예측 분석을 위한 기초 이론

정석오

1 확률분포의 개념

1.1 확률

어떤 시행에서 관측 가능한 모든 결과를 모은 집합을 표본공간 (sample space) 이라 하고, 표본공간의 부분집합을 사건 (event) 라 한다.

예제 동전을 두 번 연속으로 던지는 시행에 의해 생성되는 표본공간을 S라 하면

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

과 같다. 이 때 두 번 연달아 같은 면이 나오는 사건을 A라 하면

$$A = \{(H,H),(T,T)\}$$

이 된다. □

확률 (probability) 에 대한 엄밀한 수학적 정의가 따로 있지만 이른바 '고전적 정의'에 한정해 다루기로 한다. 확률은 주어진 사건에 대해 0과 1사이의 숫자를 대응시킨 함수로서 다음과 같이 정의된다. 사건 A의 확률 $P(A)$은

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

이다. 단, $|\cdot|$은 집합의 크기 (예: 셀 수 있는 집합의 경우 원소의 개수가 됨) 이다.

예제 (계속) $P(A) = \frac{|\{(H,H),(T,T)\}|}{|\{(H,H),(H,T),(T,H),(T,T)\}|} = \frac{2}{4}$. □

두 사건 A와 B에 대해 $A \cup B$를 합사건, $A \cap B$는 곱사건이라 한다. 합사건의 확률 $P(A \cup B)$과 곱사건의 확률 $P(A \cap B)$은 다음과 같은 등식을 만족한다.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

예제 (계속) 사건 B를 한 번 이상 앞면이 나오는 사건이라 하면 $B = \{(H,H),(H,T),(T,H)\}$이 되고, $P(B) = \frac{3}{4}$, $P(A \cap B) = P(\{(H,H)\}) = \frac{1}{4}$가 성립한다. 따라서 $P(A \cup B) = \frac{2}{4} + \frac{3}{4} - \frac{1}{4} = 1$이 된다. 실제로 $A \cup B = S$임을 확인할 수 있다. □

1.2 조건부확률

두 사건 A와 B에 대해 조건부 확률 $P(A|B)$는 다음과 같이 정의한다.

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

단 $P(B) > 0$이다. 이 정의를 이용하면 곱사건의 확률 $P(A \cap B)$에 대한 등식

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

가 성립한다. 만약 $P(A|B) = P(A)$ 또는 $P(B|A) = P(B)$를 만족하면 두 사건 A와 B이 서로 독립 (independent)이라 하며, 다음의 등식이 성립한다.

$$P(A \cap B) = P(A)P(B)$$

예제 (계속) $P(A|B) = \frac{1/4}{3/4} = \frac{1}{3}$. □

1.2.1 베이즈 정리

표본공간 S가 사건들의 열 $A_1, A_2, ..., A_N$에 의해 분할(partition)될 때(즉 $S = \cup_n A_n$이고, $i \neq j$이면 $A_i \cap A_j = \phi$), 다음의 등식이 성립한다.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{n=1}^{N} P(B|A_n)P(A_n)}, \quad i = 1, 2, ..., N$$

예제 고객의 남녀 성비가 6:4이라 한다. 그러나 클레임 비율은 남성의 경우 5%, 여성의 경우 10%이다. 어느 날 클레임 접수가 발생했을 때 해당 고객이 여성일 확률을 계산해보자. A_1과 A_2는 각각 고객이 남성인 사건과 여성인 사건을, B는 클레임이 발생하는 사건을 나타낸 것으로 하면

$$P(A_1) = 0.6, \quad P(A_2) = 0.4, \quad P(B|A_1) = 0.05, \quad P(B|A_2) = 0.1$$

이 된다. 따라서 클레임 접수가 발생했을 때 해당 고객이 여성일 확률은

$$\begin{aligned} P(A_2|B) &= \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} \\ &= \frac{0.1 \times 0.4}{0.05 \times 0.6 + 0.1 \times 0.4} \\ &\approx 0.5714 \ (> 0.4 = P(A_2)) \end{aligned}$$

와 같이 계산된다.

```
prior <- c(0.6, 0.4)
likelihood <- c(0.05, 0.10)
posterior <- prior*likelihood/sum(prior*likelihood)
probs <- data.frame(Prior = prior, Posterior = posterior)
row.names(probs) <- c("Male", "Female")
round(probs, 4)
```

```
##        Prior Posterior
## Male     0.6    0.4286
## Female   0.4    0.5714
```

□

1.3 확률변수

확률변수(random variable)란 표본공간을 정의구역으로 갖는 실함수이다. 즉, 표본공간의 모든 원소에 대해 실수인 대응값을 갖는다. 알파벳 대문자로 나타내며 시행의 결과 확률변수의 값이 실현된 것은 해당 소문자로 나타낸다.

예제 (계속) X를 앞면이 나온 횟수로 정의하면

$$X((T,T)) = 0, \quad X((T,H)) = 1, \quad X((H,T)) = 1, \quad X((H,H)) = 2$$

가 된다. □

확률변수가 취하는 값이 이산적인 경우 이산형 확률변수로, 연속적인 경우 연속형 확률변수로 분류한다. 참고로 위 예제의 확률변수 X는 이산형 확률변수이다. 확률변수는 시행의 결과에 따라 값이 결정되기 때문에 불확실성을 갖는다. 이 불확실성을 분포라는 개념을 통해 표현할 수 있다. 우선 이산형 확률변수의 경우에 대해 생각해보자. 이산형 확률변수 X가 있을 때 이 확률변수가 취할 수 있는 값 x에 대해 음이 아닌 숫자를 대응시키되 모든 가능한 x에 대해 구한 합이 1이 되도록 한 것을 X의 확률질량함수(probability mass function, pmf)라 한다. 즉, 확률질량함수 $p(x)$는 다음의 성질을 만족하게 된다.

$$\begin{aligned} &p(x) \geq 0, \quad \forall x \\ &\sum_{\forall x} p(x) = 1 \\ &P(X \in A) = \sum_{x \in A} p(x) \end{aligned}$$

예제 (계속) $x = 0, 1, 2$에 대해 함수 $p(x)$를 다음과 같이 정의하면 이는 X의

확률질량함수이다.

$$p(0) = P(X = 0) = P(\{(T, T)\}) = 1/4,$$
$$p(1) = P(X = 1) = P(\{(T, H), (H, T)\}) = 1/2,$$
$$p(2) = P(X = 2) = P(\{(H, H)\}) = 1/4$$

□

연속형 확률변수의 경우에는 특정한 하나의 값을 가질 확률(고전적 정의)은 0이 되므로 확률질량함수 형태로 확률분포를 표현할 수 없다. 대신 X의 값이 임의의 구간 $[a, b]$에 속할 확률을

$$P(X \in [a, b]) = \int_a^b f(x)dx$$

과 같이 나타낼 수 있게 해 주는 함수 $f(x)$, 즉 확률밀도함수(probability density function, pdf) 또는 간단히 줄여 밀도함수(density function, density)를 통해 확률분포를 표현한다. 확률밀도함수 $f(x)$는 다음을 만족한다.

$$f(x) \geq 0, \quad \forall x$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$
$$P(X \in A) = \int_A f(x)dx$$

실은 확률분포를 가장 일반적으로 표현할 수 있는 방법은 누적분포함수(cumulative distribution function, cdf)를 이용하는 것이다. 이를 줄여서 분포함수라 부르기도 한다. 확률변수 X가 연속형이든 이산형이든 상관없이 누적분포함수의 정의는 다음과 같다.

$$F(x) = P(X \leq x)$$

이 함수를 이용하면 앞에서 소개한 pmf와 pdf를 각각 차분과 미분을 이용해 구할 수 있다.

$$p(x) = F(x) - F(x-) \quad \text{(discrete)}$$

$$f(x) = F'(x) \quad \text{(continuous)}$$

1.4 확률분포의 특성값

1.4.1 기대값

확률분포의 특성을 숫자 몇 개로 요약하는 방법을 소개하고자 한다. 이런 숫자들을 모수(parameters)라 한다. 대표적 모수인 기대값(expectation, expected value, mean)은 분포의 중심 위치를 나타내는 대표값으로서, 확률변수 X의 기대값은 $\mathsf{E}(X)$로 나타내며 다음과 같이 정의된다.

- X가 이산형일 때
$$\mathsf{E}(X) = \sum_{\forall x} xp(x)$$
- X가 연속형일 때
$$\mathsf{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

기대값이 큰 값인 확률변수는 평균적으로 큰 값을, 기대값이 작은 확률변수는 평균적으로 작은 값을 가지는 경향성이 있다. 따라서 기대값을 구하면 굳이 시행을 통해 확률변수의 값을 실현해보지 않아도 평균적으로 어떤 값을 가지게 될 지 예상(기대)할 수 있게 된다.

1.4.2 분산과 표준편차

분산(variance)은 분포의 산포 정도를 나타내는 대표값이다. 확률변수 X의 분산은 $\mathsf{Var}(X)$로 나타내며, 다음과 같이 편차제곱의 기대값으로 정의된다.

- X가 이산형일 때
$$\mathsf{Var}(X) = \mathsf{E}\left[\{X - \mathsf{E}(X)\}^2\right] = \sum_{\forall x}\{x - \mathsf{E}(X)\}^2 p(x)$$
- X가 연속형일 때
$$\mathsf{Var}(X) = \mathsf{E}\left[\{X - \mathsf{E}(X)\}^2\right] = \int_{-\infty}^{\infty}\{x - \mathsf{E}(X)\}^2 f(x)dx$$

분산이 크면 편차제곱의 값이 평균적으로 크다는 의미이므로 실제 확률변수의 실현값이 기대값으로부터 멀리 떨어진 값으로 나타나는 가능성이 높음을 의미한다. 물론 분산이 작은 경우는 반대의 경향성을 기대한다. 즉, 분산은 확률변수에 내재된 이질성(heterogeneity) 혹은 불확실성(uncertainty)의 크기를 측정해주는 측도가 된다.

분산은 편차를 제곱한 값의 기대값이기 때문에 원래 확률변수 값과 같은 단위를 갖도록 하기 위해 제곱근을 적용한 값을 대신 사용하는 경우가 많다. 이를 표준편차(standard deviation)이라 하고 $\mathsf{sd}(X)$로 나타낸다.

$$\mathsf{sd}(X) = \sqrt{\mathsf{Var}(X)}$$

참고로

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2$$

이 성립함을 이용하면 분산을 보다 간단히 계산할 수 있음을 기억하자.

1.4.3 표준화

기대값과 분산을 각각 그리스 문자 μ와 σ^2로 간단히 나타내기도 한다. 표준편차는 자연스럽게 σ로 나타낸다. 확률변수 X에 대해

$$Z = \frac{X - \mu}{\sigma}$$

와 같은 변환을 적용하는 것을 표준화(standardization)라 한다. 표준화 결과 새로이 정의되는 확률변수 Z는

$$\mathsf{E}(Z) = 0, \quad \mathsf{Var}(Z) = 1$$

을 만족한다. 표준화는 서로 다른 확률분포를 갖는 확률변수들의 실현값을 상대적으로 비교할 때 유용하다.

예제 학생들의 영어과목 성적 X는 기대값이 60이고 표준편차가 10인 확률변수이고, 국어과목 성적 Y는 기대값이 70이고 표준편차가 15인 확률변수라 하자. 이 때 어느 학생의 영어 성적은 70점이고 국어 성적은 80점이었을 때 어느 과목

의 학업성취도가 더 높을 지 생각해보자. 이 학생의 영어 성적을 표준화한 값은 $(70-60)/10 = 1$이고, 국어 성적을 표준화한 값은 $(80-70)/15 = 2/3$이다. 즉, 실제 점수는 국어 성적이 높지만 두 과목의 성적 분포를 고려하면 영어 과목의 학업 성취도가 높은 것으로 판단된다. □

1.4.4 왜도와 첨도

왜도(skewness)는 확률분포의 비대칭성에 대한 측도로서 다음과 같이 정의된다.

$$\mathsf{Skewness}\,(X) = \mathsf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

확률분포가 대칭이면 왜도가 0이 되며, 오른쪽으로 꼬리가 길면 양수, 왼쪽으로 꼬리가 길면 음수가 된다. 첨도(kurtosis)는 확률분포의 뾰족한 정도와 꼬리의 두터운 정도를 측정해주는 대표값으로, 분포의 꼬리가 두껍고 뾰족한 성질이 강할수록 큰 값을 갖는다.

$$\mathsf{Kurtosis}\,(X) = \mathsf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

1.5 이변량 확률분포

예측분석의 대부분은 서로 다른 변수 사이의 종속성(dependence)을 이용한다. 보통 예측 대상 변수에 대한 정보를 담은 보조변수들의 패턴을 이용하기 때문이다. 따라서 서로 다른 변수 사이의 종속성을 고려한 확률모형에 대한 이해가 필수적이다.

두 이산형 확률변수 X와 Y에 대해 정의되는 확률분포

$$p(x,y) = P(X=x, Y=y)$$

를 결합확률분포(joint probability distribution)라 한다. 함수 $p(x,y)$를 결합확률밀도함수라 하는데,

$$p(x) = \sum_y p(x,y), \quad p(y) = \sum_x p(x,y)$$

가 성립한다. 두 확률변수 X와 Y가 다음을 만족하면 서로 독립(independent)이라 한다.

$$p(x, y) = p(x)p(y)$$

또한 $X = x$ 일 때 Y의 조건부확률분포는 다음과 같이 구한다.

$$P(Y = y|X = x) = \frac{p(x, y)}{p(x)} =: p(y|x)$$

1.5.1 공분산

두 확률변수 X와 Y 사이의 선형 종속성에 대한 측도인 공분산(covariance)은 다음과 같이 정의된다.

$$\mathsf{Cov}(X, Y) = \mathsf{E}\left[\{X - \mathsf{E}(X)\}\{Y - \mathsf{E}(Y)\}\right].$$

참고로,

$$\mathsf{Cov}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$$

임을 이용하면 보다 간단히 공분산을 계산할 수 있다.

1.5.2 상관계수

공분산은 두 확률변수의 척도(단위)에 따라 그 값이 크게 영향을 받으므로 표준화해 공분산을 계산해 선형 종속성의 측도로 사용하는 것이 나은 경우가 있다. 이렇게 표준화된 확률변수 사이의 공분산을 상관계수(correlation coefficient)라 하며 다음과 같이 계산한다.

$$\mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\mathsf{sd}(X)\mathsf{sd}(Y)}$$

참고로, 상관계수를 나타내는 기호로 보통 그리스 문자 ρ를 쓰는 경우가 많다.

2 주요 확률분포

특별한 구조나 특징이 있는 확률분포에 별도의 이름을 구별해 붙여 사용하면 의사소통 시 편리하다. 대표적 확률분포의 예들을 살펴보자.

2.1 베르누이분포, 이항분포

성공확률이 $\theta \in [0,1]$ 인 시행을 실시해 성공이 관측되면 1, 실패가 관측되면 0 의 값을 갖는 확률변수의 분포를 베르누이분포 (Bernoulli distribution) 이라 하고 다음과 같이 나타낸다.

$$X \sim \mathsf{Ber}(\theta)$$

확률질량함수는

$$p(x) = P(X = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1$$

이며, 기대값과 분산은 각각

$$\mathsf{E}(X) = \theta, \quad \mathsf{Var}(X) = \theta(1-\theta)$$

이 됨을 쉽게 확인할 수 있다.

성공확률이 θ 로 동일한 베르누이 시행을 n 번 독립적으로 실시할 때 성공횟수 X 의 분포를 이항분포라 하고 다음과 같이 표현한다.

$$X \sim \mathsf{B}(n, \theta)$$

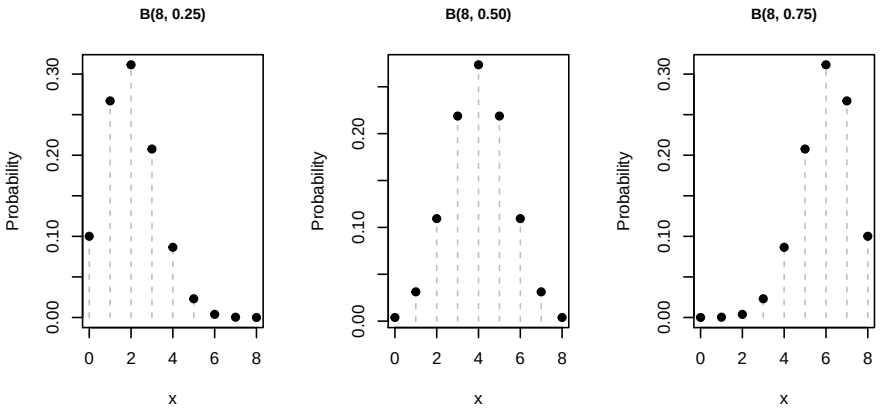
이항분포를 따르는 확률변수 X 의 확률질량함수는 다음과 같다.

$$p(x) = P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, ..., n$$

기대값, 분산은 각각

$$\mathsf{E}(X) = n\theta, \quad \mathsf{Var}(X) = n\theta(1-\theta)$$

이 된다. 베르누이 분포 $sfBer(\theta)$ 는 이항분포의 특별한 경우인 $\mathsf{B}(1, \theta)$ 이다.



예제 동전을 두 개 던질 때 둘 다 앞면이 나올 확률은 $\theta = 1/4$이다. 이러한 시행을 8회 실시해 둘 다 앞면이 나온 횟수를 X라 하면 $X \sim \mathsf{B}(8, 0.25)$가 되며, 기대값과 분산은 각각

$$\mathsf{E}(X) = 8 \times 0.25 = 2, \quad \mathsf{Var}(X) = 8 \times 0.25 \times 0.75 = 1.5$$

이다. X의 값이 3 이상이 될 확률은

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2) = 0.3216$$

이 된다.

```
1 - pbinom(2, size = 8, prob = 0.25)
```

```
## [1] 0.3214569
```

□

2.2 포아송 분포

계수(count) 데이터의 확률분포를 모형화하는 데 유용한 분포이다. 단위 시간(또는 공간) 내에 평균 발생 횟수가 $\lambda > 0$인 사건이 실제로 발생한 횟수를 X라 하면

$$X \sim \mathsf{Poi}(\lambda)$$

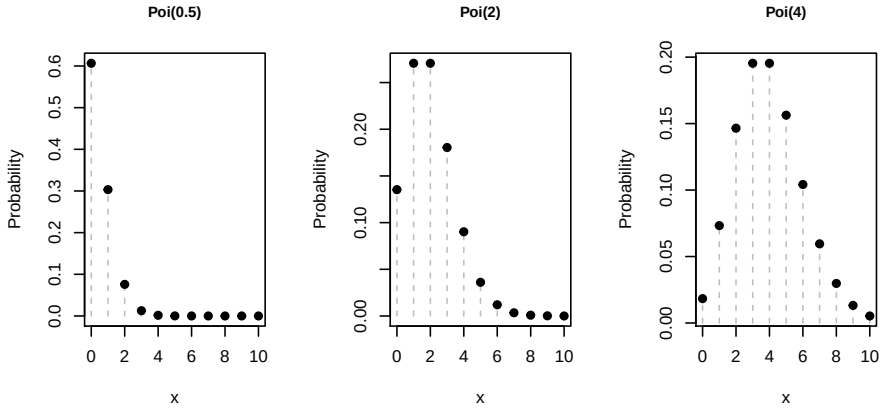
로 나타내며, 확률질량함수는

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ...$$

와 같다. 기대값, 분산은 각각

$$\mathsf{E}(X) = \lambda, \quad \mathsf{Var}(X) = \lambda$$

이 된다.



예제 경부고속도로 양재-오산 구간에서 하루 평균 0.2회의 교통사고가 일어난다고 한다. 이 구간에서 하루에 발생하는 교통 사고 횟수를 X라 하자. 내일 이 구간에서 2건 이상 교통사고가 발생할 확률을 예측해보면

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{0.2^0 e^{-0.2}}{0!} - \frac{0.2^1 e^{-0.2}}{1!} \approx 0.0175$$

이다.

```r
1 - ppois(1, lambda = 0.2)
```

```
## [1] 0.0175231
```

□

2.3 정규분포

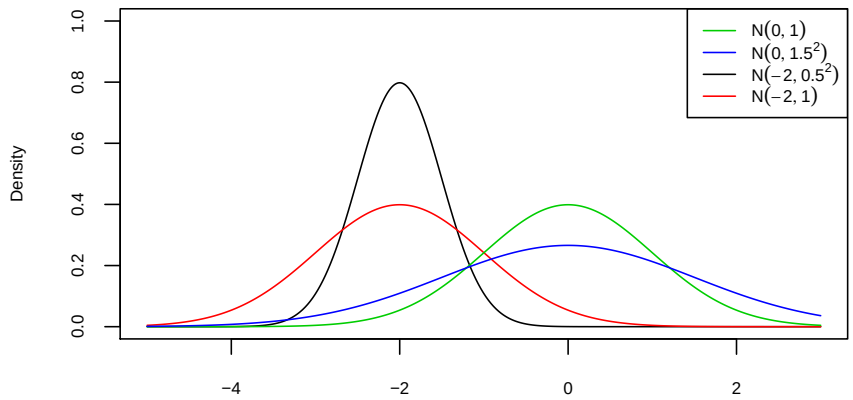
확률밀도함수가

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

인 연속형 확률변수의 분포를 평균이 μ이고 분산이 σ^2인 정규분포(normal distribution)라 하고 (기대값과 분산을 실제로 계산해보면 각각 μ, σ^2 임)

$$X \sim \mathsf{N}(\mu, \sigma^2)$$

와 같이 나타낸다. 고안한 사람의 이름을 따서 가우스 분포(Gaussian distribution)이라 부르기도 한다. 평균이 0이고 분산이 1인 정규분포를 따로 구별해 표준정규분포라 한다.



예제 초코과자를 생산하는 공장이 있다. 이 공장에서 생산하는 과자 한 봉지의 무게는 평균이 200g이고 표준편차가 1.5g인 정규분포를 따른다고 한다. 어느 날 생산된 과자 중 임의로 한 봉지를 선택해 무게를 재었을 때 198g 이하가 될 확률을 계산해보자. 과자 한 봉지의 무게를 확률변수 X로 나타내면 $X \sim \mathsf{N}(200, 1.5^2)$ 이므로

$$P(X \leq 198) = \int_{-\infty}^{198} \frac{1}{\sqrt{2\pi} \cdot 1.5} \exp\left\{-\frac{(x-200)^2}{2 \cdot 1.5^2}\right\} dx$$

를 계산하면 되는데, R 함수 `pnorm()`를 이용하면 다음과 같이 간단히 구할 수 있다.

```
pnorm(198, mean = 200, sd = 1.5)
```

```
## [1] 0.09121122
```

또는 X를 표준화한 $Z = \frac{X-200}{1.5}$의 분포가 표준정규분포임을 이용하면

$$P(X \leq 198) = P\left(\frac{X-200}{1.5} \leq \frac{198-200}{1.5}\right) = P(Z \leq -\frac{4}{3}) = \int_{-\infty}^{-4/3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

를 계산해도 된다.

```
pnorm(-4/3)
```

```
## [1] 0.09121122
```

□

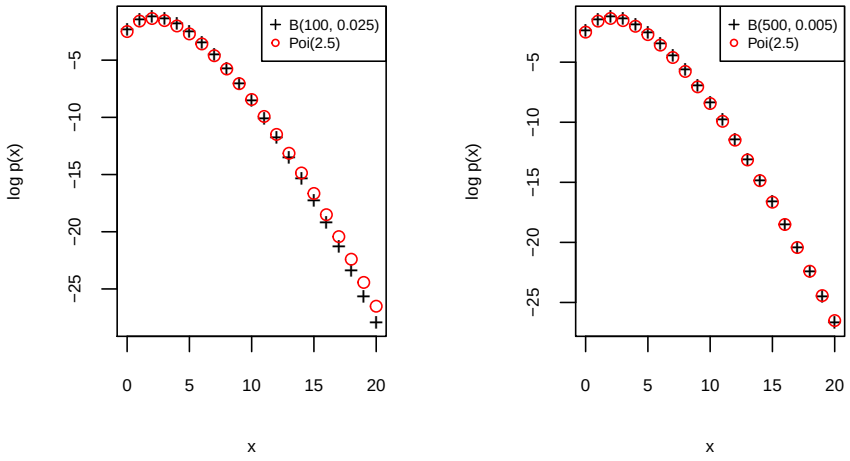
2.4 이항분포의 근사

$X \sim \mathsf{B}(n, \theta)$이라 하자. 이 때 n이 매우 큰 경우 X에 관련된 확률값을 계산하는 데 수치적으로 어려움을 만나게 된다. 따라서 정확히 확률을 계산하는 대신 근사적인 방법을 사용해 이 어려움을 해결하는 방안을 생각할 수 있다.

우선, n은 매우 큰데 θ는 거의 0에 가까운 값일 경우에는

$$\mathsf{B}(n, \theta) \approx \mathsf{Poi}(n\theta)$$

의 근사식이 성립하는데 이를 이항분포의 포아송 근사라 한다. 아래 그림은 $\mathsf{B}(100, 0.025)$와 $\mathsf{B}(500, 0.005)$를 $\mathsf{Poi}(2.5)$로 근사시킨 결과를 로그 스케일로 비교한 것이다. 근사의 정확도가 매우 우수함을 확인할 수 있다.

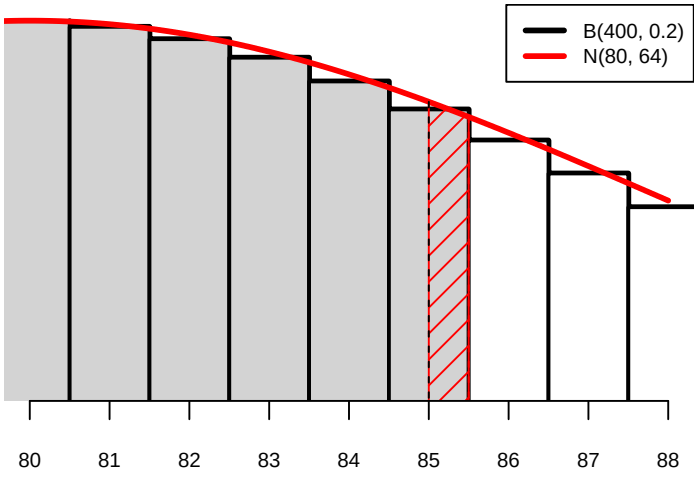


반면에 n은 매우 큰데 θ는 0에 그리 가깝지 않은 경우에는

$$\mathsf{B}(n, \theta) \approx \mathsf{N}(n\theta, n\theta(1-\theta))$$

의 근사식이 성립하는데 이를 이항분포의 정규 근사라 한다. 보통 $n\theta \geq 5$, $n(1-\theta) \geq 5$인 경우 사용하면 근사의 정밀도가 좋은 것으로 알려져 있다. 한 가지 주의할 점은 이산형 분포를 이산형 분포로 근사하는 포아송 근사와는 달리 정규 근사는 이산형 분포를 연속형 분포로 근사하는 방법이기 때문에 연속성 수정(continuity correction) 절차가 필요하다는 점에 유의해야 한다. 예를 들어 $X \sim \mathsf{B}(400, 0.2)$일 때 $P(X \leq 85)$를 정규 근사를 통해 계산하는 문제를 생각해 보자. $\mathsf{B}(400, 0.2) \approx \mathsf{N}(80, 8^2)$가 성립하므로, 새로운 확률변수 $Y \sim \mathsf{N}(80, 8^2)$을 도입해 $P(X \leq 85) \approx P(Y \leq 85)$와 같은 근사식을 생각할 수 있다. 그러나 아래 그림에서 보는 바와 같이 이 방법은

근사에 따른 오차가 상당한데(빗금 표시된 영역의 넓이만큼), 이는 이산형 분포를 연속형 분포를 사용해 근사하는 과정에서 흔히(또는 당연히) 발생하는 현상이다. 빨간 실선으로 표시된 정규분포 곡선을 85까지만 적분한 값인 $P(Y \leq 85)$ 보다는 85.5까지 적분한 값인 $P(Y \leq 85.5)$ 가 $P(X \leq 85)$(회색막대 넓이의 합)에 대해 훨씬 더 좋은 근사값이 됨을 알 수 있다.



2.5 카이제곱 분포, F-분포

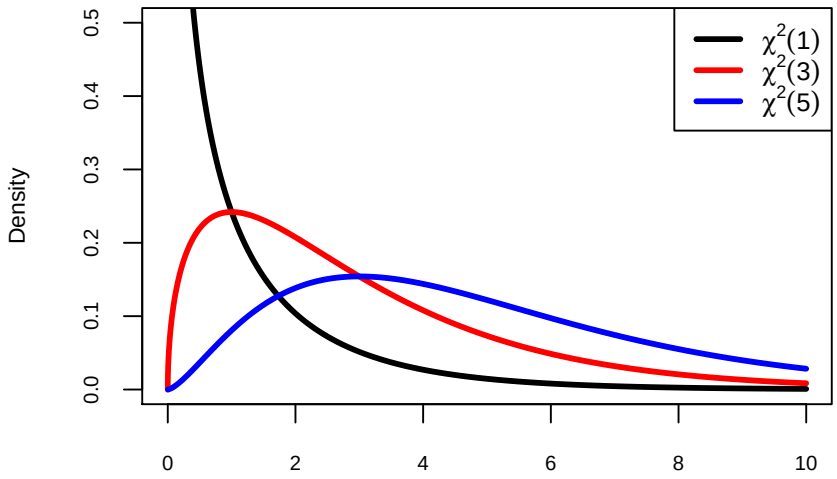
데이터의 제곱합에 관련된 두 분포를 소개하고자 한다. 서로 독립이고 표준정규분포를 따르는 확률변수들 $Z_1, Z_2, \cdots, Z_r$ 의 제곱합

$$V = Z_1^2 + Z_2^2 + \cdots + Z_r^2$$

이 따르는 분포를 자유도가 r 인 카이제곱분포(chi-squared distribution)이라 하고

$$V \sim \chi^2(r)$$

과 같이 나타낸다. 참고로 이 분포의 기대값과 분산은 각각 r, $2r$ 이다. 밀도함수의 형태는 오른쪽으로 꼬리가 긴 특징을 가지며, 자유도가 커질수록 봉우리의 위치가 오른쪽으로 이동하게 된다.



예제 $X_1, ..., X_n$ 이 서로 독립이고 $N(\mu, 1)$ 을 따를 때 제곱합 $\sum_{i=1}^{n}(X_i - \bar{X})^2$ 의 분포는 자유도가 $n-1$ 인 카이제곱분포가 된다. 단, $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ 이다. □

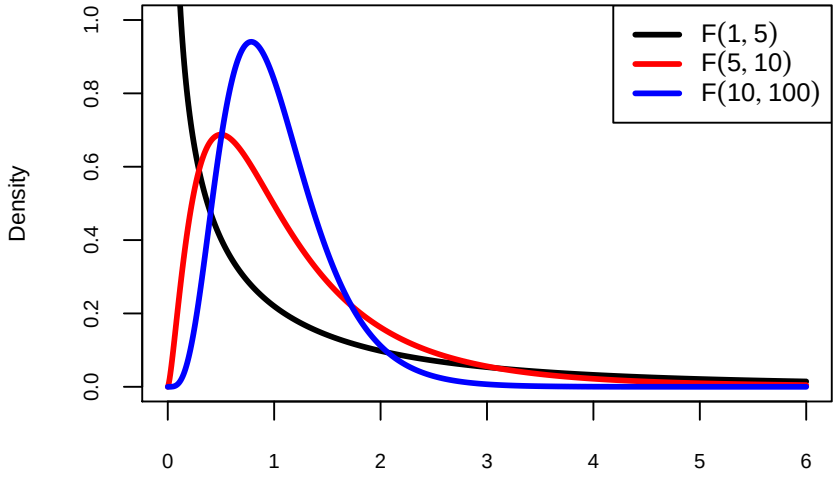
V_1, V_2 가 서로 독립이고 자유도가 각각 r_1, r_2 인 카이제곱분포를 따르는 확률변수일 때

$$F = \frac{V_1/r_1}{V_2/r_2}$$

가 따르는 분포를 F-분포라 하고

$$F \sim \mathsf{F}(r_1, r_2)$$

와 같이 나타낸다.

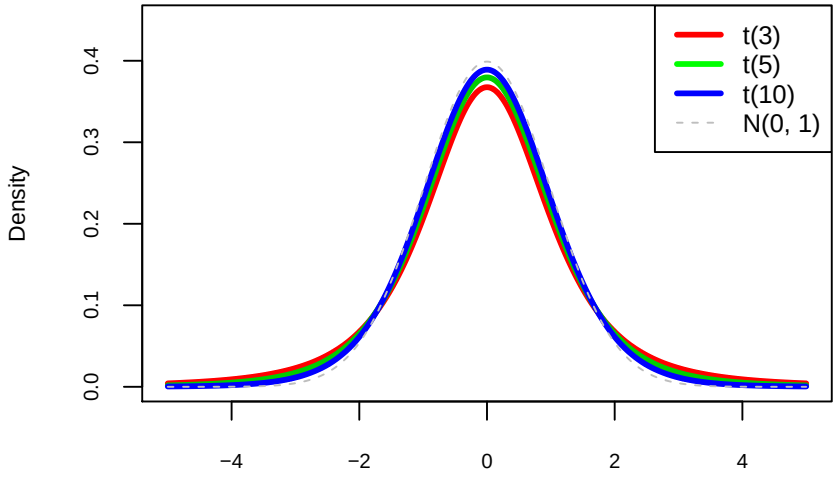


2.6 t-분포

$Z \sim N(0,1)$, $V \sim \chi^2(r)$ 이고 Z 와 V 가 서로 독립일 때

$$T = \frac{Z}{\sqrt{V/r}}$$

가 따르는 확률분포를 자유도가 r 인 t-분포라 하고, $T \sim t(r)$ 로 나타낸다. 자유도 r 이 커지면서 점점 표준정규분포 $N(0,1)$ 에 가까워지는 성질이 있다. 자유도가 작은 값인 경우 분포곡선의 형태가 표준정규분포에 비해 두터운 꼬리를 갖게 된다.



3 예측 분석의 기본 개념

3.1 오차의 원천

분석가가 추론 과정에서 맞닥뜨리는 오차는 크게 두 가지 성격으로 분류된다. 분석 방법론 내지 기술의 한계에 의해 발생하는 오차와 데이터가 내재하고 있는 불확실성에 기인한 오차이다. 전자를 체계적 오차, 후자를 비체계적 오차라 부른다. 분석가를 양궁 시합에 나선 선수에 비유하자면 체계적 오차는 훈련 프로그램, 활 또는 화살 성능 등에 의해 발생하는 오차로 이해할 수 있다. 그에 반해 비체계적 오차는 당일 날씨, 경기장 분위기 등에 의한 오차에 해당한다. 최신 알고리즘을 연구 개발하고 좋은 계산 장비를 확보하는 것은 체계적 오차를 줄이기 위한 노력이다. 반면에 데이터를 충분히 확보할 수 없는 경우, 또는 결측 및 오염이 있는 데이터를 사용해 분석해야 할 때 분석가가 맞닥뜨리는 어려움은 비체계적 오차와 관련되며

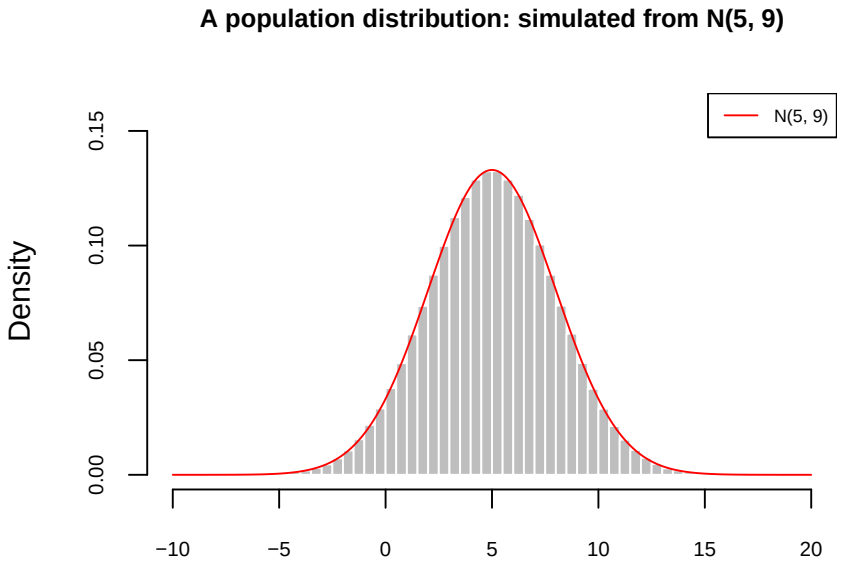
대부분의 경우 통제할 수 없는 영역에 속한 문제이기 때문에 확률 및 통계 이론의 역할이 중요해진다.

3.2 모집단, 표본, 표본분포

모집단(population)이란 분석 대상이 되는 개체(또는 해당 관측값) 모두를 다 모은 집합을 가리킨다. 우리가 실행하는 분석의 궁극적 목표는 모집단의 분포(distribution)를 규명하는 것이라 할 수 있다. 모집단 분포의 특성을 요약한 대표값을 모수(parameter)라 한다. 모집단 전체를 완벽히 파악하는 것이 현실에서는 불가능하기 때문에 일반적으로 모수의 값은 알려지지 않은 경우가 대부분이며, 따라서 추론의 대상이다. 모수의 값에 의해 모집단 분포의 특성이 완전하게 설명되는 모형을 모수적 모형이라 하며, 모수적 모형에 기반한 분석 방법을 모수적 방법이라 한다.

아래 그림은 모의실험으로 생성한 모집단의 히스토그램이다. 모의실험에 사용한 모형은 평균이 5이고 표준편차가 3인 정규분포이며, 모집단의 크기는 $N = 1,000,000$으로 설정했다.

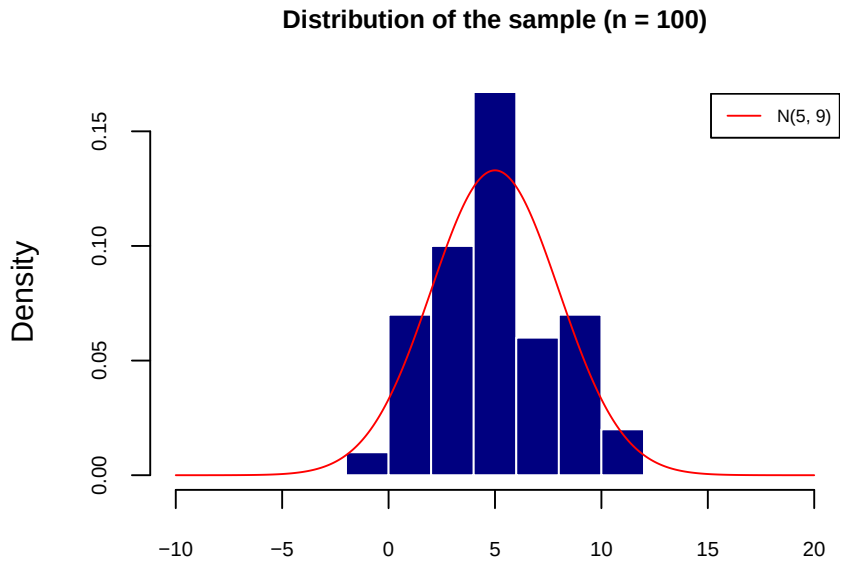
```
set.seed(1)
N <- 1000000
pop <- data.frame(id = 1:N, x = rnorm(N, mean = 5, sd = 3))
hist(pop$x, nclass = 100, probability = TRUE,
     col = "gray", border = "white",
     xlim = c(-10, 20), xlab = "", ylim = c(0, 0.16),
     main = "A population distribution: simulated from N(5, 9)",
     cex.main = 0.8, cex.axis = 0.7)
z <- seq(from = -10, to = 20, by = 0.005)
lines(z, dnorm(z, mean = 5, sd = 3), col = 2)
legend("topright", c("N(5, 9)"),
       col = c(2), lty = c(1), cex = 0.6)
```



모집단의 부분집합을 표본(sample)이라 한다. 모집단 전체를 조사해 데이터를 구하는 것이 어렵기 때문에 그 일부를 추출한 표본을 활용한다. 표본을 추출할 때 모집단을 잘 대표할 수 있는 방식으로 부분집합을 구성해야 한다. 가장 간단하고 이상적인 방식은 단순랜덤추출(simple random sampling)이다. 모수를 계산하는 절차를 흉내내어 표본에 적용한 버전을 통계량(statistic)이라 한다. 다음은 단순랜덤추출법으로 크기가 $n = 100$인 표본을 추출해 평균과 표준편차를 계산한 예이다.

```
n <- 100
sam <- pop[sample(1:N, n, replace = FALSE),] # simple random sampling
hist(sam$x, prob = TRUE, xlab = " ",
     xlim = c(-10, 20), ylim = c(0, 0.16),
     main = "Distribution of the sample (n = 100)",
     cex.main = 0.8, cex.axis = 0.7,
     col = "navy", border = "white")
```

```r
lines(z, dnorm(z, mean = 5, sd = 3), col = 2)
legend("topright", c("N(5, 9)"),
       col = c(2), lty = c(1), cex = 0.6)
```




```r
mean(sam$x)  # 모수값 = 5
```

```
## [1] 4.974753
```

```r
sd(sam$x)   # 모수값 = 3
```

```
## [1] 2.837773
```

위 예에서 확인할 수 있듯이 통계량의 값은 모수의 값과 차이를 보이기 마련이다. 그리고 위에서 생성한 모집단에서 크기가 100인 표본을 다시 추출해 평균과

표준편차를 계산한다면 조금 전의 것과 약간 다른 값을 얻게 될 것이다.

```
sam <- pop[sample(1:N, n, replace = FALSE),] # 두 번째 샘플
mean(sam$x)

## [1] 4.867998

sd(sam$x)

## [1] 3.378174

sam <- pop[sample(1:N, n, replace = FALSE),] # 세 번째 샘플
mean(sam$x)

## [1] 4.854398

sd(sam$x)

## [1] 3.031466
```

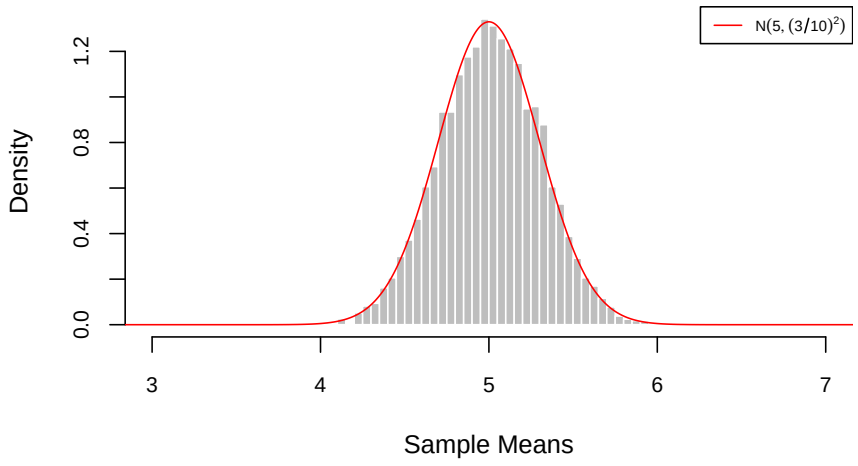
이렇듯 표본이 태생적으로 내포하고 있는 불확실성을 통계량이 물려받아 역시 불확실성을 갖게 된다. 이 불확실성을 분포로 표현한 것을 통계량의 표본분포(sampling distribution)라 한다. 많은 경우 통계학 이론을 통해 통계량의 표본분포를 이론적으로 알아낼 수 있다.

다음은 모의실험을 통해 위 예제에서 사용한 표본평균의 표본분포를 흉내낸 결과이다. 모의실험으로 생성한 표본평균들의 히스토그램과 이론적인 표본분포인 $\mathsf{N}(5, \frac{3^2}{100})$ 곡선이 매우 비슷함을 확인할 수 있다.

```
set.seed(0)
M <- 10000
m <- numeric(M)
```

```r
for ( r in 1:M ) {
  sam <- pop[sample(1:N, n, replace = FALSE),]
  m[r] <- mean(sam$x)
}

hist(m, probability = TRUE, xlab = "Sample Means",
     col = "gray", border = "white",
     xlim = c(3, 7), breaks = 50, main = "", cex.axis = 0.8)
lines(z, dnorm(z, mean = 5, sd = 0.3), col = "red")
legend("topright", expression(N(5, (3/10)^2)),
       col = c("red"), lty = c(1), cex = 0.6)
```

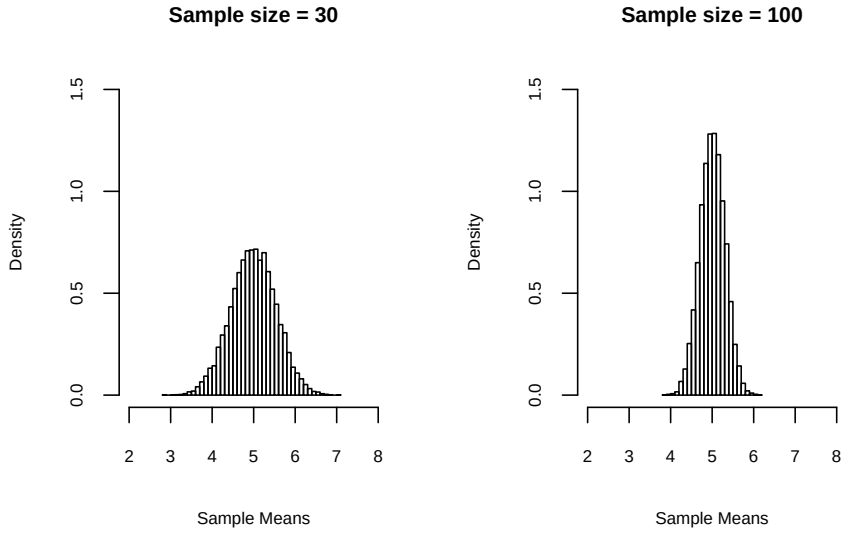


표본분포의 표준편차를 해당 통계량의 표준오차(standard error, s.e.)라 한다. 표본의 크기가 클수록 정보량이 많아지기 때문에 표준오차는 줄어들기 마련이다. 예를 들어 평균 계열의 통계량은 표본의 크기가 커짐에 따라 표본 크기의 제곱근에 반비례하는 방식으로 표준오차가 줄어든다.

```r
n <- 30
m.30 <- numeric(M)
for ( r in 1:M ) {
  sam <- pop[sample(1:N, n, replace = FALSE),]
  m.30[r] <- mean(sam$x)
}

round(c(sd(m.30)/sd(m), sqrt(100/30)), 4)

## [1] 1.8152 1.8257
```



표본분포는 데이터를 수집하는 전 단계에서 수집된 데이터를 통해 구현될 통계량의 값을 예상해 볼 수 있게 하는 예측분포(predictive distribution)로서의 의미를 갖는다. 위 예제를 살펴보면 표본 크기가 30이든 100이든 공히 히스토그램의 꼭대기점이 5 근방에 위치하고 있다. 즉, 모집단 분포가 $N(5, 3^2)$인 경우 표본을 추출해 평균을 구하면 그 값이 5 근처가 될 가능성이 매우 높다는 것을 예상할 수 있게 해준다. 또한 표본평균의 값이 5에서 멀리 떨어진 곳에서 실현될 가능성은 낮

으며, 표본의 크기가 클수록 그 경향성이 더 강해지기 때문에 그만큼 정밀한 추론이 가능해진다는 것을 예상할 수 있게 해 준다. 거꾸로, 모집단 분포의 평균값을 모르는 경우 데이터를 충분히 수집해 표본평균을 계산한다면 그 값이 모집단 평균값과 크게 차이가 나지 않을 것을 기대할 수 있게 해주는 결과이기도 하다.

앞에서 언급한 바와 같이 확률통계 이론을 이용하면 원 자료의 분포가 $N(5, 3^2)$일 때 표본평균의 분포가 $N(5, 3^2/n)$이 된다는 것을 이론적으로 증명할 수 있다. n의 값이 증가함에 따라 분산이 줄어든다는 것이다. 따라서 굳이 모의실험을 통하지 않더라도 위와 같은 추론이 가능하다. 물론 현실에서는 모집단의 분포를 알 수 없다는 문제가 추가되지만, 확률통계 이론을 이용해 자료가 추출된 모집단의 분포의 종류에 상관없이 통계량이 따르는 표본분포를 도출해 낼 수 있다. 대표적인 예가 중심극한정리(central limit theorem, CLT)로서, 표본 크기가 충분히 크면 모집단 분포의 종류에 상관없이 표본평균 통계량의 표본분포가 정규분포에 가깝다는 것은 잘 알려진 사실이다.

예제 어느 고등학교의 2학년에 재학 중인 학생들이 치른 영어시험 성적의 전체 평균은 60(점)이고 표준편차가 8(점)이라 한다. 이 학교에서 임의로 선택한 2학년 학생 100명의 영어시험 성적을 조사해 계산한 평균값이 62.5(점) 이상이 될 확률은 얼마나 될까? 중심극한정리에 따르면 100명의 성적을 평균한 값의 확률분포는 $\mathsf{N}(60, \frac{8^2}{100}) \equiv \mathsf{N}(60, 0.8^2)$과 비슷해진다. 따라서 정규확률변수 $Y \sim N(60, 0.64)$에 대한 확률 $P(Y \geq 62.5)$을 계산하면 구하고자 하는 확률에 대한 근사값을 구할 수 있다.

```
1 - pnorm(62.5, mean = 60, sd = 0.8)
```

```
## [1] 0.0008890253
```

□

모의실험 결과와 중심극한정리 둘 다 표본평균은 모평균 값 근처의 값을 취할 확률이 매우 높다는 것을 뒷받침해준다. 따라서 모평균의 값을 짐작해보고자 할 때 표본평균을 사용하는 것은 매우 합리적인 생각이라 할 수 있다.

4 기초적 예측 분석 : 평균 연비 예측

이 절에서는 R 내장데이터인 `mtcars`를 예제 데이터로 사용한다. 이 데이터셋은 1974년에 미국의 한 자동차 관련 잡지가 발표한 자료의 일부를 발췌한 것으로, 당시 32종의 자동차에 대한 디자인 및 성능에 대한 10가지 지표값으로 구성되어 있다.

```
mtcars
```

```
##                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
```

```
## AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird   19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2      26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa       30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L     15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino       19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora      15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E         21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

예제 데이터가 포함하고 있는 여러 변수 중 `mpg`(연비)와 `am`(변속기 종류 : 0 = 자동, 1 = 수동)을 분석에 사용하게 될 것이다.

다음 세 가지 연구질문에 대한 해답을 찾는 과정을 통해 예측 모형의 개념을 익혀보기로 하자.

1. 평균 연비는 얼마일까?
2. 변속기 종류에 따른 평균 연비의 차이는 얼마일까?
3. 자동변속기에 비해 수동변속기의 평균 연비가 좋다고 볼 수 있는가?

4.1 평균 연비 추정

우선 연비(`mpg`)의 평균값을 추정해보자. 이 데이터에는 겨우 32종의 자동차에 대한 정보만 포함되어 있기 때문에 이 세상에 존재하는 모든 자동차의 연비의 평균값을 데이터에 있는 연비의 평균값 즉 표본평균($\bar{x}$)의 값으로 추정하는 것이 다소 불안할 수 있다. 그러나 앞에서 언급한 표본분포의 개념 특히 중심극한정리에 따르면 표본평균은 모평균 근처에서 그리 멀지 않은 곳의 값이 될 가능성이 매우 높다는 것을 기대할 수 있으므로 32종 자동차의 연비를 평균한 값으로 이 세상에 존재하는 모든 자동차의 연비의 평균값을 추정하기로 하자.

```r
mean(mtcars$mpg)
```

```
## [1] 20.09062
```

이 추정치에 대한 불확실성을 표현하는 표준오차는 표본표준편차 s를 표본 크기의 제곱근 $\sqrt{n}$으로 나눈 값, 즉 $s/\sqrt{n}$으로 주어진다.

```r
print(n <- nrow(mtcars))
```

```
## [1] 32
```

```r
sd(mtcars$mpg)/sqrt(n)
```

```
## [1] 1.065424
```

4.2 변속기 종류와 평균 연비 사이의 관계 1

변속기 종류(`am`)에 따른 평균 연비의 차이를 추정해보자. 역시 표본평균의 차이 $\bar{x}_{\mathrm{a}} - \bar{x}_{\mathrm{m}}$의 값으로 평균 연비의 차이를 추정하는 것이 자연스럽다.

```r
print(n1 <- nrow(mtcars[mtcars$am == 0,]))
```

```
## [1] 19
```

```r
print(n2 <- nrow(mtcars[mtcars$am == 1,]))
```

```
## [1] 13
```

```r
print(m1 <- with(mtcars, mean(mpg[am == 0])))
```

```
## [1] 17.14737
```

```r
print(m2 <- with(mtcars, mean(mpg[am == 1])))
```

```
## [1] 24.39231
```

```r
m2 - m1
```

```
## [1] 7.244939
```

변속기별 평균 연비 추정치에 대한 표준오차를 계산하면 평균 연비의 차이에 대한 통계적 유의성을 평가할 근거가 생긴다.

```r
se1 <- with(mtcars, sd(mpg[am == 0]))/sqrt(n1)
se2 <- with(mtcars, sd(mpg[am == 1]))/sqrt(n2)
print(c(se1, se2))
```

```
## [1] 0.8795722 1.7102804
```

변속기별 평균 연비의 표준오차가 각각 $s_1/\sqrt{n_1} = 0.88$, $s_2/\sqrt{n_2} = 1.71$로서, 평균 연비 차이 추정치인 7.24에 비해 매우 작은 값임을 확인했다. 따라서 변속기별로 평균 연비는 차이가 있을 것으로 판단되며, 수동변속기를 장착한 자동차가 자동변속기를 장착한 경우보다 연비가 평균적으로 높은 것으로 보인다. 참고로 두 가지 변속기 사이의 평균연비의 차이를 표준오차로 나눈 것을 t-통계량이라 한다.

$$t = \frac{\bar{x}_{\mathrm{a}} - \bar{x}_{\mathrm{m}}}{\hat{\mathrm{se}}(\bar{x}_{\mathrm{a}} - \bar{x}_{\mathrm{m}})} = \frac{\bar{x}_{\mathrm{a}} - \bar{x}_{\mathrm{m}}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

```r
print(t.stat <- (m2 - m1)/sqrt(se2^2 + se1^2))
```

```
## [1] 3.767123
```

이 값은 두 변속기 간 평균 연비의 수학적 차이를 데이터에 내재된 불확실성에 대해 보정한 값으로서 다음에 살펴볼 가설 검정에서 중요하게 활용될 것이다.

4.3 변속기 종류와 평균 연비 사이의 관계 2

앞 절의 마지막 부분에서 세 번째 연구질문인 자동변속기의 평균 연비가 수동에 비해 높은 지 그렇지 않은 지에 대해 간단히 논의한 바 있다. 이 연구질문에 대한 답변을 보다 체계적이고 객관적인 방식으로 제시하는 방법에 대해 생각해보자. 이 문제는 다음의 두 가지 가설 중에 하나를 선택하는 가설검정(hypothesis tesing) 문제이다.

- 가설 1: 가속기 종류와 평균 연비는 무관하다
- 가설 2: 수동 자동차의 평균 연비가 자동가속기 자동차의 평균 연비에 비해 높다

가설 1과 같이 새로운 주장에 반하는 기존의 주장을 귀무가설(null hypothesis)라 하고 보통 H_0으로 표시한다. 그리고 가설 2처럼 새로이 주장하고자 하는 가설을 대립가설(alternative hypothesis)이라 하고 H_1으로 나타낸다. 이들 두 가설은 한 쪽이 참이면 다른 한 쪽은 반드시 거짓이기 때문에 어느 한 쪽을 채택(또는 기각)하는 의사결정을 실행하게 되면 그 결과는 표 4.1과 같이 요약된다.

	의사결정	
실제	H_0를 채택	H_0를 기각
H_0가 참	OK	오류 (제1종)
H_0가 거짓	오류 (제2종)	OK

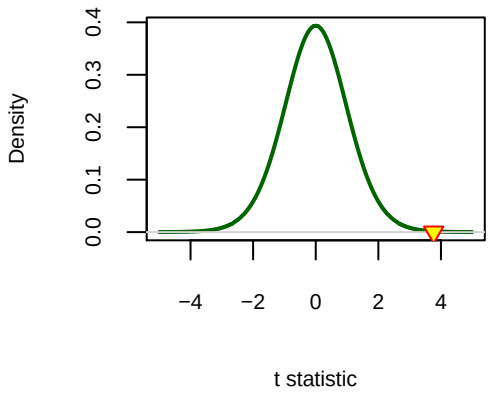
표 4.1: 가설 검정의 오류

통계적 가설검정에서는 두 가지 종류의 오류 중 제1종의 오류(Type-I error)를 범하게 될 확률 즉 귀무가설이 참인데 귀무가설을 기각하는 오류를 범할 확률에 대해 허용가능한 상한을 정해둔 상태에서, 제2종의 오류(Type-II error)를 범할 확률을 최소로 하는 규칙을 찾게 된다. 제1종의 오류를 범할 확률에 대한 허용상한을 유의수준(significance level)이라 하며 보통 α로 나타낸다.

$$P(\text{Reject } H_0 \,|\, H_0 : \mathsf{TRUE}) \leq \alpha$$

예제의 가설 검정은 데이터로 구한 변속기별 평균 연비의 차이에 기반해 시행하는 것이 가장 자연스러울 것이다. 다만, 단순히 변속기별 평균 연비의 수학적 차이를 사용하는 것보다는 데이터에 내재된 불확실성을 고려한 차이, 즉 앞 절에서 언급한 t-통계량을 사용하는 것이 합리적이다. 이처럼 가설 검정에 사용하는 통계량을 특별히 검정통계량(test statistic)이라 부른다. 이 예제의 경우 검정통계량의 값이 클수록 귀무가설보다는 대립가설을 지지하는 증거가 되므로, 검정통계량이 기준치 이상의 값을 가지면 귀무가설을 기각하는 것으로 검정 규칙을 수립하는 것이 자연스럽다. 이 때 사용하는 기준치를 기각치라 한다.

이 예제 데이터를 가지고 실제로 구한 검정통계량의 값은 3.7671이었다. 일단 양수이므로 수동변속기가 자동변속기에 비해 평균 연비면에서 유리한 것으로 보인다. 그러나 이 차이는 충분히 큰 값일까, 기각치를 얼마로 정하는 것이 좋을까에 대한 해답이 필요하다. 이를 위해 귀무가설이 참일 때 즉 평균 연비의 차이가 0일 때, 검정통계량 (t)의 값이 3.7671 이상이 될 확률은 얼마나 될 지, 즉 검정 규칙에 사용할 기각치로 3.7671을 쓸 때 제1종의 오류를 범할 확률은 얼마나 될 지 생각해보자. 이 확률을 계산하려면 귀무가설이 참일 때 검정통계량의 표본분포를 알아야한다. 통계이론에 의하면 이 예제의 경우 귀무가설이 참일 때 검정통계량 t의 분포는 자유도가 18.332인 t 분포(t distribution)가 된다. 이처럼 귀무가설이 참일 때 검정통계량이 데이터로 구한 값보다 극단적인 값을 가질 확률을 p-값(p-value)이라 한다. 즉 '데이터로 구한 t-값이 3.7671 이상이면 귀무가설을 기각한다'라는 가설 검정 규칙을 사용할 때 이 규칙이 제1종의 오류를 범할 확률에 해당한다.



아래 결과를 살펴보면 이 예제의 p-값은 0.00069(0.069%)로서 지극히 작은 값이다. 귀무가설이 참일 때(실제로는 평균 연비의 차이가 없을 때) '우연히' t-값이 3.7671 이상의 값이 될 확률이 이렇게 작은 값이라는 것을 의미한다. 그러나 우연에 의해 이런 사건이 발생했다고 보는 것은 합리적 판단이 아니다. 이러한 확률을 도출한 귀무가설을 기각하고 가설 2, 즉 대립가설을 채택하는 것이 합리적인 의사결정이다.

```
pt(t.stat, df = 18.332, lower.tail = FALSE)
```

```
## [1] 0.000686833
```

만약 위 결과와 달리 p-값이 0.3이었다면 무슨 의미가 있을까? 30% 정도의 확률로 발생할 가능성이 있는 사건을 경험하게 되면, 일상적으로 일어날 만한 일이 일어났으며 특별할 것 없다는 반응을 보이는 것이 보통이다. 따라서 p-값이 0.3과 같이 큰 값이었다면 귀무가설을 기각하기보다는 채택하는 쪽으로 결론을 내리는 것이 옳다. 반대로 p-값이 작으면 귀무가설을 기각하는 규칙을 사용하는 것이 자연스럽다. 문제는 p-값에 대한 기준치를 정하는 것인데, 제1종의 오류를 범할 확률에 대한 허용 상한 즉 유의수준 α를 사용한다. 보통 유의수준은 $\alpha = 0.05(= 5\%)$로

정하지만 문제의 특성에 따라 다르게 정할 수 있다. 이 예제의 경우 유의수준을 5%로 정하면 p-값이 이보다 작으므로 귀무가설을 기각, 즉 수동가속기의 연비가 자동가속기에 비해 평균적으로 좋다는 결론을 내리게 된다. 다음은 R의 `t.test()` 함수를 이용해 가설 검정을 실시한 예이다. 위에서 얻은 결과와 일치함을 확인할 수 있다.

```
with(mtcars, t.test(mpg[am == 1], mpg[am == 0],
                    alternative = "greater"))


##
##  Welch Two Sample t-test
##
## data:  mpg[am == 1] and mpg[am == 0]
## t = 3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.913256      Inf
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```