면 다음과 같다.

표본수	성관계 파트너 수의 평균	성관계 파트너 수의 중앙값
10	8.3	9
50	10.5	7.5
200	12.2	8
760	11.4	7

[35~44세 남성에 관한 성생활 실태조사 결과와 resampling을 통한 요약치]
(The Art of Statistics, p194)

기타 출처 #2

통계량의 분포(표본분포: Sampling Distribution)

- 여기서 주목할 점은 같은 크기의 표본을 다시 뽑는다면 요약치의 값들이 변할 것이라는 점이다.
- 즉 표본의 크기가 커질수록 표본 요약치(통계량: statistics)의 값들이 모집단의 요약치(모수: parameter)에 가까워지는 건 사실이지만 대신 각 표본 요약치의 변동성은 어떻게 설명할 수 있을까?

통계량의 분포(표본분포: Sampling Distribution)

- 이 요약치의 변동성(즉 통계량의 분산)을 알아내기 위해서는 모집단에서 표본크기 N이 같은 여러 개의 표본을 생성하면 된다. 예를 들면 *N=10*인 표본을 100개 구한 후 각각의 표본에서 평균을 구한다면 우리는 100개 평균을 가지게 된다.
- 위의 100개의 평균의 히스토그램을 그린다면 *N=10*인 경우 표본평균의 변동성을 알 수 있다.

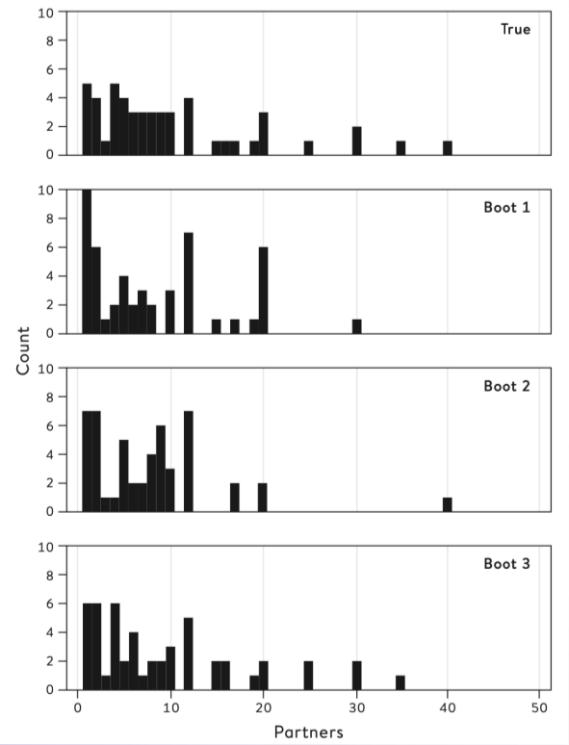
통계량의 분포(표본분포: Sampling Distribution)

- 또 다른 방법은 모집단에 관한 추가 정보, 예를 들면 모집단의 변동성(즉 모집단의 분산)을 알아야 한다. 하지만 모집단의 요약치를 알지 못하면서 그 외의 추가 정보를 아는 경우는 거의 없다. 이 경우 통계이론을 바탕으로 모집단에 관한 추가정보를 알아낼 수 있다.
- 하지만 복잡한 통계이론을 알지 못하는 경우 쉽게 통계량의 변동성을 알아내는 방법은 없을까?

통계량의 분포(표본분포: Sampling Distribution)

- 모집단과 표본의 모습이 비슷하다는 가정하에 다음과 같은 방법을 생각할 수 있다.
- 표본을 모집단이라 생각하고 표본에서 복원추출(즉 하나의 데이터를 뽑고 다시 그 데이터를 원래대로 집어넣어서 같은 자료가 여러번 뽑힐 수 있는 표본추출방법)로 같은 크기의 재표본(resample)을 뽑는다. 이 방법을 사용할 경우 원래 표본에서 같은 크기의 재표본을 여러 개 뽑을 수 있다.

통계량의 분포(표본분포: Sampling Distribution)

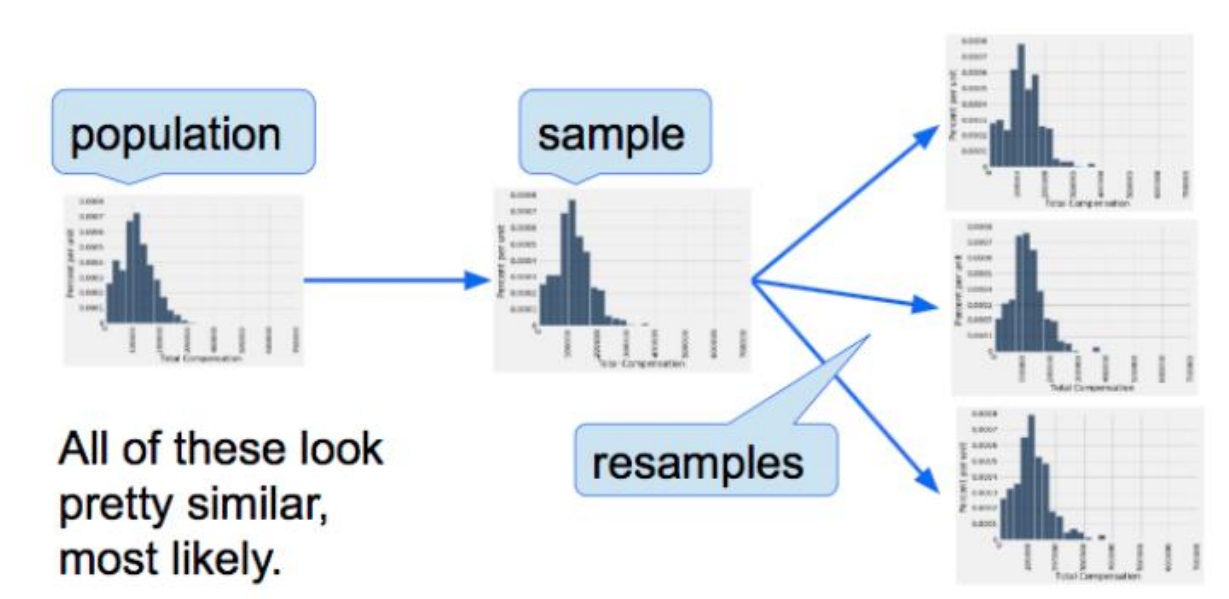


[붓스트랩 분포]
(The Art of Statistics, p194)

- 왼쪽 그림의 맨 위쪽 그림은 760명에게서 뽑은 크기가 50인 표본의 히스토그램이다.
- 아래 3개의 그림은 맨 위 그림의 크기가 50인 표본으로부터 3번의 복원추출한 재표본을 히스토그램으로 제시하였다.
- 아래 3개의 그림에 나오는 재표본들의 평균은 8.4, 9.7, 9.8이며 맨 위의 표본의 평균은 10.5이다.

기타 출처 #3

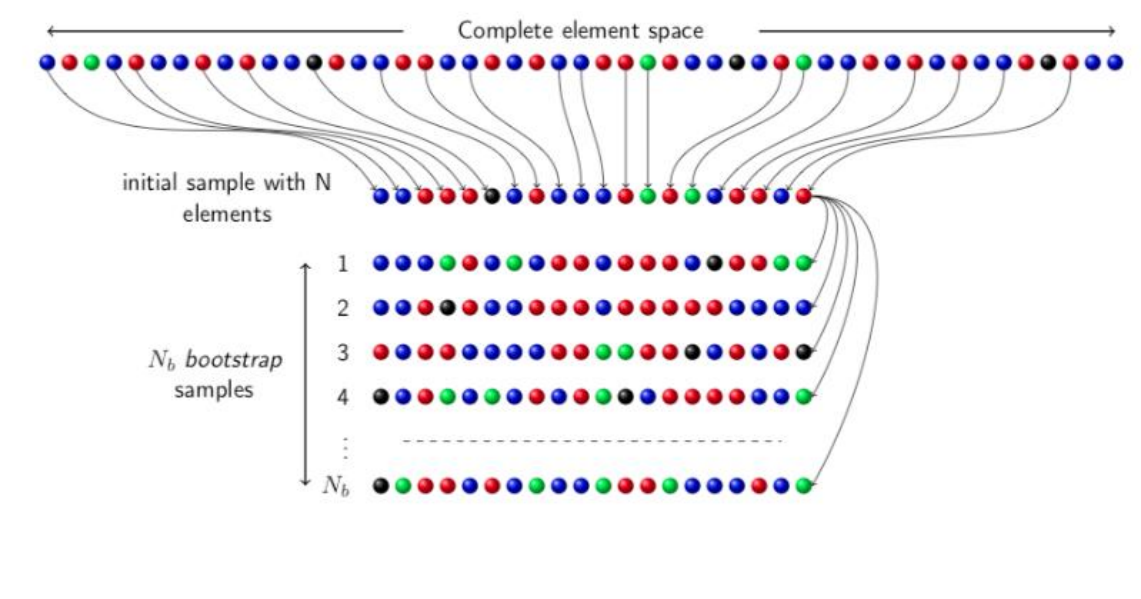
붓스트랩(Bootstrap)



[붓스트랩의 원리]
(Computational and Inferential Thinking)

기타 출처 #4

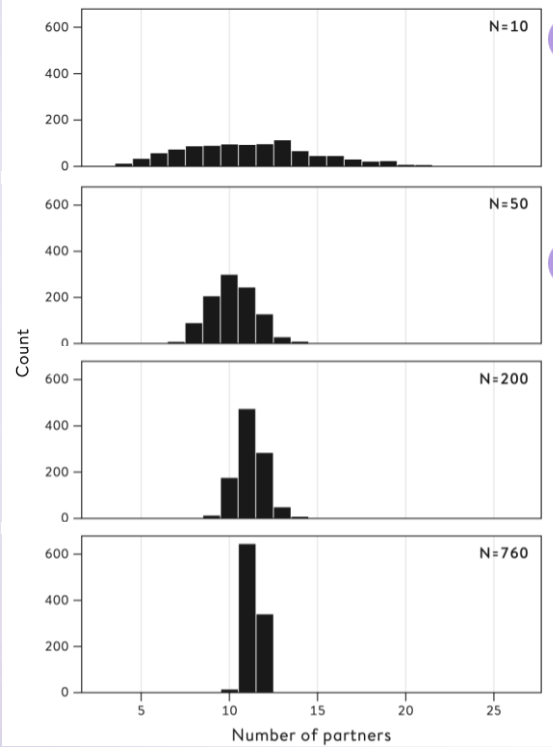
붓스트랩(Bootstrap)



[붓스트랩의 원리]
(TEXample.net)

기타 출처 #5

붓스트랩(Bootstrap)



[1,000개의 붓스트랩 표본을 이용한 표본평균의 분포]
(The Art of Statistics, p198)

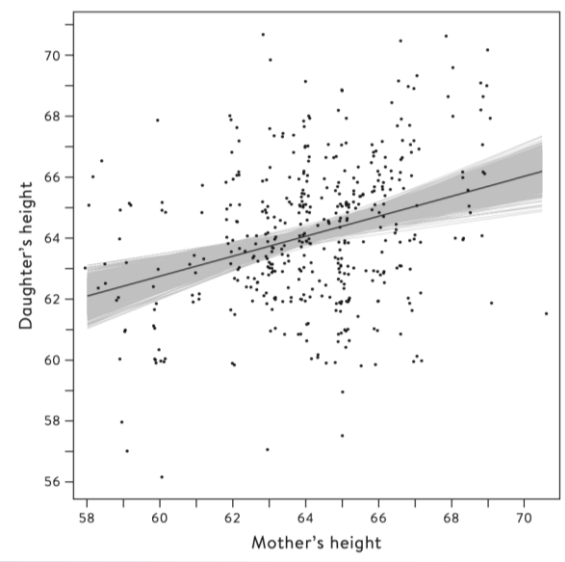
- 왼쪽 그림은 크기가 10, 50, 200, 760인 재표본 1,000개의 평균값의 분포를 나타낸다.
- 붓스트랩 95% 불확실성 구간은 재표본 평균의 95%가 포함되는 범위를 의미한다.

표본 수	성관계 파트너 수의 평균	붓스트랩 95% 불확실성 구간
10	8.3	(5.3, 11.5)
50	10.5	(7.7, 13.8)
200	12.2	(10.5, 13.8)
760	11.4	(10.5, 12.2)

[35~44세 남성에 관한 성생활 실태조사 결과와 resampling을 통한 요약치]
(The Art of Statistics, p194)

기타 출처 #2, 6

붓스트랩을 이용한 회귀분석

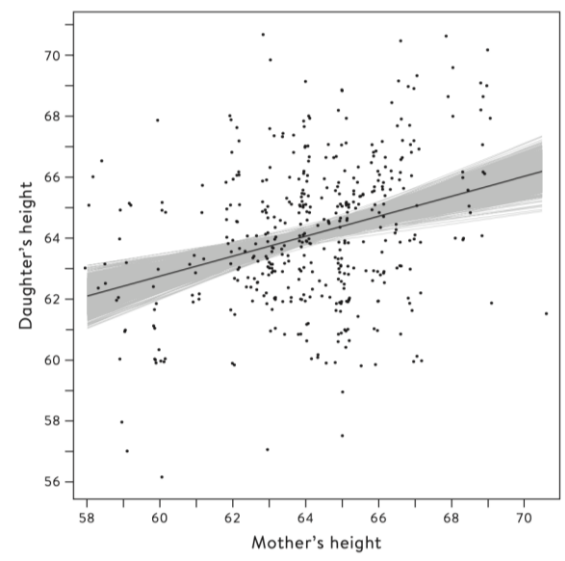


[Galton의 모녀 키자료에 적합한 회귀직선(검은 직선)과 20개의 붓스트랩 표본을 이용하여 생성한 회귀직선(회색)]
(The Art of Statistics, p202)

- 골턴의 자료에서 어머니의 키를 예측변수로 사용하고 딸의 키를 반응변수로 사용하여 회귀분석을 적합 시키면 기울기 0.33인 회귀직선식을 구할 수 있다.
- 433개의 쌍으로 이루어진 모녀 키 자료를 20번의 비복원추출을 통해서 20개의 재표본을 구하자.

기타 출처 #7

붓스트랩을 이용한 회귀분석



[Galton의 모녀 키자료에 적합한 회귀직선(검은 직선)과 20개의 붓스트랩 표본을 이용하여 생성한 회귀직선(회색)]
(The Art of Statistics, p202)

- 각각의 재표본에서 회귀분석을 적합하면 20개의 회귀직선의 기울기를 구할 수 있다.
- 이 경우 기울기의 95% 불확실성 구간은 0.22에서 0.44이다.

기타 출처 #7

오늘의 강의 요점

- 통계량과 표본의 분포
- 재표본을 이용한 통계량의 변동성 파악
- 붓스트랩 불확실성 구간

출처

#1~3 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

#4 Computational and Inferential Thinking
https://www.inferentialthinking.com/chapters/13/2/Bootstrap.html

#5 TEXample.net https://texample.net/media/tikz/examples/PDF/bootstrap-resampling.pdf

#6~7 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

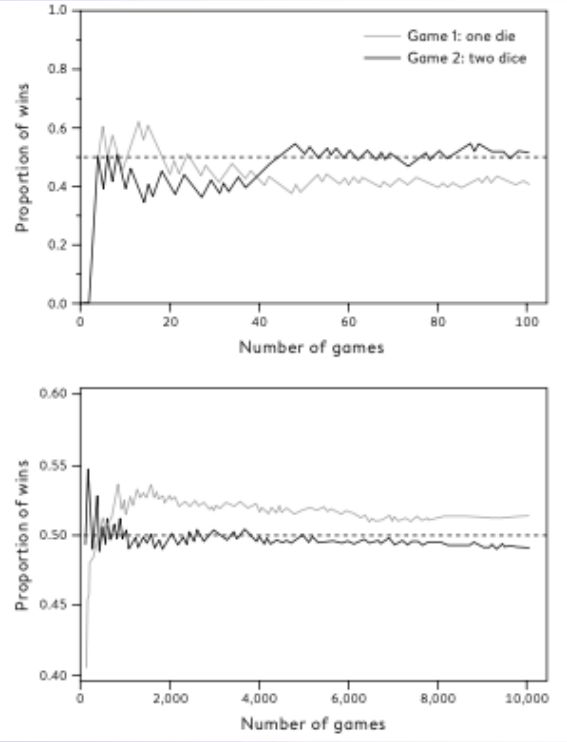
확률로 풀어보는 불확실성

2. 확률의 기원과 법칙

슈발리에 드 메레(Chevalier de Méré)의 질문

- **1650년대 프랑스 작가 슈발리에 드 메레는 다음과 같은 도박 문제를 고심하고 있었다.**
  - 게임 1: 최대 4번까지 공정한 주사위를 한 개 던지는데 6이 나오면 이긴다.
  - 게임 2: 최대 24번까지 공정한 주사위를 두 개 던지는데 둘다 6이 나오면 이긴다.
- **어느 게임이 더 유리한 게임일까?**

슈발리에 드 메레(Chevalier de Méré)의 질문

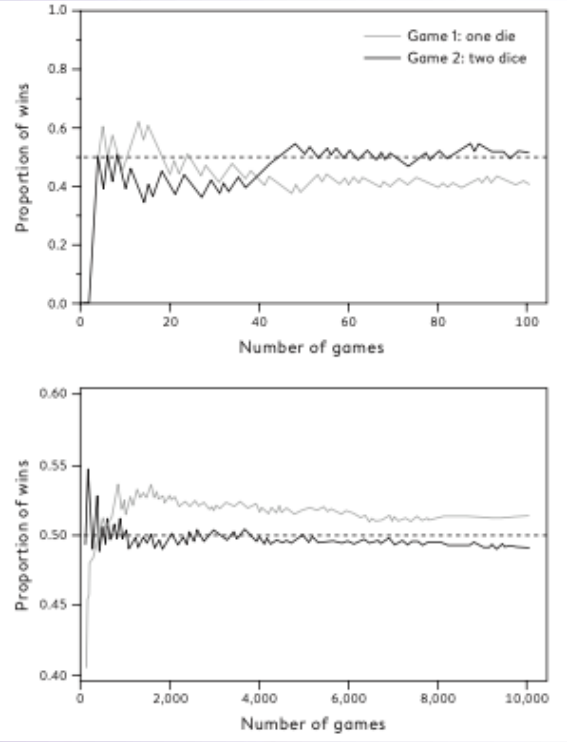


[게임 1과 게임 2의 컴퓨터 시뮬레이션 결과]
(The Art of Statistics, p206)

- 컴퓨터 시뮬레이션을 통해 실제 어느 게임이 더 유리한지 알아보자.
- 왼쪽 위 그림은 처음 100번 시뮬레이션 결과 게임 2가 유리하다는 것을 보여준다.

기타 출처 #1

슈발리에 드 메레(Chevalier de Méré)의 질문

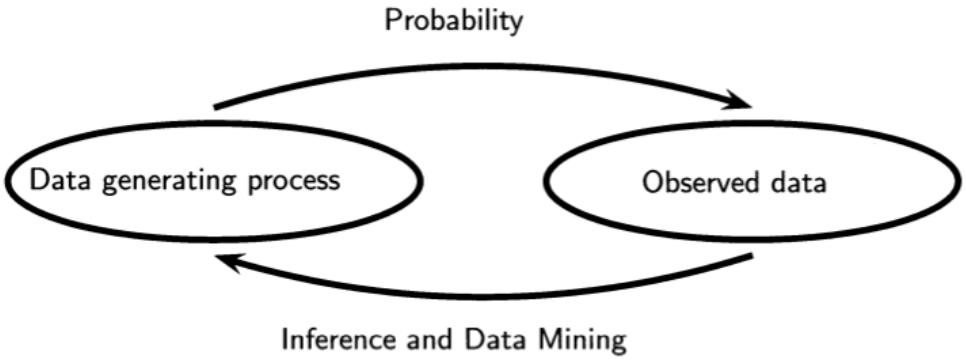


[게임 1과 게임 2의 컴퓨터 시뮬레이션 결과]
(The Art of Statistics, p206)

- 하지만 400번 정도를 기점으로 게임 1의 승률이 더 높아지며 장기적으로 게임 1이 이길 확률은 약 52%이고 게임 2는 약 49%이다.
- 이 문제에 대한 해답을 구하기 위해 드 메레가 도움을 청한 사람은 파스칼이었고 파스칼의 그의 친구 페르마와 같이 확률론의 기반을 다지게 되었다.

기타 출처 #1

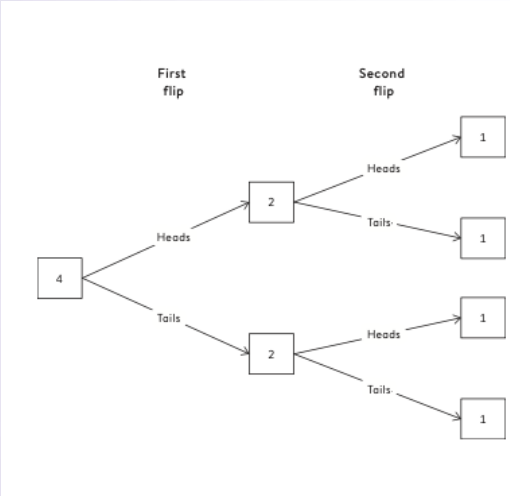
확률과 통계



[확률과 통계적 추론]
(All of Statistics p. ix)

기타 출처 #2

기대돗수나무(Expected Frequency Tree)

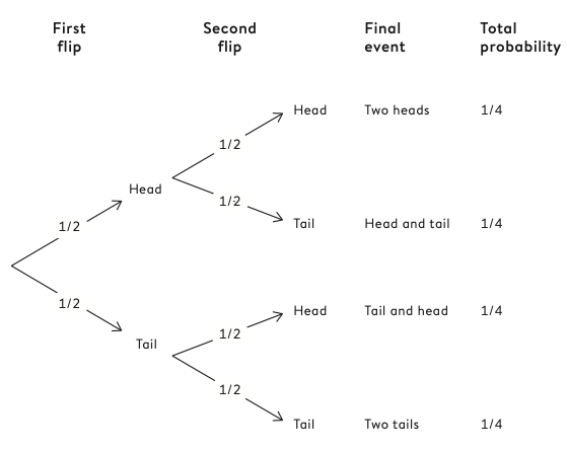


[기대돗수나무]
(The Art of Statistics, p211)

- 2012년 영국의 국회의원 97명에게 “동전을 두 번 던졌을 때 앞면이 2번 나올 확률이 얼마입니까?”라는 질문을 한 결과 절반 이상인 60명이 정답을 말하지 못했다!
- 이 확률을 쉽게 계산하기 위해 왼쪽 그림과 같이 가능한 모든 경우를 고려하는 기대돗수나무를 생각할 수 있다.

기타 출처 #3

확률나무(Probability Tree)



[확률나무]
(The Art of Statistics, p212)

- **기대돗수나무에서 각 가지에 나오는 경우의 횟수를 비율로 나타내면 왼쪽 그림과 같은 확률나무로 바꿀 수 있다. 확률나무를 통해 다음과 같은 확률 법칙이 성립함을 알 수 있다.**

1. 확률은 0과 1사이다.
2. 여사건 법칙
3. 덧셈법칙
4. 곱셈법칙

기타 출처 #4

확률의 법칙

- 사건 A가 일어날 확률을 $Pr(A)$라 하자. 첫 번째 법칙은

  $0 \leq Pr(A) \leq 1$.

- 여사건의 법칙은 $Pr(A^c) = 1 - Pr(A)$ 를 의미한다. 여기서 A^c는 A의 여사건으로 사건 A가 일어나지 않는 경우를 의미한다.

확률의 법칙

- 덧셈법칙은 2개의 사건이 동시에 일어날 수 없는 경우 두 사건을 관측할 확률은 각각의 확률의 합으로 나타난다는 것을 의미한다. 즉 만약 사건 A와 B가 서로 동시에 일어날 수 없다면

$$Pr(A\ or\ B) = Pr(A) + Pr(B).$$

- 곱셈법칙은 두 개의 사건이 서로 독립이라면(서로 영향을 주지 않는다면) 두 사건이 동시에 일어날 확률은 각각의 확률의 곱이다. 즉 만약 사건 A와 B가 독립이면

$$Pr(A\ and\ B) = Pr(A) \cdot Pr(B).$$

다시 슈발리에 드 메레의 질문…

- 게임 1: 최대 4번까지 공정한 주사위를 한 개 던지는데 6이 나오면 이긴다.
- 2번째 법칙 여사건의 확률을 이용하여 게임 1에서 승률은

  1- (4번 모두 6 이외의 숫자가 나올 확률)임을 알 수 있다.
- 여기서 각각 주사위를 던질 때 나오는 눈의 개수는 서로 영향을 주지 않으므로 곱셈의 법칙을 사용하여 (4번 모두 6 이외의 숫자가 나올 확률) = (모두 6 이외의 숫자가 나올 확률)^4
- 따라서 게임 1의 승률은 $1-\left(\frac{5}{6}\right)^4 = 0.5177$

다시 슈발리에 드 메레의 질문…

- 게임 2: 최대 24번까지 공정한 주사위를 두 개 던지는데 둘 다 6이 나오면 이긴다.
- 비슷하게 여사건의 법칙을 사용해서 게임 2의 승률은

  1 – (24번 모두 둘다 6이 나오지 않을 확률)임을 알 수 있다.
- 여기서 둘 다 6이 나오지 않은 확률은 $1 - \frac{1}{6} \cdot \frac{1}{6} = \frac{35}{36}$
- 다시 곱셈의 법칙을 사용하면 게임 2의 승률은

  $1 - \left(\frac{35}{36}\right)^{24} = 0.4914$

오늘의 강의 요점

- **확률나무**
- **확률의 4가지 법칙**

1. 확률은 0과 1사이다.
2. 여사건 법칙
3. 덧셈법칙
4. 곱셈법칙

출처

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

#2 Larry Wasserman, (2004), All of Statistics, Springer

#3~4 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

확률로 풀어보는 불확실성

3. 조건부 확률과 확률변수

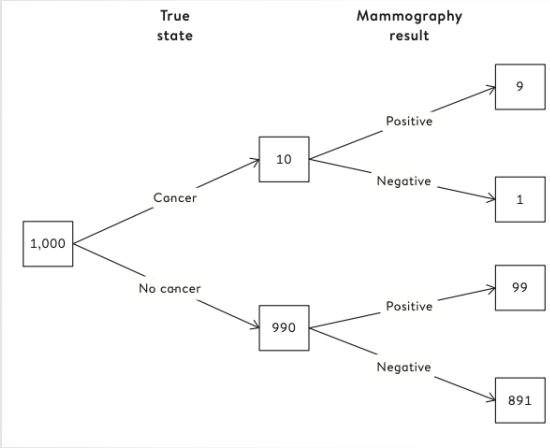
유방촬영술(맘모그램) 언제부터 받아야 하나요?

- 한국은 만 40세부터 2년마다 유방암 진단을 위해 유방촬영술 받기를 권장한다.
- 유방암 검사 시 유방촬영술의 결과는 약 90% 정확하다. 즉 암이 있는 경우와 없는 경우를 올바르게 분류하는 것이 각각 90% 정도 정확하다는 의미이다.
- 검사를 받은 여성 중 1%가 실제 암이 있다고 가정하자.(우리나라의 유방암 환자 수는 2010년 기준 10만 명당 67.2명이다.) 다음 질문에 답하여라.

  → 유방촬영술 결과가 양성이 나올 확률은?

  → 유방촬영술 결과가 양성일 때 실제 유방암에 걸려 있을 확률은?

유방촬영술(맘모그램) 언제부터 받아야 하나요?



[1,000명의 유방암 검사결과]
(The Art of Statistics, p215)

- 첫 번째 질문의 경우 왼쪽 기대 돗수나무에서 1,000명 중 검사결과가 양성인 것은 마지막 마디에서 위에서 첫 번째와 세 번째 경우로 합계는 9+99=108명이다. 즉 검사결과가 양성일 확률은 108/1,000=0.11이다.
- 검사 결과가 양성인 경우 실제 암환자일 경우는 9/108=0.08이다. 즉 8%에 불과하다!

기타 출처 #1

검사의 오류(Prosecutor's Fallacy)

- 암에 걸렸을 때 진단 결과가 양성일 확률은 90%이지만 진단 결과가 양성일 때 암에 걸려있을 확률은 8%에 불과하다.
- 하지만 현실에서는 이 2가지 확률을 혼동하는 경우가 많은데 이런 유형의 혼동을 검사의 오류라고 한다.
- 법정에서 검사가 DNA 증거를 논할 때 "피고가 결백하다면 피고의 DNA가 현장에서 발견된 DNA가 일치할 확률은 10억 분의 1이다."라고 얘기해야 할 사항을 "DNA 증거를 고려할 때 피고가 결백할 확률은 10억 분의 1이다."라고 얘기하는 것과 같다.

도대체 확률이란 무엇인가?

- 고전적 확률(classical probability): 주사위 던지기와 동전 던지기와 같이 모든 결과가 나올 확률이 동일하다는 전제하에 특정 사건이 나올 확률을 계산한다.
- 나열 확률(enumerative probability): 모든 가능한 경우를 생각하고 그중 내가 관심이 있는 사건이 일어나는 비율을 생각한다. 예를 들면 검은색 공 3개와 빨간 공 2개가 들어 있는 상자에서 공을 하나 꺼낼 때 빨간 공이 나올 확률은 2/5이다.

도대체 확률이란 무엇인가?

- 장기 빈도 확률(long-run frequency probability): 동일한 사건이 반복적으로 일어날 때 발생하는 비율을 의미한다. 하지만 모든 사건이 반복적으로 일어나지는 않는다.
- 성향(propensity): 특정 사건이 일어날 진짜 가능성을 의미한다. 하지만 본인이 전지전능하지 않은 경우 이 “성향”을 알아내는 것은 (거의) 불가능하다.
- 주관적 확률(subjective probability): 내가 월드컵에서 한국이 4강까지 진출할 경우 10만 원을 주는 도박에 만 원을 걸었다고 하자. 이 경우 나의 주관적 확률은 0.1이다

확률변수란?

- **확률은 다음과 같은 상황을 설명할 때 필요하다.**
  - 데이터가 컴퓨터(혹은 난수표)에 의해서 임의로 생성된다고 할 때
  - 이미 존재하는 데이터를 임의로 선택하고자 할 때
  - 임의성은 없지만 마치 데이터를 임의로 생성되었다고 가정할 때

확률변수

- 동전을 던져서 앞면이 나오면 1이라 기록하고 뒷면이 나오면 0이라고 하자.
- 이처럼 특정 결과를 숫자와 연관시키는 규칙을 만들 수 있다. 이러한 규칙을 확률변수라고 한다.
- 구체적으로 우리가 특정 사건(동전던지기)을 생각하고 특정 사건의 가능한 모든 결과물(앞면, 뒷면)의 집합을 표본공간(sample space)라고 한다.
- 확률변수는 이러한 표본공간에 속한 각각의 원소에 특정 숫자를 대입한 값이라고 생각하면 된다.

확률분포

- 확률분포는 확률변수가 특정 값을 가질 확률을 나타낸다. 예를 들면 동전던기기의 경우

$$Pr(X = 0) = Pr(X = 1) = 0.5$$

- 확률의 법칙에 따라서 확률분포에 나온 값을 모두 합할 경우 1이 되어야 한다.
- 동전던지기와 같이 2가지 가능한 결과물을 가진 확률변수의 확률분포를 베르누이 분포라고 한다.

하루에 살인사건이 7번씩이나 일어났다고?

- 2013년 4월부터 2016년 3월까지 1,095일 동안 영국과 웨일스에서 총 1,545건의 살인사건이 있었다. 즉 하루평균 1.41건의 살인사건이 발생하였다.
- 하루당 살인사건이 제일 많이 발생한 건수는 6건이었다.
- 이렇게 시공간이 정해진 상황에서 일어나는 사건의 횟수를 확률변수로 생각할 수 있다.

하루에 살인사건이 7번씩이나 일어났다고?

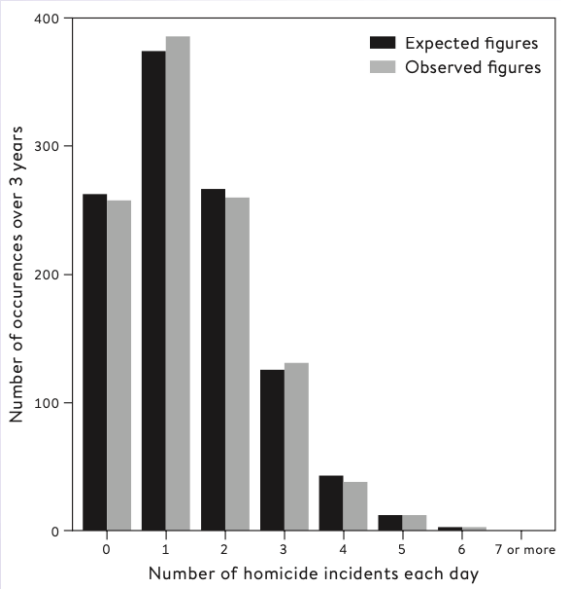
- 이러한 확률변수가 따르는 분포를 포아송 분포(Poisson distribution)이라고 한다.
- 이 분포는 특정 축구팀의 경기당 득점 수나 매년 말에 채여 사망하는 프러시아 장교의 수에 이르기까지 광범위한 사건을 모형화하는데 사용된다.
- 포아송 분포는 어떤 사건이 일어날 기회는 엄청나게 많지만 각 사건이 일어날 가능성이 아주 적은 경우에 사용된다.

하루에 살인사건이 7번씩이나 일어났다고?

- 정규분포의 경우 모수가 2개(평균과 분산)이지만 포아송 분포의 경우 모수는 평균 하나만 있다. 즉 평균만 알면 이 분포의 형태를 알 수 있으며 평균이 m일 경우 각 사건의 일어날 횟수 x의 확률은 다음과 같이 주어진다.

$$Pr(X = x) = \frac{e^{-m} m^{x}}{x!}$$

하루에 살인사건이 7번씩이나 일어났다고?



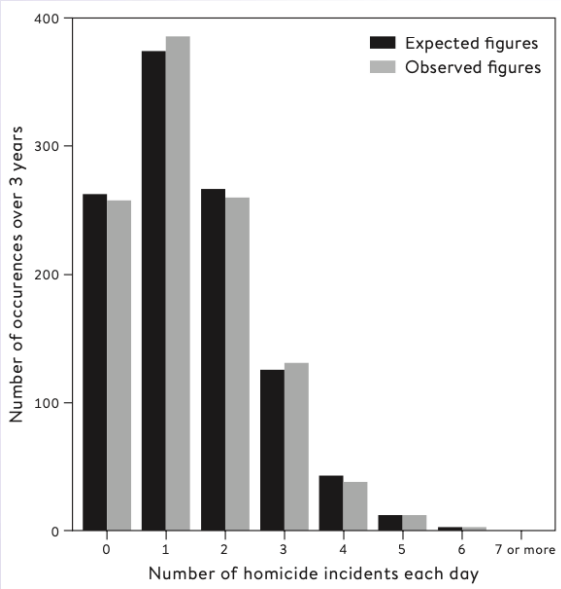
[2013년부터 2016년까지 영국과 웨일스에서 기록된 일일 살인 사건의 횟수와 기대 횟수]
(The Art of Statistics, p225)

- 왼쪽 그림은 일일 살인사건의 횟수에 대한 자료와 포아송 분포를 바탕으로 각 일일 살인사건 횟수를 예측한 값을 비교한 것이다.
- 포아송 분포를 이용할 경우 하루 7건 이상의 살인사건이 일어날 확률은 다음과 같다.

$$\sum_{x=7}^{\infty} \frac{e^{-m}m^x}{x!} = 1 - \sum_{x=0}^{6} \frac{e^{-m}m^x}{x!} = 0.0007$$

기타 출처 #2

하루에 살인사건이 7번씩이나 일어났다고?



[2013년부터 2016년까지 영국과 웨일스에서 기록된 일일 살인 사건의 횟수와 기대 횟수]
(The Art of Statistics, p225)

- 즉 평균 1,535일, 혹은 약 4년에 한 번 정도 이런 일이 일어날 수 있음을 의미한다.

기타 출처 #2

오늘의 강의 요점

- 조건부 확률
- 확률변수와 확률분포
- 포아송분포

출처

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

확률로 풀어보는 불확실성

Lab 8. 붓스트랩과 다양한 확률분포

미국 동전 페니가 발행된 평균 년도는 얼마일까?

- 2019년도에 미국에서 사용되는 동전 페니들의 평균 발행 년도는 얼마인지 알아보기 위해 통계학자 앨버트 킴은 본인이 살고있는 동네의 인근 은행에 가서 페니 50개를 받아왔다.

[앨버트 킴이 인근 은행에서 페니를 받아오는 모습]

[은행에서 받아온 페니들의 발행연도]

기타 출처 #1, 2

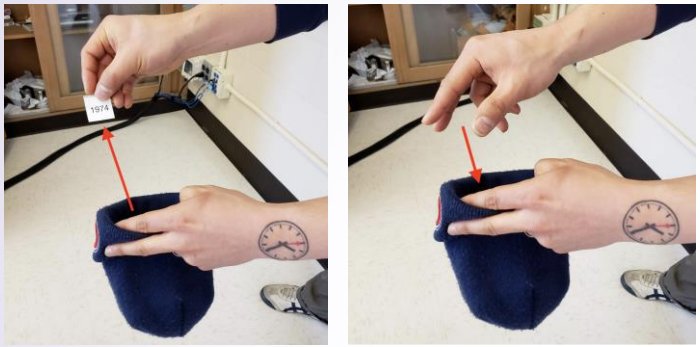
Resampling

[페니 발행년도를 적은 종이]

[번호가 적힌 종이를 모자에 집어넣자!]

기타 출처 #3, 4

Resampling



[Step 1. 모자에서 종이를 뽑고 년도를 기록하자 Step 2. 모자에 종이를 다시 넣자]

[번호가 적힌 종이를 모자에 집어넣자!]

- Step 1과 2를 50번 반복한 후 아래의 그림과 같은 결과를 얻는다.

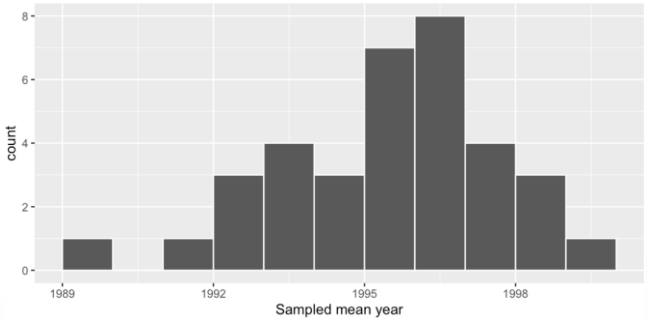
기타 출처 #2, 5, 6

친구에게 부탁하기

- 앞의 과정을 35명의 친구들에게 반복하도록 하자. 즉 앨버트가 가져온 동전에서 붓스트랩 표본을 하나씩 만드는 것이다.

Arianna	Artemis	Bea	Camryn	Cassandra
1988	2018	2016	2002	2015
2002	1988	1971	1997	1976
2015	1999	1986	2002	2015
1998	2015	2002	2013	1981
1979	1962	1992	1997	1988
1971	2004	1976	1979	1985
1971	2018	2015	2018	1979
2015	1988	1985	1971	1971
1988	2013	1976	1998	1978
1979	1988	1999	1996	1979
1982	2008	2013	1999	1986
2004	1983	1997	1983	1974

[앨버트 친구들의 붓스트랩 표본의 값]



[각 표본에서 얻은 평균 35개를 이용해서 그림 히스토그램]

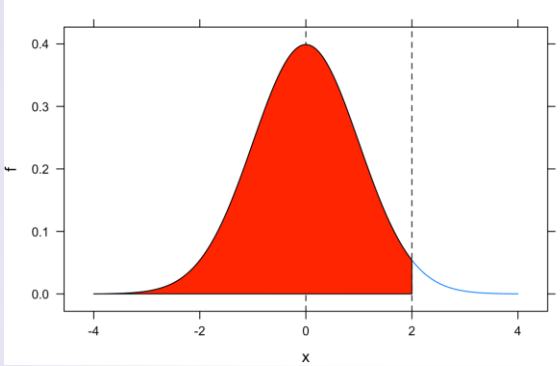
기타 출처 #7, 8

정규분포

- **R을 이용하여 다양한 확률분포의 모양, percentile, quantile을 파악하거나 그 확률분포를 따르는 확률변수를 생성할 수 있다.**
- **가장 많이 사용하는 확률분포인 정규분포를 사용하여 설명해보자. 평균이 mean이고 표준편차가 sd일 경우**

  - 확률밀도함수(probability density function)
    : dnorm(x, mean, sd)
  - 누적밀도함수(cumulative distribution function)
    : pnorm(q, mean, sd)
  - n개의 정규분포를 따르는 확률변수 생성: rnorm(n, mean, sd)
  - Quantile function(누적밀도함수의 역함수)
    : qnorm(p, mean, sd)

정규분포

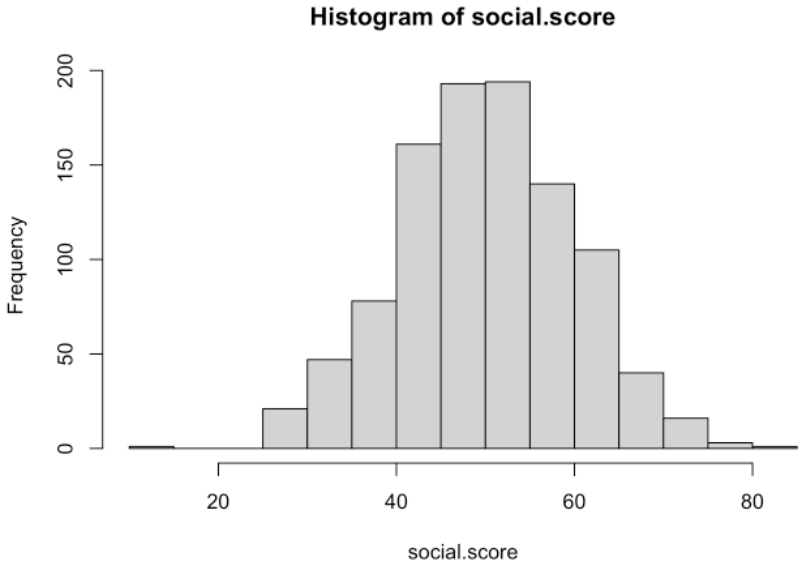


- 위의 그림에서 빨간색이 나타내는 면적이 pnorm(2,0,1)=$\int_0^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = 0.9772$
- qnorm(0.9772,0,1)=2
- dnorm(0,0,1)=$e^{-0/2}/\sqrt{2\pi}$ =0.3989
- 표준정규분포에서 1,000개를 생성: rnorm(1000,0,1)

정규분포

수능 사탐의 표준점수는 평균이 50이고 표준편차가 10인 정규분포를 따른다고 한다. 이 분포에서 1000개의 표본을 생성하고 히스토그램을 그려보자.

```
social.score <-rnorm(1000,mean=50, sd=10)
hist(social.score)
```



정규분포

사탐에서 1등급 컷(즉 96 percentile)에 해당하는 점수는 얼마인가?

```
qnorm(0.96, mean=50, sd=10)
```

```
## [1] 67.50686
```

만약 내 표준점수가 64점이라면 내 점수는 몇등급에 해당하는가?

```
pnorm(64, mean=50,sd=10)
```

```
## [1] 0.9192433
```

4%와 11%사이이므로 2등급에 해당한다.

포아송분포

- **포아송분포에서는 다음과 같은 명령어를 사용한다.**
  - 확률밀도함수(probability density function)
    :dpois(x, lambda)
  - 누적밀도함수(cumulative distribution function)
    : ppois(q, lambda)
  - n개의 정규분포를 따르는 확률변수 생성: rpois(n, lambda)
  - Quantile function(누적밀도함수의 역함수)
    : qpois(p, lambda)

포아송분포

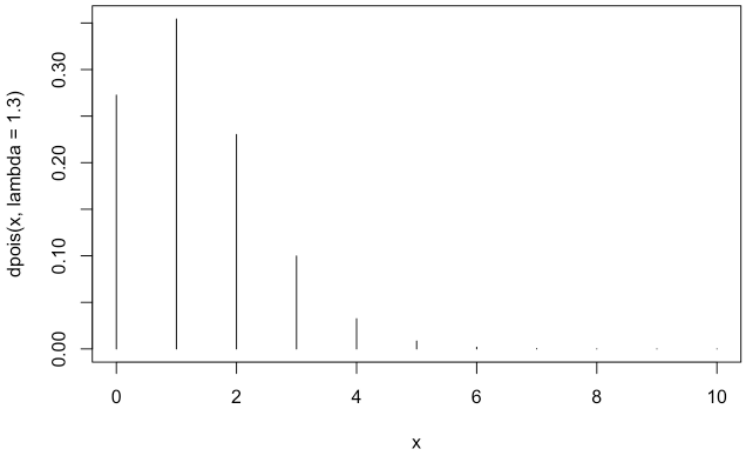
예린은 같은 회사 동료 정국과 사랑에 빠진다. 사랑하는 커플은 시간당 평균 1.3회 키스를 한다고 알려져 있다. 두사람이 한시간 동안 같이 있으면서 키스를 2회 이상할 확률은?

```
1-ppois(1,lambda=1.3)
```

```
## [1] 0.3731769
```

평균 1.3인 포아송 분포의 확률밀도 함수를 그려보자.

```
x<-c(0:10)
plot(x, dpois(x,lambda=1.3), type="h")
```



오늘의 강의 요점

- **붓스트랩**
- **여러 가지 확률분포**
  - 정규분포
  - 포아송분포

출처

#1~8 https://moderndive.com/8-confidence-intervals.html