KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

4. 확률과 통계적 추론 I: 확률, 모집단과 표본

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

10:10-11:00

확률(개념, 연산, 분포)에 대해 이해하기: 정규, 이항, 포아송, 음이항 분포 등에 대한 실습

- Gailmard Ch. 4, CH. 6의 확률측도, 사건연산, PMF/CDF/PDF, 조건부/독립, Data Generating Process (DGP)와 분포를 연결

11:10-12:05

표본 $\rightarrow$ 모집단 추론의 다리 놓기, Part I.
K & W Ch. 7의 모집단, 표본, 표집분포, CLT, 신뢰구간 등에 대한 이해

12:15-12:45

표본 $\rightarrow$ 모집단 추론의 다리 놓기, Part II.
ModernDive Ch. 7, 8

Recap the Last Class!

Pop-up quizes!

각 변수의 유형을 구분해보자.

gender	sleep	school	countries	bodytemp
male	5.0	Elem	13	36.4
female	7.0	Elem	7	37.4
male	5.5	High	1	37.1
female	3.0	Col	9	36.5

Recap the Last Class!

Pop-up quizes!

각 변수의 유형을 구분해보자.

- `gender`: 명목형-이항형
- `sleep`: 연속형-등간형
- `school`: 명목형-분류형(혹은 순위형)
- `countries`: 명목형-분류형
- `bodytemp`: 연속형-비율형

Recap the Last Class!

Pop-up quizes!

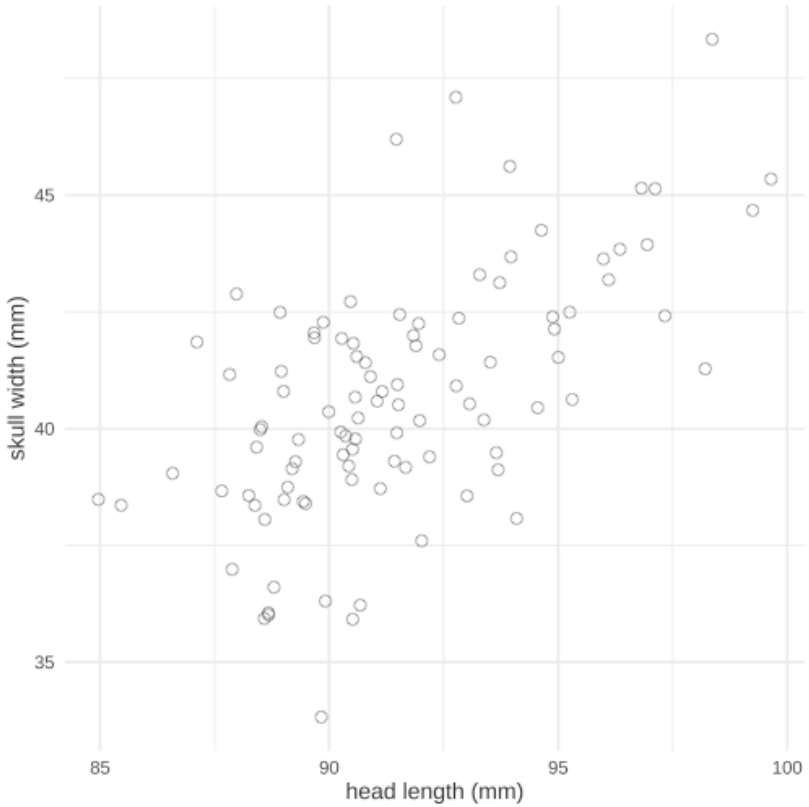
산포도를 보고 올바른 진술을 고르시오.

(a) 관계 없음(독립)

(b) 정(positive) 관계

(c) 길수록 두개골은 더 넓어진다

(d) 더 넓을수록 길다



Part I. 확률과 분포(Probability and Distribution)

확률과 통계적 추론 I: 확률, 모집단과 표본

확률, 분포, 그리고 가설

확률모델: 우리가 어떠한 이론적 기대인 모델을 가지고 있을 때, 어떠한 데이터(현실)를 관측할 확률에 대해 알려줌.

동전 던지기

- 앞면을 관측할 확률을 대략 0.5일 것이라고 기대
- 정확히는 동전을 반복해서 던졌을 때, 데이터가 어떻게 나타날 것(동전의 앞/뒷면이 반반)이라는 이론을 가지고 있다는 것

확률과 통계적 추론 I: 확률, 모집단과 표본

확률, 분포, 그리고 가설

$$\Pr(Y|M) = \Pr(\text{데이터}|\text{모델})$$

- 즉, 확률이란 무한정 반복되는 무작위한 과정에서 우리가 어떠한 결과를 관측할 횟수의 비율(proportion of times)
- **무작위한 과정(random process)**: 무슨 일이 일어날 것이라는 건 알지만, 그 결과가 어떤 것인지는 알지 못하는 상태
    - 동전 던지기, 주사위 굴리기 등은 결과가 무엇이든 도출될 것은 알지만 구체적인 결과를 불확실

확률과 통계적 추론 I: 확률, 모집단과 표본

확률의 유형

사건 A 와 B

결합확률(Joint probability)

$\Pr(A \cap B)$ 또는 $\Pr(A \text{ and } B)$

한계확률(Marginal probability)

$\Pr(A)$ 또는 $\Pr(B)$

조건부확률(Conditional probability)

한계확률 대비 결합확률의 비율

$$\Pr(A|B) = \Pr(A \cap B) / \Pr(B)$$

확률과 통계적 추론 I: 확률, 모집단과 표본

확률의 유형: 한계확률과 결합확률

Table: 한계확률과 결합확률 (1)

구분	A_{Fight}	A_{Comply}	총합
B_{Fight}	0.3	0.2	0.5
B_{Comply}	0.1	0.4	0.5
총합	0.4	0.6	1

A와 B라는 국가는 서로 갈등 중

- A가 싸우기로 결정할 확률은?
  /B가 싸우기로 결정할 확률은?
- A가 상대방의 요구에 순응할 확률은?
  /B가 상대방의 요구에 순응할 확률은?
- A와 B가 모두 싸우기로 결정할 확률은?
  /A와 B가 모두 상대방에게 순응할 확률은?

확률과 통계적 추론 I: 확률, 모집단과 표본

확률의 유형: 한계확률과 결합확률

Table: 한계확률과 결합확률 (1)

구분	A_{Fight}	A_{Comply}	총합
B_{Fight}	0.3	0.2	0.5
B_{Comply}	0.1	0.4	0.5
총합	0.4	0.6	1

A와 B라는 국가는 서로 갈등 중

- $\Pr(A_{\text{Fight}}) = 0.4$
  $\Pr(B_{\text{Fight}}) = 0.5$
- $\Pr(A_{\text{Comply}}) = 0.6$
  $\Pr(B_{\text{Comply}}) = 0.5$
- $\Pr(A_{\text{Fight}} \cap B_{\text{Fight}}) = 0.3$
  $\Pr(A_{\text{Comply}} \cap B_{\text{Comply}}) = 0.4$

확률과 통계적 추론 I: 확률, 모집단과 표본

확률의 유형: 한계확률과 결합확률

Table: 한계확률과 결합확률 (2)

구분	A	B	총합
남성	40	60	100
여성	65	35	100
총합	105	95	200

유권자들에게 두 가지 질문:
(1) A 후보 vs B 후보 투표
(2) 남성 vs 여성

- 응답자가 여성일 확률은?
- 응답자가 남성일 확률은?
- 여성 중에서 A 후보에게 투표했을 조건부 확률은?
- 응답자가 여성이면서 A 후보에게 투표했을 확률은?

확률과 통계적 추론 I: 확률, 모집단과 표본

확률의 유형: 한계확률과 결합확률

Table: 한계확률과 결합확률 (2)

구분	A	B	총합
남성	40	60	100
여성	65	35	100
총합	105	95	200

유권자들에게 두 가지 질문:
(1) A 후보 vs B 후보 투표
(2) 남성 vs 여성

- $\Pr(\text{여성}) = 100/200 = 0.5$
- $\Pr(\text{남성}) = 100/200 = 0.5$
- $\Pr(A \mid \text{여성}) = 65/100 = 0.65$
- $\Pr(A, \text{여성}) = 65/200 = 0.325$

확률과 통계적 추론 I: 확률, 모집단과 표본

독립(Independence)

어떤 사건의 결과를 알아도 다른 사건의 결과에 정보가 없다면 독립

- 동전 던지기 1회 결과는 2회 결과에 영향을 주지 않음 $\rightarrow$ 독립
- 반대로 정보가 있다면 종속. 예를 들어, 카드 덱에서 에이스를 한 장 뽑았다면 다음 에이스 확률은 변함 $\rightarrow$ 종속

확률과 통계적 추론 I: 확률, 모집단과 표본

확률변수(Random variables)

무작위 사건의 결과를 수량화하여 나타냄.

- X 는 변수, x는 그 변수의 구체적인 값을 나타냄.
  - $X = \{1, 2, 3, \ldots, x\}$
  - $\Pr(X = x)$: 확률변수 X 가 x 라는 구체적인 값을 가질 확률
  - 이때, x 는 이산형(정수) 혹은 연속형(실수)일 수 있음.

확률과 통계적 추론 I: 확률, 모집단과 표본

기대값(Expectation)

확률변수의 평균적 결과 = 기대값

- 모집단을 대표하는 값이나, 우리는 현실에서 가지고 있는 데이터인 표본을 대표하는 값인 평균으로 이 기대값을 대응함.

$$\mu = E(X) = \sum_{i=1}^{k} x_i \Pr(X = x_i)$$

1~6의 눈을 가진 주사위의 기대값을 구하면 어떻게 될까?

$$E(X) = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$$

확률과 통계적 추론 I: 확률, 모집단과 표본

이산형 확률 변수의 기대값

R-code | Plot | Expected value

```
library(tidyverse)
discrete <- tibble(x = c(rep("0", 60), rep("1", 20),
                "2", "3", "4", rep("5", 10),
                "6", "7", "8", "9", rep("10", 3)))
discrete |> ggplot(aes(x = x)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(breaks = c(seq(0, 0.6, 0.1)), labels = scales::percent) +
  scale_x_discrete(limits = c(unique(discrete$x)))
```

확률과 통계적 추론 I: 확률, 모집단과 표본

분포(Distribution)

확률분포(Probability distribution)는 임의의 사건이나 값 x_i가 나타날 확률 p_i와의 체계적 관계를 나타냄.

- 어떤 확률변수 X가 취할 수 있는 가능한 값들을 $x_1, x_2, \ldots, x_n$이라고 할 때, 각 값이 발생할 확률을 $p_1, p_2, \ldots, p_n$으로 정의
- 즉, 확률분포는 다음과 같은 쌍 (x_i, p_i)들의 집합으로 표현

$$\Pr(X = x_i) = p_i, \quad (i = 1, 2, \ldots, n)$$

- 이때, 확률변수는 다음의 조건을 만족해야 함.
  - (1) 확률은 0 이상이어야 함: $p_i \geq 0 \quad \forall i$, (2) 확률의 전체 합은 1: $\sum_{i=1}^{n} p_i = 1$

확률과 통계적 추론 I: 확률, 모집단과 표본

분포(Distribution): 예시

동전 던지기

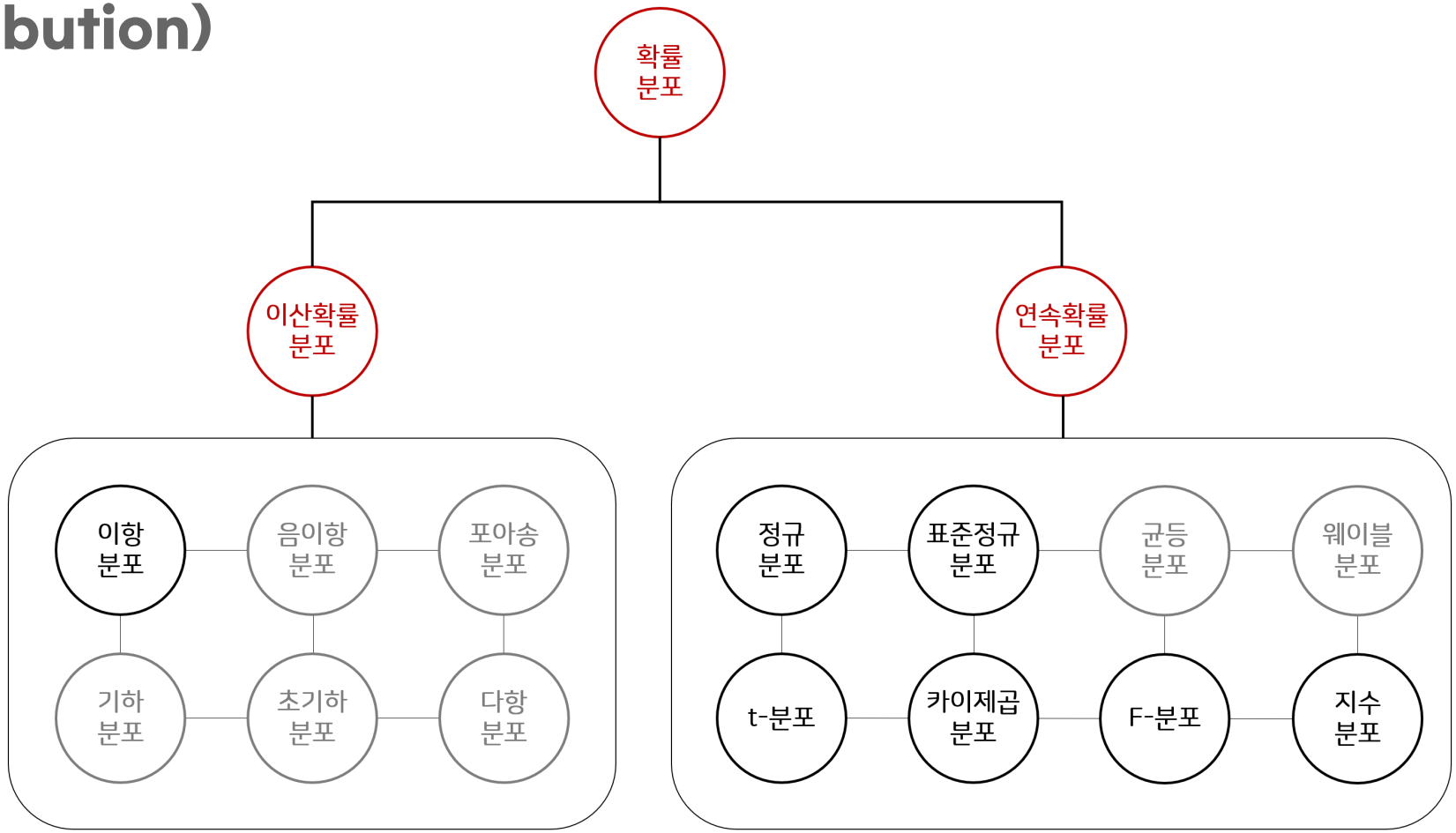
$x_i \in \{\text{앞}, \text{뒤}\}, p_1 = 0.5, p_2 = 0.5 \rightarrow$ 확률분포: $\{(\text{앞}, 0.5), (\text{뒤}, 0.5)\}$

주사위 던지기

$x_i \in \{1, 2, 3, 4, 5, 6\}, p_i = \frac{1}{6} \rightarrow$ 확률분포: $\{(1, 1/6), (2, 1/6), \ldots, (6, 1/6)\}$

확률과 통계적 추론 I: 확률, 모집단과 표본

분포(Distribution)



확률과 통계적 추론 I: 확률, 모집단과 표본

분포(Distribution)

확률질량함수(Probability Mass Function, PMF)

이산확률분포(Discrete distribution): 베르누이분포, 이항분포, 포아송분포

이산형 확률변수 X에 대해 $p(x) = P(X = x)$ 의 관계를 보여주는 함수

확률밀도함수(Probability Density Function, PMF)

연속확률분포(Continuous distribution): 정규분포, 지수분포

연속형 확률변수 X에 대한 확률을 나타내는 함수 $f(x)$. $\Pr(a \leq X \leq b) = \int_a^b f(x), dx$

확률과 통계적 추론 I: 확률, 모집단과 표본

분포(Distribution)

누적분포함수(Cummulative Distribution Function, CDF)

주어진 확률변수의 값이 특정한 값보다 크거나 작을 확률을 나타내는 함수

- 이산형: $F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} p_i$
- 연속형: $F(x) = \int_{-\infty}^{x} f(t)dt$
  - t 는 $-\infty$부터 x까지의 구간에서 흘러가는 변수를 의미
  - x는 내가 누적 확률을 계산하고 싶은 지점(상한)

확률과 통계적 추론 I: 확률, 모집단과 표본

누적분포함수(Cummulative Distribution Function, CDF)

R-code Plot

```
x <- seq(-5, 5, length = 100)
plot(x, dnorm(x), type = "l", col = "#DB3A2F",
     ylab = "Density", xlim = c(-5, 5), ylim = c(0, 1))
text(-3, 0.2, "PDF of Normal Distribution", col = "#DB3A2F")
par(new=TRUE)
plot(x, pnorm(x), type = "l", col = "#275D8E",
     ylab = "Density", xlim = c(-5, 5), ylim = c(0, 1))
text(2, 0.5, "CDF of Normal Distribution", col = "#275D8E")
```

확률과 통계적 추론 I: 확률, 모집단과 표본

데이터 생성 과정(Data Generating Process, DGP)와 분포 선택

분포 = 가정한 데이터 생성 과정(DGP)

- 정규: 연속적인 값
- 이항/로짓: 이진 선택(있다/없다)
- 포아송/음이항: 카운트 · 과산포: 횟수
- 지수/웨이블: 지속시간

체계적 부분 + 확률적 부분의 모형 관점

확률과 통계적 추론 I: 확률, 모집단과 표본

분포와 R

R을 사용하면 쉽게 밀도, 누적분포함수, 분위, 확률값 등을 구할 수 있음.

`dname`: 밀도(density)

`pname`: 누적분포함수(CDF)

`qname`: 분위(quantile)

`rname`: 확률값(random values)

확률과 통계적 추론 I: 확률, 모집단과 표본

분포와 R

R을 사용하면 쉽게 밀도, 누적분포함수, 분위, 확률값 등을 구할 수 있음.

dnorm	pnorm	qnorm
rnorm	dbinom	pbinom
qbinom	rbinom	

```
dnorm(x = 3, mean = 2, sd = 5)
```

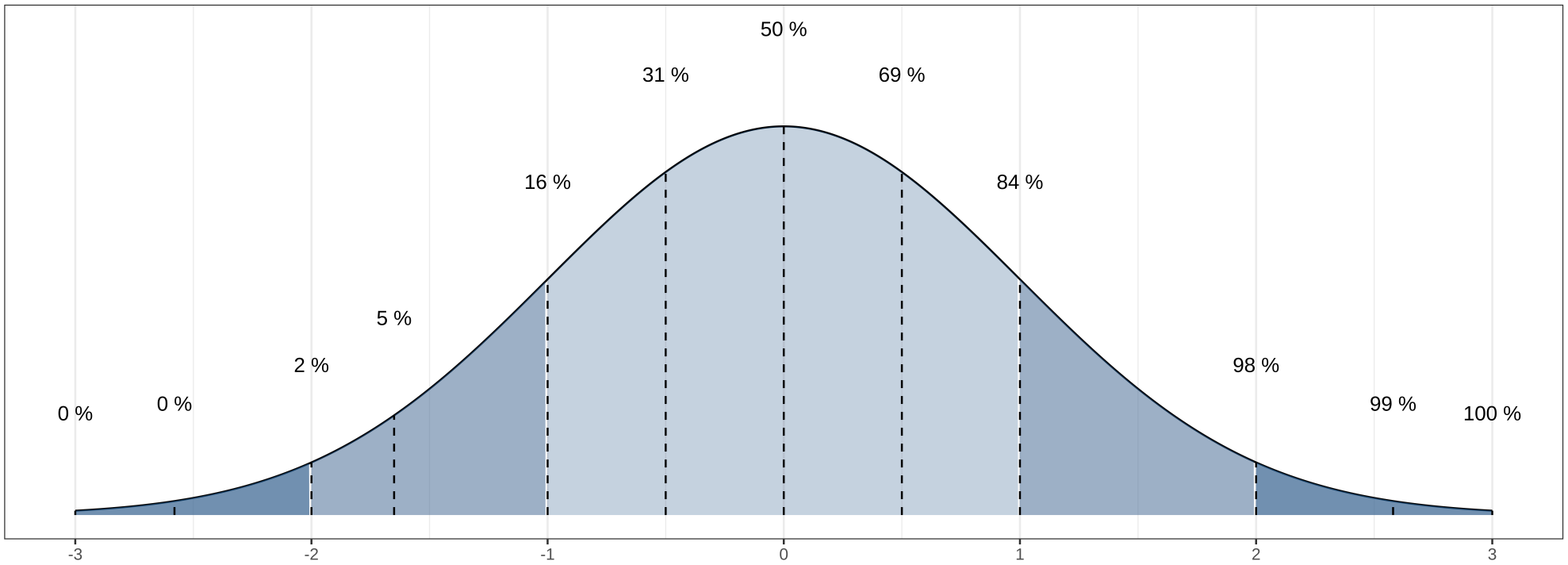
```
## [1] 0.07820854
```

Command	Distribution
*binom	Binomial
*t	t
*pois	Poisson
*f	F
*chisq	Chi-Squared

확률과 통계적 추론 I: 확률, 모집단과 표본

정규분포(Normal distribution)

표준정규분포: $\pm 1SD \approx 68\%$, $\pm 2SD \approx 95\%$, $\pm 3SD \approx 99\%$



확률과 통계적 추론 I: 확률, 모집단과 표본

정규분포(Normal distribution)

평균 50, SD 10인 정규분포에서 70의 위치?

```
pnorm(70, mean = 50, sd = 10, lower.tail = TRUE)
```

```
## [1] 0.9772499
```

```
pnorm(70, mean = 50, sd = 10, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

```
1 - pnorm(70, mean = 50, sd = 10, lower.tail = TRUE)
```

```
## [1] 0.02275013
```

확률과 통계적 추론 I: 확률, 모집단과 표본

정규분포(Normal distribution)

평균 50, SD 10인 정규분포에서 70의 위치?

70은 하위 약 97.7% (상위 2.3%)

확률과 통계적 추론 I: 확률, 모집단과 표본

Student's t: t-분포

정규와 유사하나 꼬리가 두꺼움.

- 자유도(degree of freedom)가 커질수록 정규에 수렴
  - 추정 과정에서 **독립적으로 변할 수 있는 정보 조각의 수**
  - 자유도가 작다는 것의 의미는? 추정한 모수(제약)가 생기면, 데이터 값들 사이에 **구속조건**이 생겨서 자유롭게 변할 수 있는 차원이 감소
  - 표본분산에서 $n-1$로 나누어주는 것에서 -1은 표본평균인 $\bar{x}$를 이미 추정했다는 사실 이 하나의 제약을 만들어 $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ 이 항상 성립.
  - 따라서 $(x_i - \bar{x})$ 들은 **완전히 독립적이지 않다** $\rightarrow$ 자유도는 n이 아니라 $n-1$

확률과 통계적 추론 I: 확률, 모집단과 표본

Student's t: t-분포

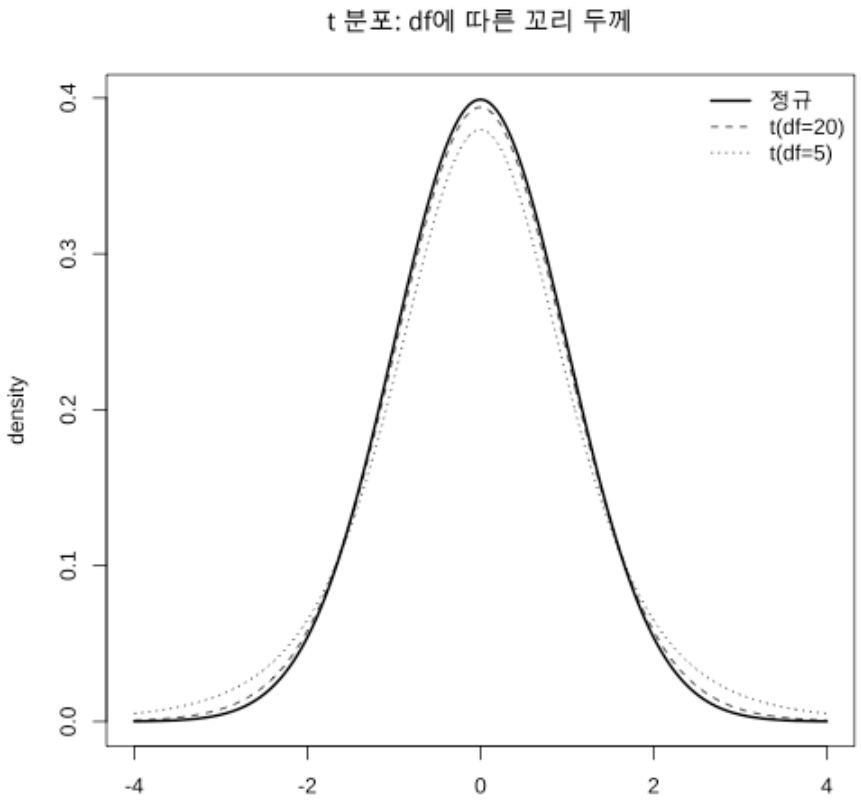
정규와 유사하나 꼬리가 두꺼움.

- 자유도(degree of freedom)가 커질수록 정규에 수렴
  - t-분포에서의 자유도
  - 일표본(one-sample) 평균 검정/CI: 모분산을 모르므로 s를 추정 $\rightarrow$ 제약 1개
  - $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}, \quad t \sim t_{n-1}$
- 평균 신뢰구간의 오차한계에 $t(se)$ 등장(다음 주차에 자세히 배울 것임)
  - 자유도가 작을수록 꼬리 두꺼움 $\rightarrow$ 더 큰 임계값 $\rightarrow$ 더 넓은 신뢰구간

확률과 통계적 추론 I: 확률, 모집단과 표본

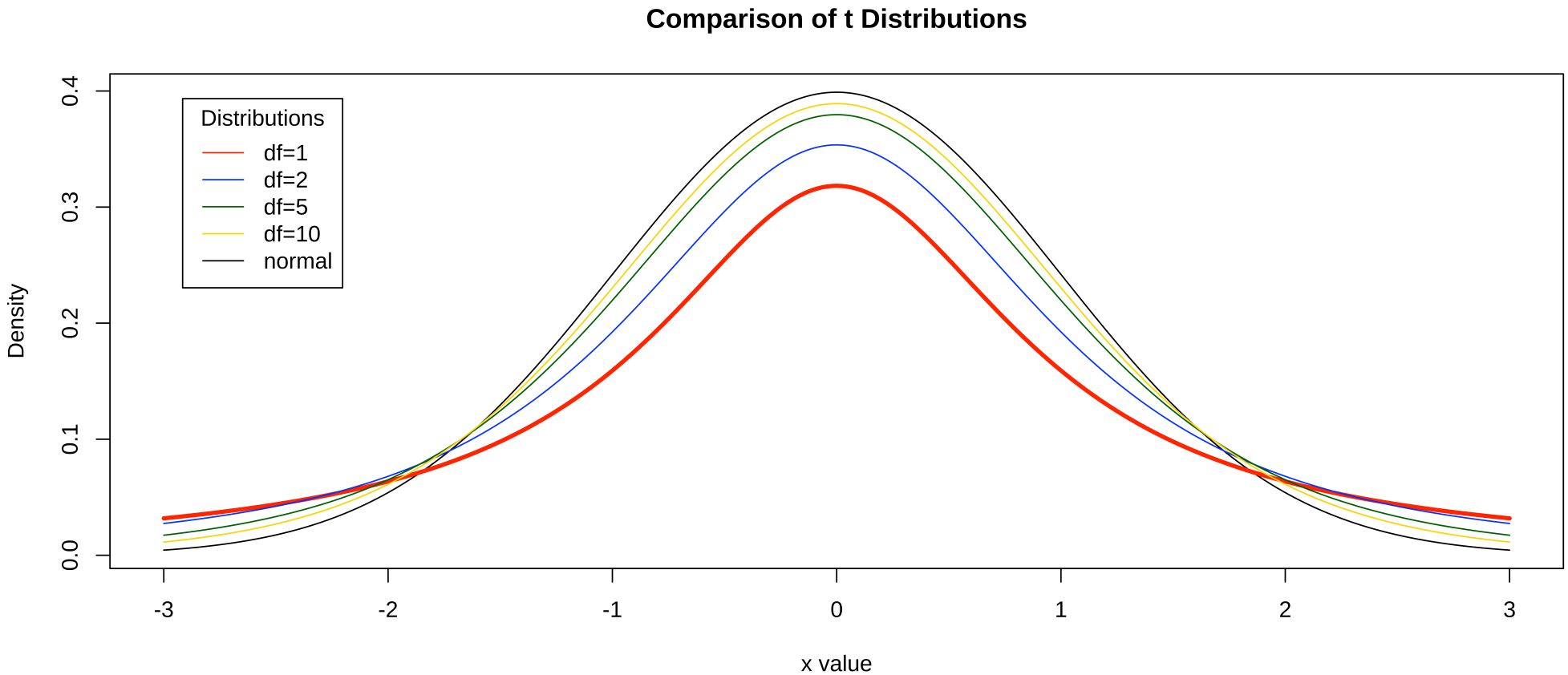
Student's t: t-분포

```
x <- seq(-4, 4, length.out = 400)
plot(x, dnorm(x), type="l", lwd=2,
     ylab="density", xlab="",
     main="t 분포: df에 따른 꼬리 두께")
lines(x, dt(x, df=20), lty=2)
lines(x, dt(x, df=5),  lty=3)
legend("topright",
       c("정규", "t(df=20)", "t(df=5)"),
       lwd=c(2,1,1), lty=c(1,2,3),
       bty="n")
```



확률과 통계적 추론 I: 확률, 모집단과 표본

Student's t: t-분포



확률과 통계적 추론 I: 확률, 모집단과 표본

이항분포(Binomial distribution)

n 번 시행에서 k 번 성공할 확률

- 각 시행은 **독립**이며 결과는 **성공(1)** 또는 **실패(0)**
- 모든 시행의 성공확률은 **고정** p
- 관심 변수: $Y =$ 성공 횟수

확률질량함수(PMF)

$$\Pr(Y = y) = \binom{n}{y} p^{y}(1-p)^{n-y},$$
$$y = 0, 1, \ldots, n$$

조합적 직관

- n번 중 **어느 y개의 위치에서 성공이 나오는지** 경우의 수: $\binom{n}{y}$
- 각 배치는 $p^{y}(1-p)^{n-y}$의 확률

확률과 통계적 추론 I: 확률, 모집단과 표본

이항분포(Binomial distribution)

n 번 시행에서 k 번 성공할 확률

이항분포의 특성

기대값: $\mathbb{E}[Y] = np$

분산: $\mathrm{Var}(Y) = np(1-p)$

누적분포(CDF):

$$\Pr(Y \leq y) = \sum_{i=0}^{y} \binom{n}{i} p^i (1-p)^{n-i}$$

확률과 통계적 추론 I: 확률, 모집단과 표본

이항분포(Binomial distribution)

n 번 시행에서 k 번 성공할 확률

```
pbinom(27, size=100, prob=0.25, lower.tail = TRUE)
```

```
## [1] 0.7223805
```

27번 성공은 하위 72.2%, 상위 27.8%

확률과 통계적 추론 I: 확률, 모집단과 표본

이항분포(Binomial distribution)

R-code: Parameters | R-code: Plot | Plot

```r
# 파라미터
n <- 20; p <- 0.3
# PMF/CDF/분위/난수
dbinom(6, size=n, prob=p)          # P(Y=6
```

```
## [1] 0.191639
```

```r
pbinom(6, size=n, prob=p)          # P(Y<=
```

```
## [1] 0.6080098
```

```r
qbinom(.95, size=n, prob=p)        # 95% 분
```

```
## [1] 9
```

```r
rbinom(5, size=n, prob=p)          # 난수 5
```

```
## [1]  7  6 11  6  8
```

확률과 통계적 추론 I: 확률, 모집단과 표본

이항분포(Binomial distribution)

Pop-up Quiz

$n = 50, p = 0.2$에서 Y의 기대값과 분산은?

- **HINT**: 이항분포 $Y \sim \text{Binomial}(n, p)$ 일 때, $\mathbb{E}[Y] = np$, $Var(Y) = np(1-p)$
- $E[Y] = 10$, $\text{Var}(Y) = 8$

확률과 통계적 추론 I: 확률, 모집단과 표본

포아송분포(Poisson distribution)

단위 시간/공간에서 희소한 사건의 발생 건수

희소한 사건의 발생 건수, 평균 = 분산 = λ

- 시간(혹은 공간) 축에서 사건이 드물고 독립적으로 발생
- 평균 발생률(강도) λ (단위당 평균 사건수)

확률질량함수

$$\Pr(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, 2 \ldots$$

확률과 통계적 추론 I: 확률, 모집단과 표본

포아송분포(Poisson distribution)

단위 시간/공간에서 희소한 사건의 발생 건수

희소한 사건의 발생 건수, 평균 = 분산 = λ

- 시간(혹은 공간) 축에서 사건이 드물고 독립적으로 발생
- 평균 발생률(강도) λ (단위당 평균 사건수)

포아송분포의 특성

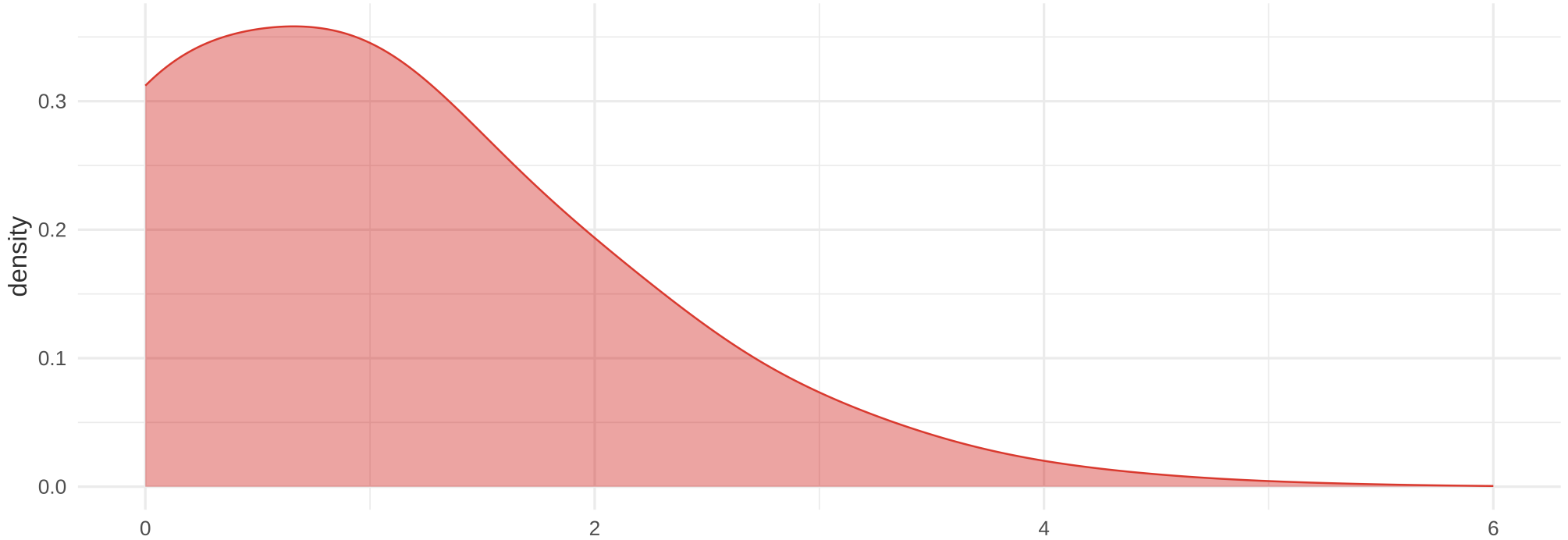
기대값: $\mathbb{E}[Y] = \lambda$

분산: $\mathrm{Var}(Y) = \lambda$ (평균=분산)

누적분포: $\Pr(Y \leq y) = \sum_{i=0}^{y} e^{-\lambda} \frac{\lambda^i}{i!}$

확률과 통계적 추론 I: 확률, 모집단과 표본

포아송분포(Poisson distribution)



확률과 통계적 추론 I: 확률, 모집단과 표본

포아송분포(Poisson distribution)

R-code: Parameters | R-code: Plot | Plot

```r
lambda <- 3
# PMF/CDF/분위/난수
dpois(4, lambda)                    # P(Y=4)
```

```
## [1] 0.1680314
```

```r
ppois(4, lambda)                    # P(Y<=4)
```

```
## [1] 0.8152632
```

```r
qpois(.95, lambda)                  # 95% 분위
```

```
## [1] 6
```

```r
rpois(5, lambda)                    # 난수 5개
```

```
## [1] 3 4 5 2 5
```

```r
# 포아송 과정의 시간확장: 길이 t 구간의 평균은 lamb
t <- 2; dpois(4, lambda*t)      # 길이 2에서
```

```
## [1] 0.1338526
```

확률과 통계적 추론 I: 확률, 모집단과 표본

음이항분포(Negative binomial distribution)

n 번째 시도에서 k 번째에 성공할 확률

- 각 시행은 **독립**이고, 결과는 **성공(1)** 또는 **실패(0)**
- 각 시행의 성공확률은 **항상** p (고정)
- N = k**번째 성공이 처음 나오는 시점(시도 횟수)** $\rightarrow$ 지원: $n = k, k+1, \ldots$
- 동치표현: $X = N - k$ = k **번째 성공 전까지 실패 횟수** $\rightarrow$ 지원: $x = 0, 1, 2, \ldots$

확률과 통계적 추론 I: 확률, 모집단과 표본

음이항분포(Negative binomial distribution)

이항분포와의 관계

이항분포는 **정해진 시도 n에서 성공이 몇 번?**을 묻는 반면,

음이항분포는 **성공을 k번 얻을 때까지 몇 번 시도?**를 묻는 분포

Part II. 모집단과 표본(Population and Sample)

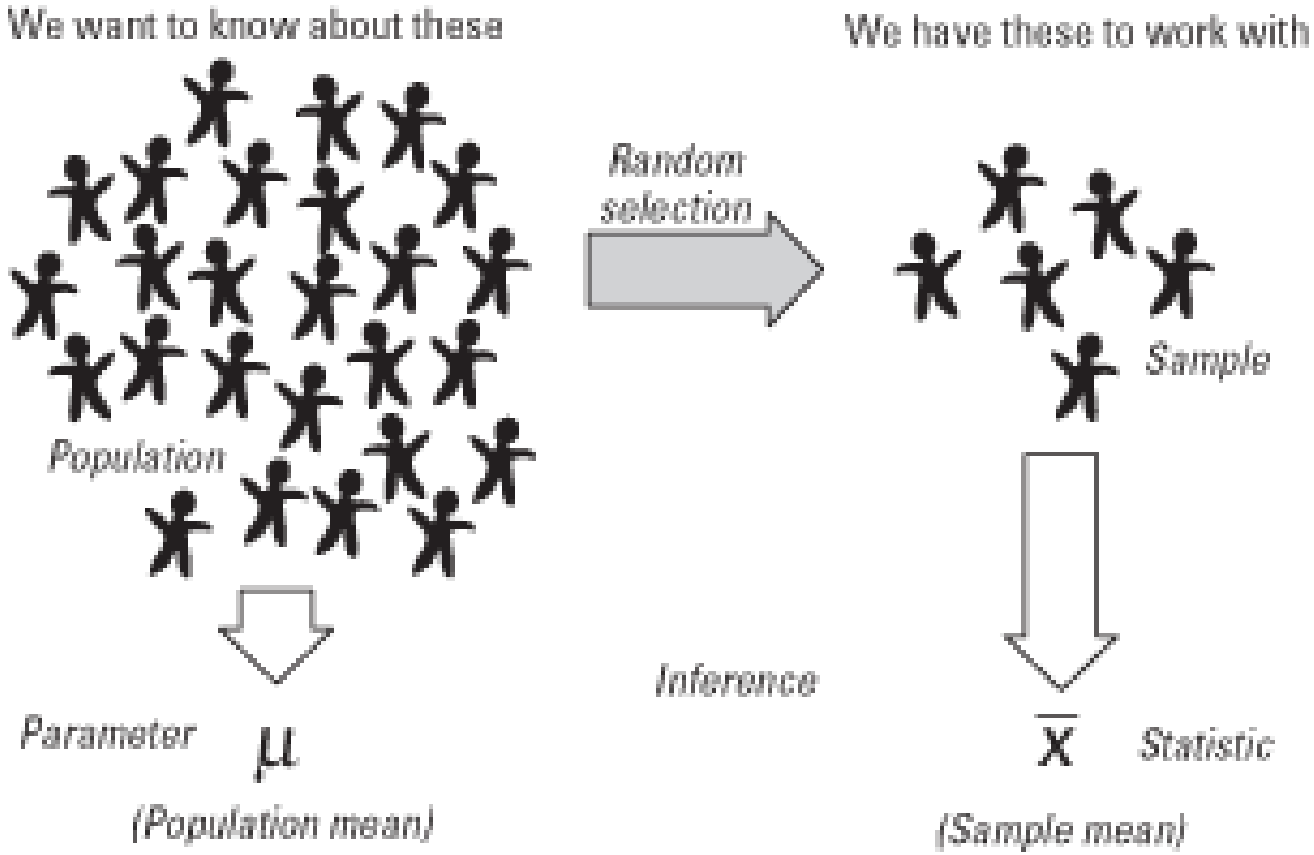
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본

우리는 모집단에 대한 이론적 기대를 가진다.

"모집단은 ~할거야!" 모집단을 직접 관측/획득할 수 없으므로, 표본을 바탕으로 모집단에 대한 추론을 수행

"표본이 ~하니, 모집단도 ~할거야!" 그러나 이러한 추론은 단정적(deterministic)이지 않음!



확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본

표본: 모집단의 무작위 부분집합

표본 통계량의 **표집분포**가 추론의 핵심

중심극한정리(Central Limit Theorem, CLT)

표본평균의 분포 $\approx$ 정규분포, $SE = \sigma/\sqrt{n}$

무작위 표집이 이상적 (편의표집은 위의 가정이 깨질 수 있음.)

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본

중심극한정리(Central Limit Theorem, CLT)

R-code | R-code: Plot | Plot

```r
set.seed(2025)
N <- 1e6
pop <- tibble(x = rexp(N, rate = 1/5))  # 평균=5, 한쪽 꼬리
clt_means <- function(n, B=5000){
  replicate(B, mean(sample(pop$x, n)))
}
m_small <- clt_means(5)
m_mid   <- clt_means(30)
m_big   <- clt_means(200)
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: CLT 미니 시뮬레이션

R-code | R-code: Plot | Plot

```
one_sample <- tibble(x = sample(1:6, size = 600, replace = TRUE))
mean(one_sample$x)
```

```
## [1] 3.388333
```

```
rep_means <- replicate(2000, mean(sample(1:6, size = 600, replace = TRUE)))
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표집변동성

그릇(모집단)에서 숟가락(n)으로 공을 퍼서 $\hat{p}$ 추정

여러 번 반복하면 $\hat{p}$ 가 매번 달라짐 $\rightarrow$ 표집변동성

R-code-1 R-code-2 Plot

```
set.seed(2025); N <- 100000
p_true <- 0.34
pop <- tibble(ball = rbinom(N, 1, p_true)); n <- 50
one_scoop <- sample(pop$ball, size = n, replace = TRUE)
mean(one_scoop)
```

```
## [1] 0.28
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본 크기에 따른 표집분포 폭 비교

R-code Plot

```
compare_n <- function(n, B = 1000, p = 0.34){
  replicate(B, mean(rbinom(n, 1, p)))
}
set.seed(2025)
d <- list(
  n_50  = compare_n(50),
  n_200 = compare_n(200),
  n_800 = compare_n(800)
) |>
  enframe(name = "group", value = "props") |>
  unnest(props)
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본을 통한 모집단 추론 맛보기

모집단을 우리가 모른다고 가정

- 각각의 변수들로부터 50개의 관측치들을 무작위 표본을 추출
- 이 무작위 표본을 5,000번 반복하여 뽑고 그렇게 뽑힌 5,000개의 무작위 표본들의 평균 분포를 그래프로 시각화
- 사망률에 대한 무작위 표본 1개를 뽑아보고, 그 표본 이름은 `samp1.mort` 라고 지정

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본을 통한 모집단 추론 맛보기

```
library(WDI)
WDI.data <-
  WDI(country = "all",
      indicator = c("SH.DYN.NMRT","DC.DAC.USAL.CD", "SH.VAC.TTNS.ZS",
                    "SP.URB.TOTL.IN.ZS", "NE.TRD.GNFS.ZS"),
      start = 1990, end = 2005, extra = FALSE, cache = NULL)

# 사망률
samp1.mort <- sample(WDI.data$SH.DYN.NMRT, 50)

# 표본 평균과 모집단 평균을 비교
mean(samp1.mort, na.rm = TRUE) # 사망률 표본의 평균
```

```
## [1] 24.73767
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본을 통한 모집단 추론 맛보기

```
## na.rm 옵션은 표본에 결측치가 있을 수 있을 때 중요하게 사용
## 결측치 빼고 평균을 구하란 뜻
mean(WDI.data$SH.DYN.NMRT, na.rm = TRUE) # 사망률 모집단의 평균
```

```
## [1] 22.51515
```

5,000개의 표본들을 루프 (loop)를 가지고 뽑아서 그 각각의 평균들을 벡터로 저장

- 이를 위해서 먼저 빈깡통, 빈벡터 (Blank vector)를 만듦.

```
# 빈 벡터 만들기
sample_means50.mort <- rep(NA, 5000)
# 루프로 5,000개의 표본평균을 구해 저장하기
for(i in 1:5000) {
  samp <- sample(WDI.data$SH.DYN.NMRT, 50)
  sample_means50.mort[i] <- mean(samp, na.rm = TRUE)}
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본을 통한 모집단 추론 맛보기

여기에서 루프의 의미는 원래 우리가 가지고 있던 WDI의 사망률 지표에서 50개씩 꺼낸 표본을 1개라고 할 때, 이와 같은 과정을 5,000번 반복하라는 것

- 그러면 5,000개의 표본을 얻게 되고, 이 중 당연히 결측치도 있을 수 있으므로 `na.rm()` 옵션으로 결측치를 제외하여 평균을 구하라고 코딩하는 것
- 평균을 구할 때, 요소 중에 NA(결측치)가 있으면 전체 평균 값도 NA로 계산되므로 이걸 처리
- 이렇게 구한 5,000개의 표본들의 평균들이 어떻게 분포되어 있는지 확인

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 표본을 통한 모집단 추론 맛보기

```
mean(sample_means50.mort, na.rm = TRUE)
```

```
## [1] 22.5144
```

```
mean(WDI.data$SH.DYN.NMRT, na.rm = TRUE)
```

```
## [1] 22.51515
```

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 다르게 생각해보기

비모수 부트스트래핑

우리의 표본은 모집단으로부터 무작위 추출된 것

따라서 만약 표본으로부터 무작위 *복원(replacement)* 추출을 한다면, 그 결과는 모집단으로부터 무작위 추출한 또 다른 표본이라고 할 수 있을 것

즉, 표본으로부터 무작위 복원추출을 한다는 것은 모집단으로부터 표본을 추출하는 행위를 것을 모방하는 것

확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 다르게 생각해보기

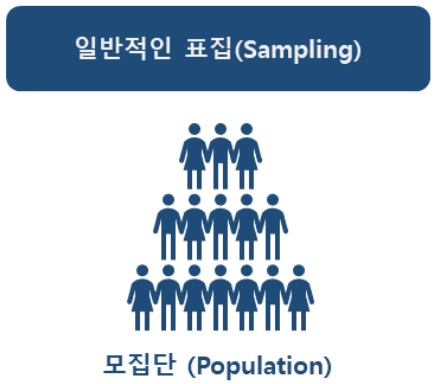
비모수 부트스트래핑

비모수 부트스트래핑을 통해서

- 표집분포를 얻을 수 있으며,
- 표집분포에 대한 함수도 어떤 것이든 얻을 수 있음.
- $g = 1, \ldots, G$번 표본재추출을 한다고 할 때, 총 G개의 결과값을 저장하고 그 결과를 요약해서 보여줄 수 있게 됨: 평균, 표준편차, 히스토그램, 신뢰구간 등

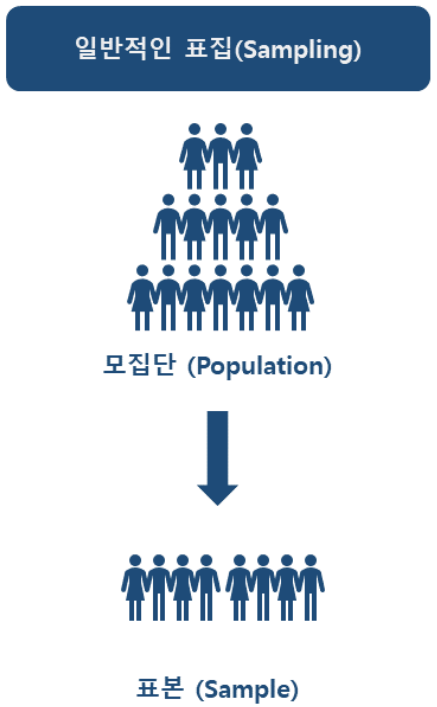
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



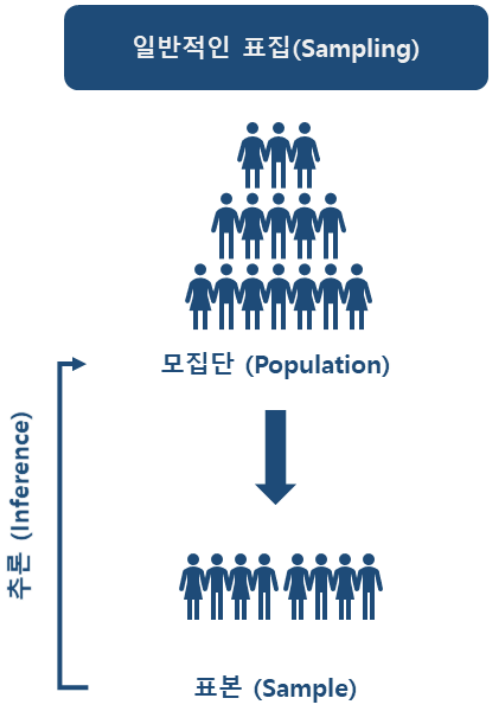
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



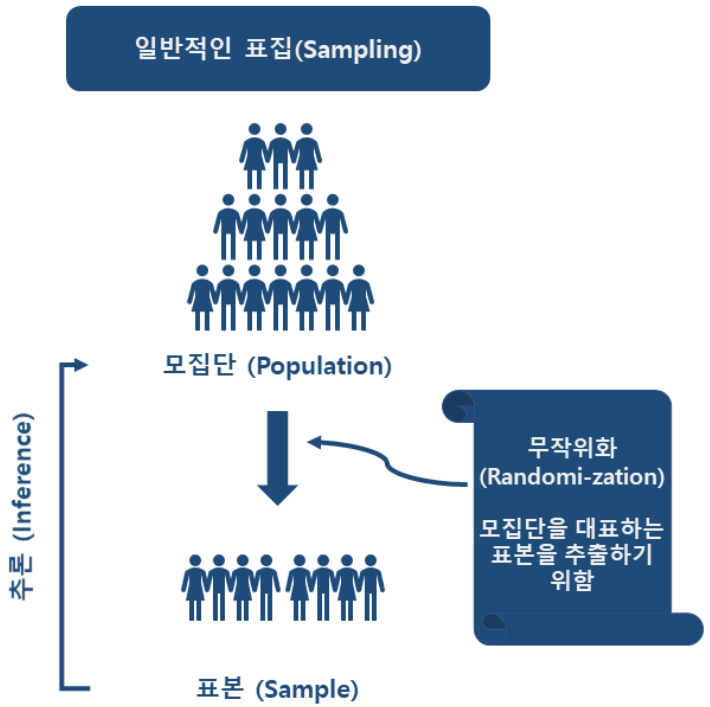
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



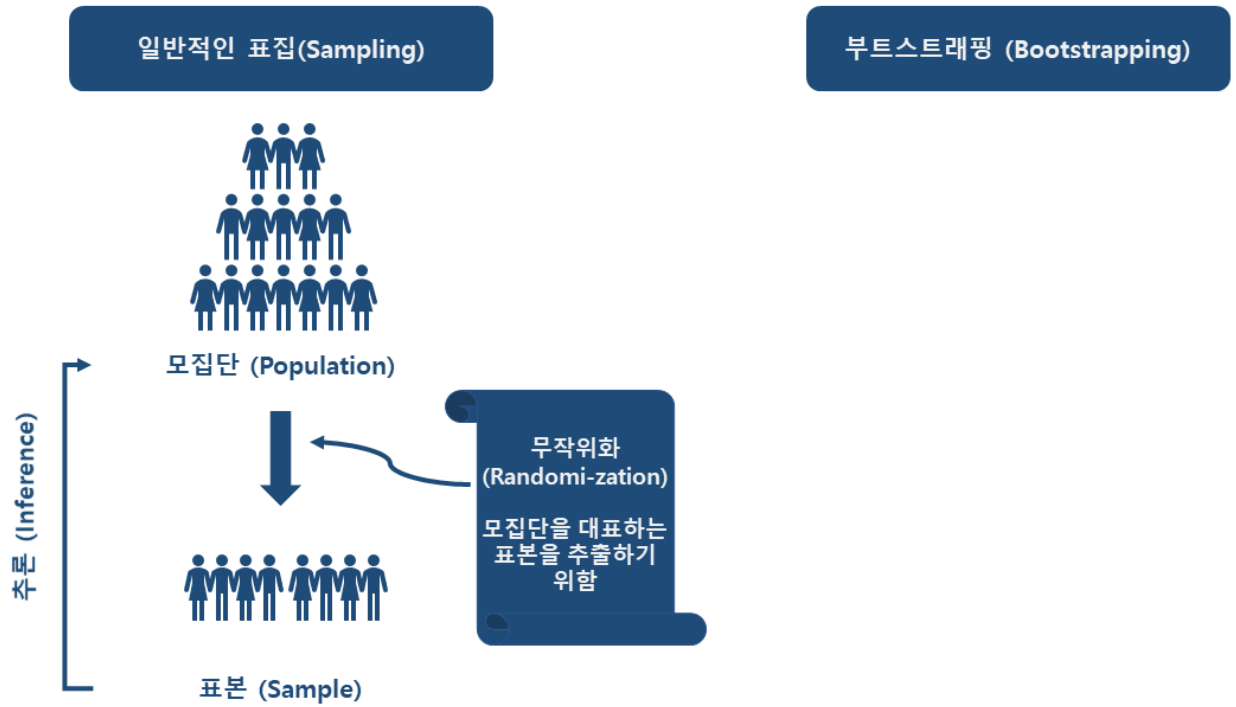
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



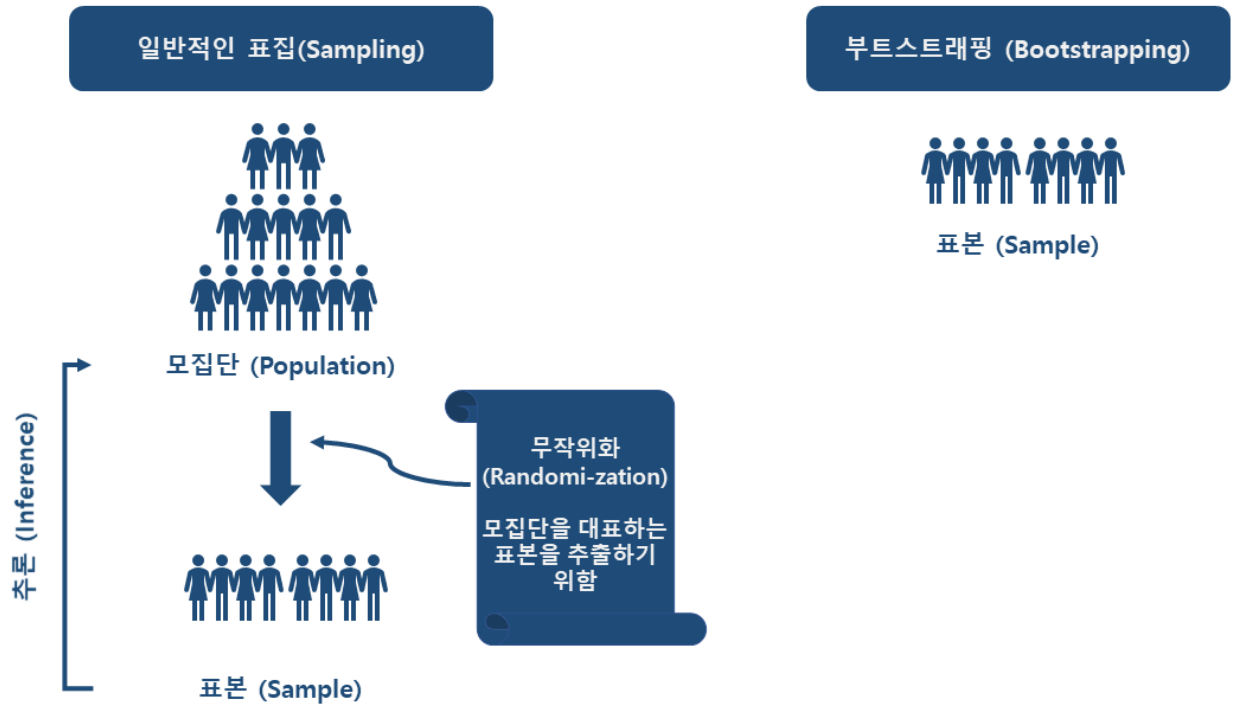
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



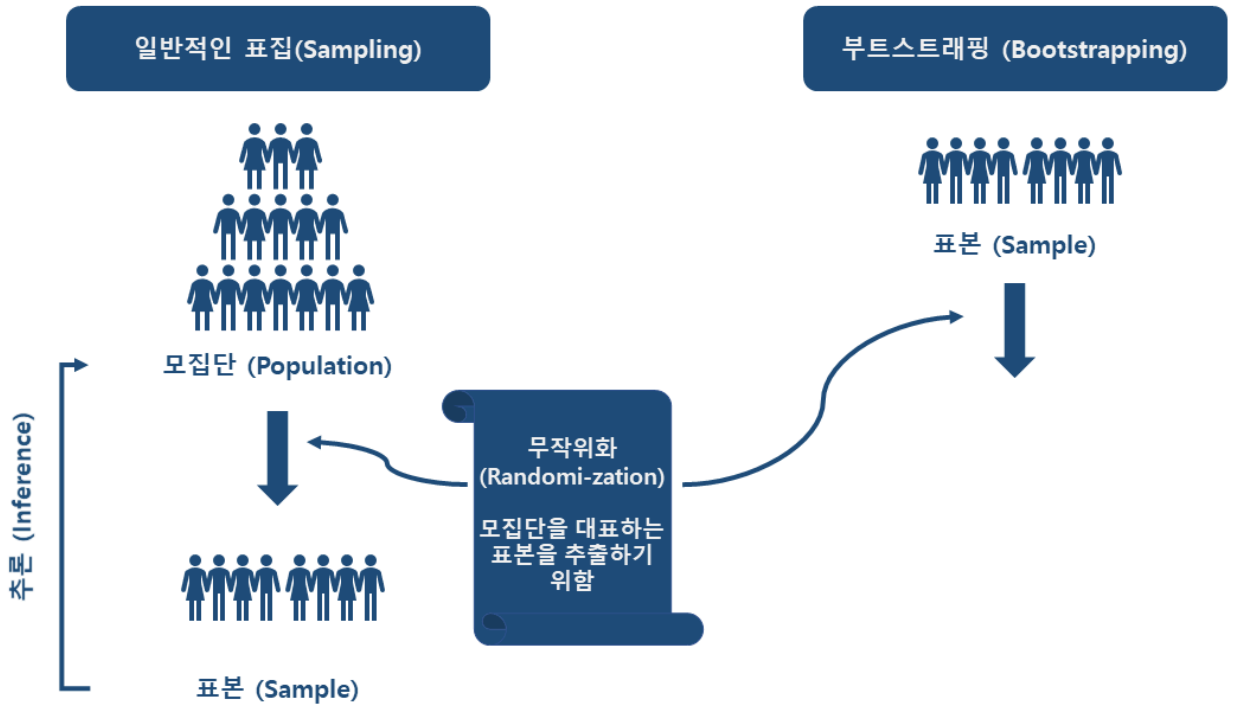
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



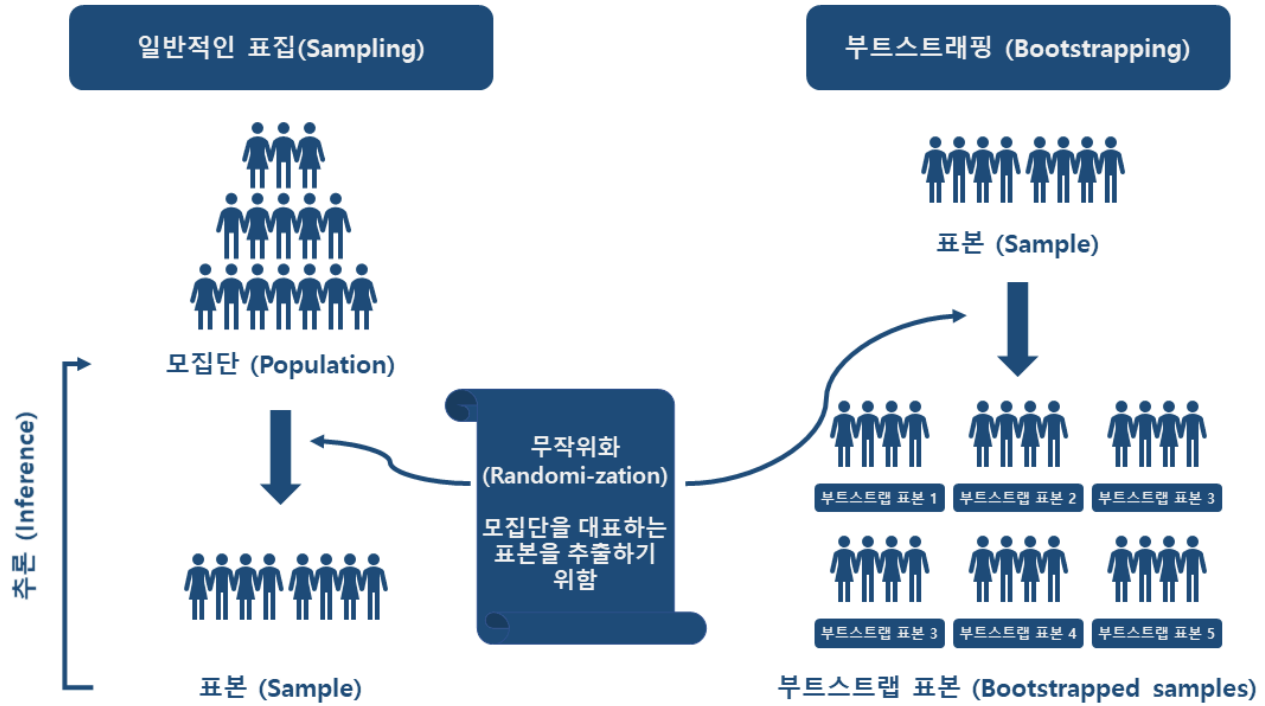
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



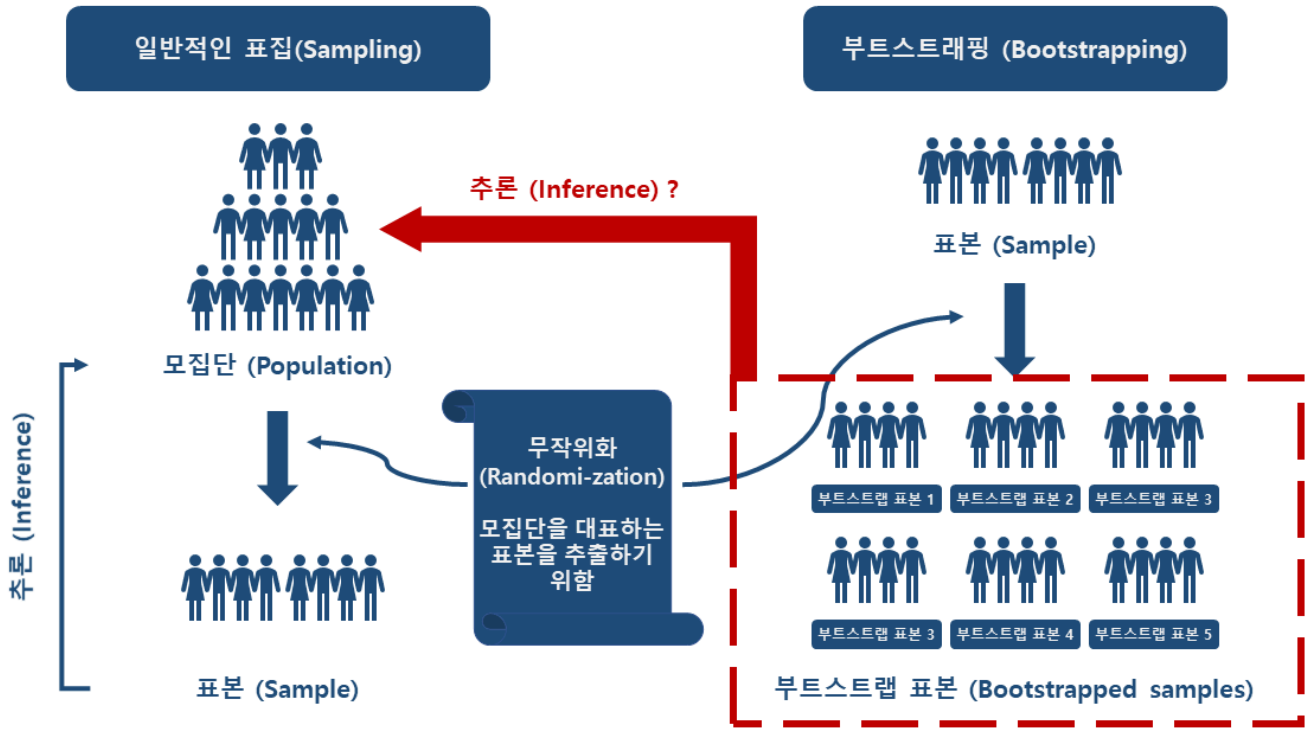
확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



확률과 통계적 추론 I: 확률, 모집단과 표본

모집단과 표본: 비모수 부트스트래핑



Part III. R을 이용한 시각화 실습 및 질의응답

다음 강의에는 정규분포와 신뢰구간에 대해 살펴볼 것

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405