Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

02

확률과 통계

경북대학교 배준현 교수
(joonion@knu.ac.kr)

- 확률: *probability*
  - 생각할 수 있는 모든 경우의 수 중에서
    - 우리가 관심을 갖는 경우의 수가 차지하는 비율
  - 로또복권의 1등 당첨 확률
    - 모든 경우의 수: ${}_{45}C_{6} = 8{,}145{,}060$, 관심 있는 경우의 수: 1
  - 반드시 한 가지 조건을 만족해야 함:
    - 1에서 45까지의 모든 숫자가 뽑힐 확률이 동일하다.

확률 용어:

- 시행: *trial*
  - 다양한 결과가 나올 수 있는 어떤 것을 실제로 하는 것. 예) 동전 던지기
- 표본공간: *sample space*
  - 시행의 결과로 나올 수 있는 가능한 모든 결과의 집합. 예) {앞, 뒤}
- 사건: *event*
  - 가능한 결과들 중 어떤 요구사항을 만족하는 것. 예) 앞면이 나옴
- 배반사건: *disjoint event*
  - 동시에 일어날 수 없는 두 사건. 예) 앞면/뒷면이 동시에 나옴
- 여사건: *complementary event*
  - 어떤 사건이 일어나지 않은 것. 예) 앞면이 나오지 않음

- **수학적 확률**: *mathematical* probability
  - 가능한 모든 경우 중에서 관심 있는 경우의 비율이 얼마냐?
  - 예) 두 개의 주사위를 동시에 던졌을 때, 두 값을 곱해서 홀수가 나올 확률은?
    - 모든 경우의 수: $6 \times 6 = 36$
    - 두 값을 곱해서 홀수가 나올 경우의 수: $3 \times 3 = 9$
    - 두 주사위의 곱이 홀수일 확률: $9/36 = 1/4 = 0.25$
    - 단, 두 주사위의 각 숫자가 나올 가능성이 같아야 한다는 조건을 만족해야 함.
  - 동일한 가능성의 가정:
    - 표본공간의 모든 경우가 나올 가능성이 같아야 한다.

- **통계적 확률:** *statistical* probability
  - 전체 시행 횟수 중에서 특정 사건이 일어난 횟수의 비율
    - 전체 시행 횟수를 n, 특정 사건이 일어날 횟수를 r이라고 하면
    - 그 사건이 일어날 통계적 확률 $= \frac{r}{n}$
  - 통계적 확률은 수학적 확률과 정확히 일치하지 않을 수 있음
    - 하지만, 시행 횟수를 늘릴수록 통계적 확률이 수학적 확률에 근접함

- **큰 수의 법칙**: law of large numbers
  - 표집 오차: *sampling error*
    - 시행 횟수가 적을 때는 통계적 확률이 수학적 확률에서 벗어남
  - 기댓값: *expected value*
    - 표본평균이 자료의 크기가 커짐에 따라 한없이 가까워지는 특정값
    - 시행 횟수가 많아질수록 통계적 확률은 기댓값에 가까워짐.
  - 동전 던지기를 10회 시행했을 경우:
    - 앞 뒤 앞 앞 뒤 뒤 뒤 앞 앞 앞
    - 앞면이 나올 기댓값 $= \ 0.5$
    - 앞면이 나올 통계적 확률 $= \ 6/10 \ = \ 0.6$
    - 표집오차 $= |0.6 - 0.5| = 0.1$

▪ 베르누이 시행: *Bernoulli trial*

- 가능한 결과가 두 개 밖에 없고, 성공의 확률이 정해져 있는 확률 시행
  - 예) 동전 던지기

- `rbinom(n, size, prob)`
  - `n`: number of observations.
  - `size`: number of trials.
  - `prob`: probability of success on each trial.

```
?rbinom

x <- rbinom(10, 1, 0.5)
x
sum(x)/10
mean(x)

x <- rbinom(100, 1, 0.5)
mean(x)

x <- rbinom(10000, 1, 0.5)
mean(x)
```

- **몬테카를로 시뮬레이션:** *Monte Carlo Simulation*
  - 충분히 큰 횟수의 시행을 통해서 복잡한 확률을 계산하는 방법
  - 시행 횟수가 늘어남에 따라 통계적 확률은 수학적 확률에 한없이 가까워진다.

  - `sample(x, size, replace=FALSE, prob=NULL)`
    - `x`: a vector of one or more elements from which to choose.
    - `size`: a non-negative integer giving the number of items to choose.
    - `replace`: should sampling be with replacement?
    - `prob`: a vector of probability weights for obtaining the elements.

▪ 복원 추출과 비복원 추출:

- 복원 추출: sampling *with replacement*
  - 표본공간에서 표본을 추출한 다음 원래대로 돌려놓고 다음 뽑기를 하는 방법
- 비복원 추출: sampling *without replacement*
  - 표본을 추출하고나서 원래대로 돌려놓지 않고 다음 뽑기를 하는 방법

```
sample(1:10, 10, replace=T)
sample(1:10, 10, replace=F)
```

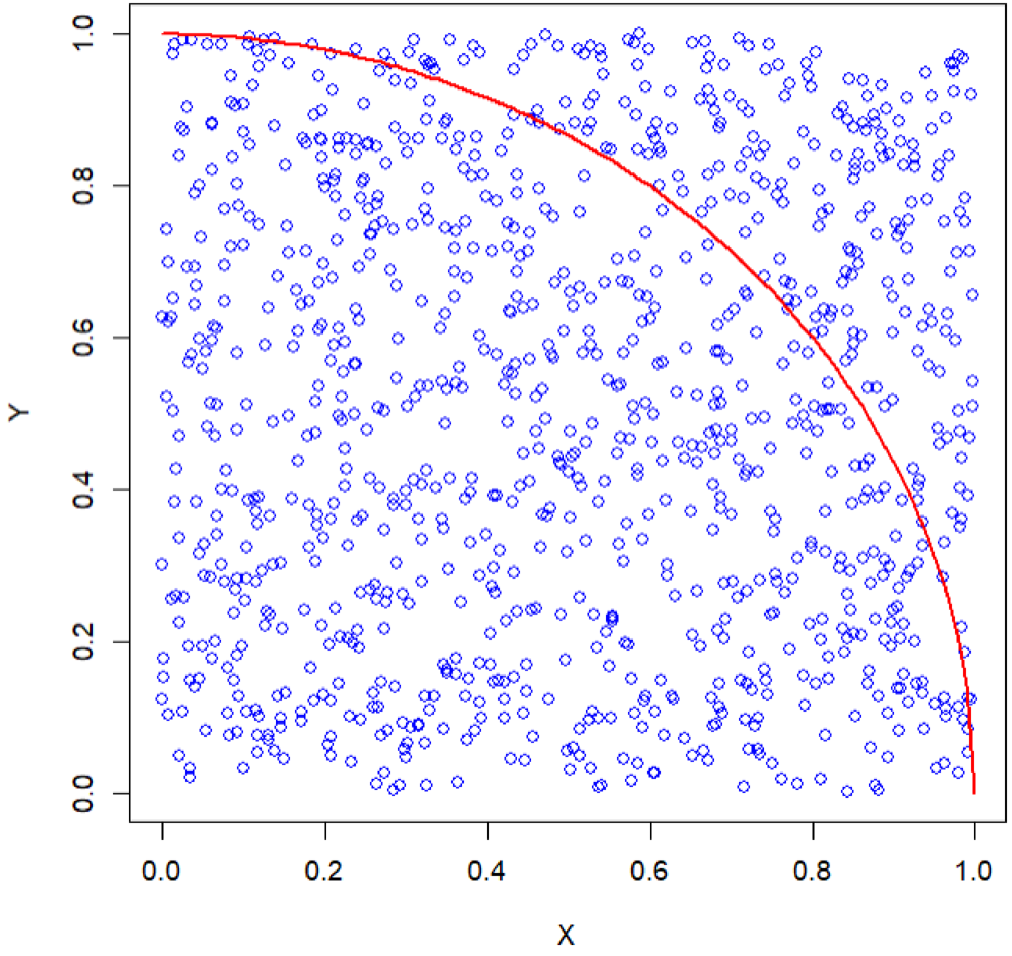
- **난수 생성:** *Random Number Generation*
  - 난수 생성기의 조건:
    - 수의 분포가 확률적으로 균일해야 하고,
    - 다음에 나올 값을 예측할 수 없어야 한다.
  - 컴퓨터를 이용한 난수 생성: 의사 난수 (*pseudo random*)
    - 완전한 난수는 아니지만 난수의 조건을 충족하는 알고리즘을 적용

  - `runif(n, min = 0, max = 1)`
    - `n`: number of observations.
    - `min`, `max`: lower and upper limits of the distribution.

- **몬테카를로 시뮬레이션으로 원주율(π) 계산하기**

```
n_sim <- 1000
x <- vector(length = n_sim)
y <- vector(length = n_sim)
res = 0
for (i in 1:n_sim) {
    x[i] <- runif(1)
    y[i] <- runif(1)
    if (x[i]^2 + y[i]^2 < 1) {
        res <- res + 1
    }
}
4 * res / n_sim

circle <- function (x) sqrt(1 - x^2)
plot(x, y, xlab='X', ylab='Y', col='blue')
curve(circle, from = 0, to = 1, add=T, col='red', lwd=2)
```



- **조건부 확률**: *conditional probability*
  - 어떤 사건이 참일 때 특정 사건의 확률이 얼마인지를 일컫는 개념
  - $P(B|A)$: 사건 A가 일어났을 때의 사건 B의 조건부 확률
    - $P(B|A) = \frac{P(A \cap B)}{P(A)}$, 단, $P(A) > 0$.
  - 확률의 곱셈정리
    - $P(A) > 0, P(B) > 0$일 때,
    - $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

▪ 조건부 확률 시뮬레이션:

- 주사위를 던져서 홀수가 나왔을 경우(A)에 5가 나올(B) 확률
  - $P(A) = 3/6 = 1/2$
  - $P(B) = 1/6$
  - $P(A \cap B) = 1/6$
  - $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/2} = \frac{1}{3} = 0.333\ldots$
  - $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/6} = 1$

```r
n_sim <- 10000
n_odd <- 0
n_five <- 0

for (i in 1:n_sim) {
    trial = sample(1:6, 1)
    if (trial %% 2 == 1) n_odd <- n_odd + 1
    if (trial == 5) n_five <- n_five + 1
}

n_five / n_odd
p_odd <- n_odd / n_sim
p_five <- n_five / n_sim
p_five / p_odd
```

- 베이즈 정리: *Bayes' Theorem*

$$P(B|A) = \frac{P(B|A)P(B)}{P(A)}$$

- **베이즈 정리:** *Bayes' Theorem*

likelihood

prior probability

$$P(B|A) = \frac{P(B|A)P(B)}{P(A)}$$

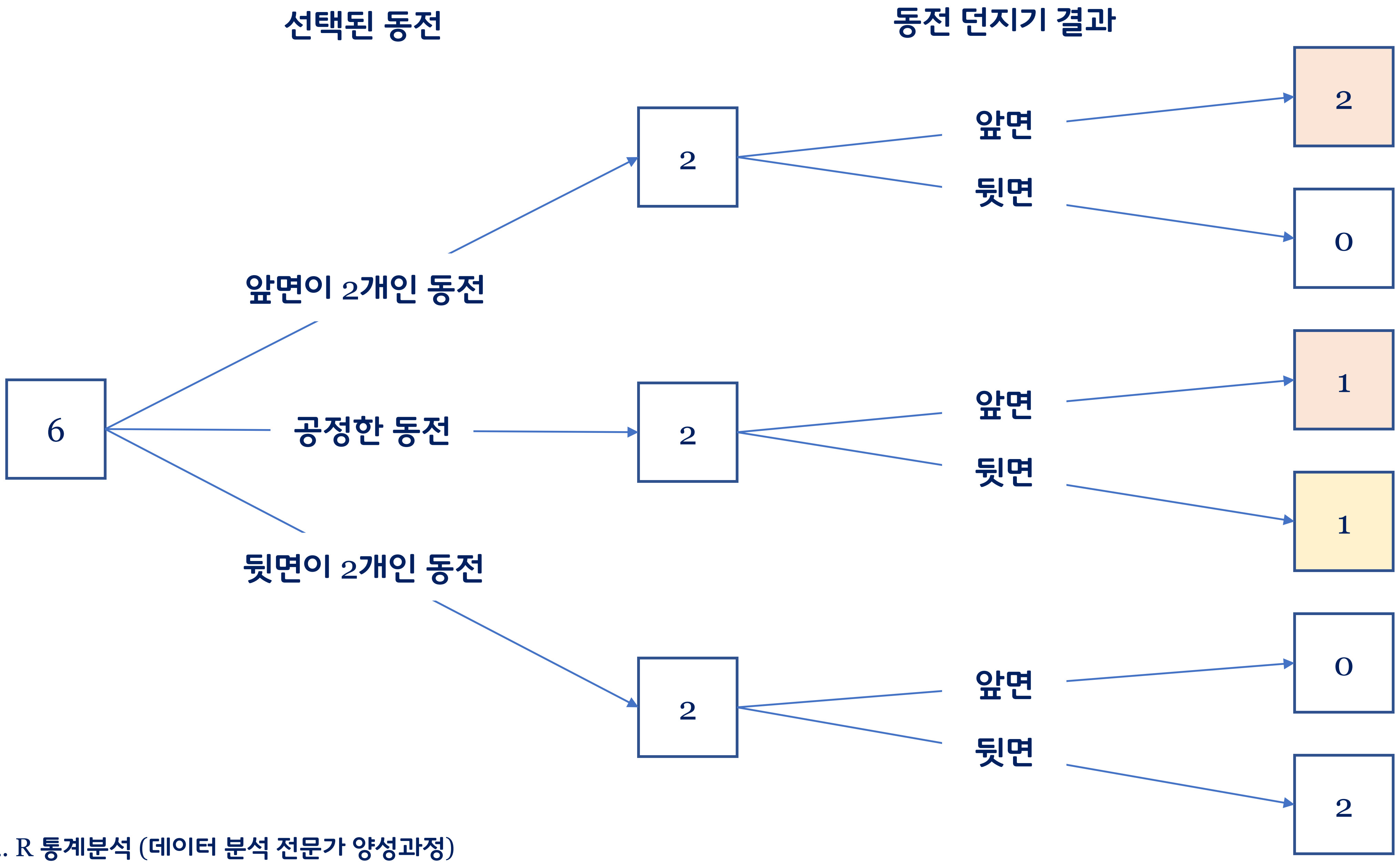
posterior probability

evidence

- 베이즈 정리의 확률 해석:
  - 확률은 사건의 발생에 대한 기대치의 계산과,
    - 실제로 그것이 발생할 것으로 기대되는 가능성 간의 비율이다.
  - 즉, 과거의 데이터를 보면 미래를 예측할 수 있다.
  - 빈도주의와 베이즈주의: *Frequentism* .vs. *Bayesianism*
    - 로널드 피셔 .vs. 토마스 베이즈
    - 동전 던지기를 해서 연속으로 열 번 앞면이 나온 후에
      - 다시 그 동전을 던졌을 때 앞면이 나올 확률은?
    - 내일 아침에 해가 동쪽에서 뜰까, 서쪽에서 뜰까?

▪ 연습문제:

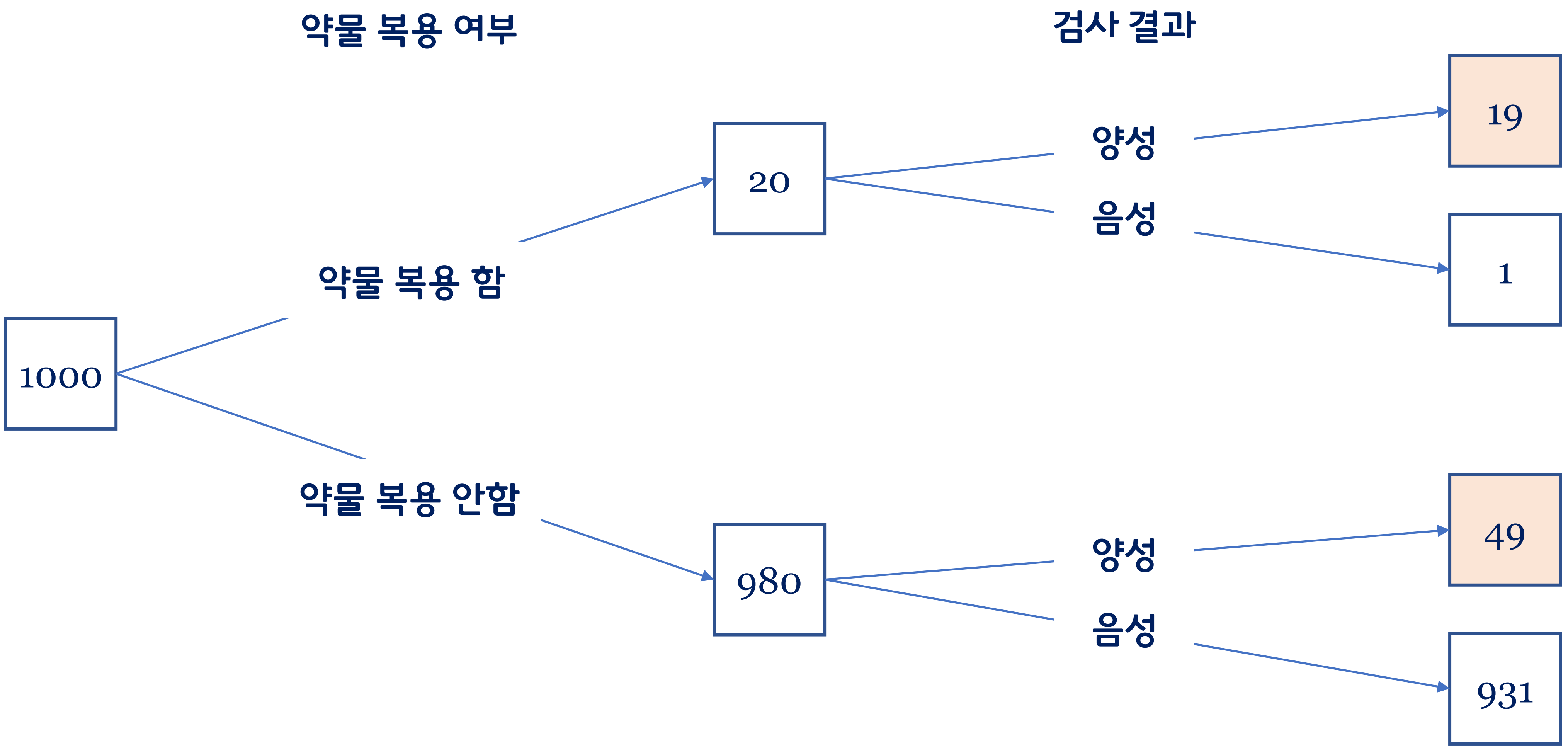
- 주머니에 동전 세 개가 있다. 하나는 앞면만 둘이고, 하나는 앞면과 뒷면이 각각 하나씩 있고, 하나는 뒷면만 둘이다. 당신이 임의로 동전을 하나 골라서 그것을 던졌는데 앞면이 나왔다면, 그 동전의 다른 면이 앞면일 확률은 얼마인가?
  - 답: 2/3



선택된 동전
동전 던지기 결과
6
앞면이 2개인 동전
공정한 동전
뒷면이 2개인 동전
2
2
2
앞면
뒷면
앞면
뒷면
앞면
뒷면
2
0
1
1
0
2

▪ 연습문제:

- 스포츠 경기에서 실시되는 금지약물 복용 검사의 정확도가 95%라고 가정하자. 즉 약물 복용자의 95%와 비복용자의 95%를 정확하게 분류한다고 하자. 즉, 선수 50명당 1명 꼴로 금지약물을 복용하고 있는데, 1,000명의 선수에 대한 약물 검사를 했다고 하자. 만약, 한 선수의 검사 결과가 양성이라면, 그 선수가 정말로 금지약물을 복용하고 있을 확률은 얼마인가?
  - 답: 19/68=28%

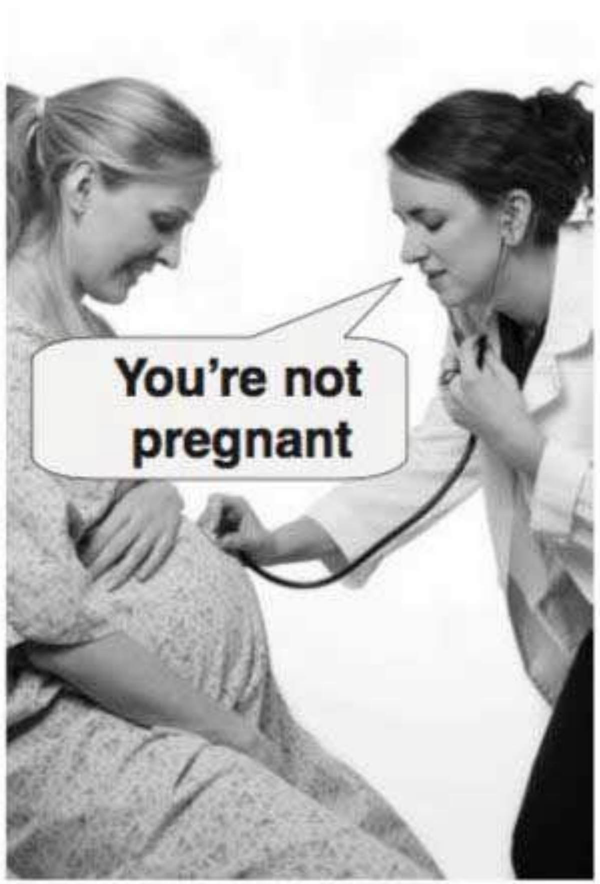

약물 복용 여부
검사 결과
1000
약물 복용 함
약물 복용 안함
20
980
양성
음성
양성
음성
19
1
49
931

- **혼동 행렬:** *confusion matrix*
  - 예측값이 실제값을 얼마나 정확히 예측했는지를 보여주는 행렬

		예측값	
		Positive	Negative
실제값	Positive	TP	FN
	Negative	FP	TN

Type I error
(false positive)

Type II error
(false negative)



▪ 혼동 행렬의 평가 지표: *evaluation metrics*

- 민감도 (재현율): *sensitivity* (recall)
  - $recall = \frac{TP}{TP+FN}$
  - 양성을 옳게 양성으로 진단할 확률(2종 오류를 저지르지 않을 확률)
- 특이도: *specificity*
  - $specificity = \frac{TN}{TN+FP}$
  - 음성을 옳게 음성으로 진단할 확률(1종 오류를 저지르지 않을 확률)
- 정확도: accuracy
  - $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- 정밀도: precision
  - $precision = \frac{TP}{TP+FP}$

연습문제:

- 민감도와 특이도가 모두 99%인 암 진단 검사가 있다.
  - 검사 결과 암 환자로 판정된 사람들 중에서 정말 암 환자일 확률은?
- 답: $1/11 = 0.0909\ldots$
  - 암 진단을 받은 환자 중에서 실제 암 환자는 10%도 되지 않는다고?
  - 1종 오류와 2종 오류의 확률이 둘 다 충분히 낮다 하더라도
    - 상황에 따라 직관에 반하는 결과가 나올 수 있다. 어떤 상황일까?

- 질병 진단 시뮬레이션:
  - 유병율이 0.1%인 질병 검사를 몬테카를로 시뮬레이션으로 검사하시오.
    - 단, 이 질병 검사의 민감도와 특이도는 둘 다 99%라고 가정한다.
    - 유병율: 검사 환자 중에서 실제로 질병이 있을 확률

```r
n_sim <- 10000      # 검사 횟수
prevalence <- 0.001 # 유병율

# 검사의 민감도, 특이도
sensitivity <- 0.99
specificity <- 0.99

n_total_positive <- 0  # 전체 질환 케이스 수
n_true_positive <- 0   # 실제 환자 수
n_false_positive <- 0  # 오진 케이스 수
```

```r
for (i in 1:n_sim) {
    # 유병율에 따라 실제 병의 유무를 할당함
    disease <- rbinom(1, 1, prevalence)
    if (disease == 0) { # 실제 병이 없는 경우
        diagnosis <- rbinom(1, 1, 1-specificity)
        if (diagnosis == 1) {
            n_total_positive <- n_total_positive + 1
            n_false_positive <- n_false_positive + 1
        }
    }
    if (disease == 1) { # 실제 병이 있는 경우
        diagnosis <- rbinom(1, 1, sensitivity)
        if (diagnosis == 1) {
            n_total_positive <- n_total_positive + 1
            n_true_positive <- n_false_positive + 1
        }
    }
}
```

```r
n_total_positive # 양성의 수
n_true_positive  # 진양성의 수
n_false_positive # 위양성의 수

n_false_positive / n_total_positive
```

Any Questions?