9.1 가설검정(Testing Hypotheses)

§. 귀무가설과 대립가설 (The Null and Alternative Hypotheses)

Definition 9.1.1 귀무가설과 대립가설 / 기각
통계적 가설의 종류로 H_0 : **귀무가설(null hypothesis)**, H_1 : **대립가설(alternative hypothesis)**라고 부른다. 그리고 모수공간 $\Omega = \Omega_0 \cup \Omega_1$, $\Omega_0 \cap \Omega_1 = \varnothing$ 에 대하여 모수가 $\theta \in \Omega_1$이면 H_0을 **기각(reject)한다**고 하고, $\theta \in \Omega_0$이면 H_0을 **기각할 수 없다(not to reject)**고 말한다.

▷ **예제 9.1.2 : Egyptian Skulls**
다양한 기간동안 이집트인들의 여러 두개골 측정 자료가 보고되었다고 한다. 그 중 어느 기간은 거의 기원전 4000년경에 측정되었다. 우리는 관찰된 두개골 너비의 측정값이 정규(normal) 확률변수로써 분산σ가 26이고 μ는 알려지지 않았다고 모형화하였다. 여기서 관심있는 것은 오늘날의 두개골 너비 측정값(140mm)과 어떻게 비교할 것인지 파악하고자 한다.

모수공간을 $\Omega = R^+$라 두고, $\Omega_0 = [140, \infty)$, $\Omega_1 = (0, 140)$이라 하자. 이 경우, 우리는 다음과 같이 가설을 세울 것이다.

$$H_0 : \mu \geq 140$$
$$H_1 : \mu < 140$$

좀 더 구체적으로, 우리는 알려지지 않은 두개골 측정값의 평균과 분산을 가정하였다고 하자. 이는 각 측정치가 μ와 σ^2를 갖는 정규확률변수라고 하는 것과 같다. 모수는 2차원 벡터로써 $\theta = (\mu, \sigma^2)$를 갖는다. 그러면 우린 오직 μ에 대해 관심 있으므로, $\Omega_0 = [140, \infty) \times (0, \infty)$, $\Omega_1 = (0,140) \times (0, \infty)$ 이다.
이제 우린 더 다양한 이론들을 살펴보면서 이 예제와 더불어 가설검정이 무엇인지 파악할 것이다.

▲

Definition 9.1.2 단순가설과 복합가설
$\theta \in \Omega_i$ 에 대하여 $\dim(\Omega_i) = 1$ 이면 H_i 는 **단순가설(simple hypothesis)** 이라고 부른다. 만약 $\dim(\Omega_i) > 1$ 이면 H_i 는 **복합가설(composite hypothesis)** 이라고 부른다.

Definition 9.1.3 단측가설과 양측가설
$\dim(\Omega) = 1$ 이라고 하자. $H_0 : \theta \leq \theta_0$ 또는 $H_0 : \theta \geq \theta_0$인 경우 각각 $H_1 : \theta > \theta_0$ 또는 $H_1 : \theta < \theta_0$이면 **단측가설(One-sided)** 이라고 부른다. $H_0 : \theta = \theta_0$ 인 경우 $H_1 : \theta \neq \theta_0$ 이면 **양측가설(Two-sided)** 이라고 부른다.

§. 임계역 및 기각역과 검정통계량 (Critical / Reject Region and Test Statistics)

Definition 9.1.4 임계역

$\vec{X}=(X_1,\ldots,X_n)$을 알려지지 않은 모수 θ를 갖는 분포로부터 취해진 확률표본이라 가정하자. 확률표본에서 얻는 데이터 벡터 $\vec{x}=(x_1,\ldots,x_n)$, 통계량 T에 대하여 $S_0=\{\vec{x}:-c\le T\le c\}$, $S_1=S_0^C$ 이라 정의하면 S_1을 **임계역(Critical Region)** 이라고 말한다. 이 때 상수 c는 선택될 수 있다.

▷ **예제 9.1.3 : Testing Hypotheses about Mean of a Normal Distribution with Known Variance.**

위의 정의와 마찬가지로 가정하자. 대신 모수가 $\theta=(\mu,\sigma^2)$인 정규분포를 갖는다고 가정하자. 다음과 같이 가설을 설정하고 이를 살펴보고자 한다.

$$H_0:\mu=\mu_0$$
$$H_1:\mu\neq\mu_0$$

우리는 $\overline{X_n}$이 μ_0과 얼마나 멀리 떨어져 있는지에 따라 H_0을 기각하고자 한다. 예를 들어, 상수 c를 선택하여 $\overline{X_n}$과 μ_0의 거리가 c보다 큰 경우 H_0을 기각할 것이다. 이를 표현하는 하나의 방법으로는 모든 가능한 데이터 벡터 $\vec{x}=(x_1,\ldots,x_n)$의 집합인 표본공간 S를 다음과 같이 두 개의 집합으로 분할하는 것이다.

$$S_0=\{\vec{x}:-c\le\overline{X_n}-\mu_0\le c\},\ S_1=S_0^C$$

그러면 우린 $\vec{X}\in S_1$일 때 H_0을 기각하고, $\vec{X}\in S_0$일 때 H_0을 기각할 수 없다. 이 과정을 표현하는 간단한 방법은 통계량 $T=\left|\overline{X}-\mu_0\right|$를 정의하고 $T\ge c$일 때 H_0을 기각한다고 정의하는 것이다.

▲

Definition 9.1.5 검정통계량 / 기각역

$\vec{X}$가 모수 θ에 의존하는 분포로부터 취해진 확률표본이라 하자. $T=r(\vec{X})$는 통계량이고, R은 실직선의 부분집합이라 하자. 가설 $H_0:\theta\in\Omega_0$, $H_1:\theta\in\Omega_1$에 대한 검정절차에서 명제 "$T\in R$일 때 H_0을 기각한다."를 만족하면 T 를 **검정통계량(test statistic)** 이라 부르고, R 을 검정의 **기각역(rejection region)** 이라 부른다.

▷ **예제 9.1.4 : Rain from Seeded Clouds.**

예제 9.1.1의 내용을 가정하자. 우린 26개의 seeded cloud가 정규확률변수로써 알려지지 않은 모수 $\theta=(\mu,\sigma^2)$를 갖는 log-rainfalls를 모형화 하였다. 우린 어찌 되었든 $\mu>4$인 것에 관심이 있는 상태이다. 이를 모수벡터의 관점에서 서술하면, $\{(\mu,\sigma^2):\mu>4\}$에 θ가 들어가도록 하는 것이다. 이제 가설을 세우면 $H_0:\mu\le 4$ vs $H_1:\mu>4$ 이다. 우린 검정통계량으로 이전 예제와 같은 것을 사용할 수도 있지만, $U=n^{1/2}(\overline{X_n}-4)/\sigma'$을 사용하고자 한다. 이 경우에 U가 크다면 H_0을 기각할 수 있다. 이는 U에서 대응되는 $\overline{X_n}$이 4와 비교해서 커지기 때문이다.

▲

§. 검출력 함수와 제 1종, 2종 오류 (The Power Function and Types of Error)

Definition 9.1.6 검출력 함수

δ를 검정절차라고 하자. 그러면 함수 $\pi(\theta|\delta)$를 검정 δ의 **검출력 함수(Power function)**라고 부른다. 만약 S_1이 δ의 임계역이면, 검출력함수 $\pi(\theta|\delta)$는 다음의 관계에 의해 결정된다.

$$\pi(\theta|\delta) = \Pr(\overrightarrow{X} \in S_1|\theta) \text{ for } \theta \in \Omega$$

만약 δ가 검정통계량 T와 기각역 R로 표현된다면, 검출력 함수는 다음의 관계에 의해 결정된다.

$$\pi(\theta|\delta) = \Pr(T \in R|\theta) \text{ for } \theta \in \Omega$$

▷ **예제 9.1.5 : Testing Hypotheses about Mean of a Normal Distribution with Known Variance.**

예제 9.1.3의 경우에서, 검정 δ는 검정통계량 $T = \left|\overline{X_n} - \mu_0\right|$과 기각역 $R = [c, \infty)$을 기반으로 하고 있다. $\overline{X_n}$의 분포는 평균 μ와 분산 σ^2/n을 갖는 정규분포를 따른다. μ는 우리가 σ^2를 알고 있다고 가정하였기 때문에 이 문제에서의 모수가 된다. 검출력 함수는 이러한 정규분포에서 계산될 수 있다, 우선 Φ를 표준정규분포의 누적분포함수(cdf)라 하자. 그러면 검출력 함수는 다음과 같이 구해진다.

$$\begin{aligned}\pi(\mu|\delta) = \Pr(T \in R|\mu) &= \Pr(\overline{X_n} \geq \mu_0 + c|\mu) + \Pr(\overline{X_n} \leq \mu_0 - c|\mu) \\ &= 1 - \Phi\left(n^{1/2}\frac{\mu_0 + c - \mu}{\sigma}\right) + \Phi\left(n^{1/2}\frac{\mu_0 - c - \mu}{\sigma}\right)\end{aligned}$$

Figure 9.1.은 $c = 1,2,3$인 경우의 검정에 대한 검출력 함수를 나타낸 것이다. 이 때, $\mu_0 = 4$, $n = 15$, $\sigma^2 = 9$ 로 두고 구하였다.

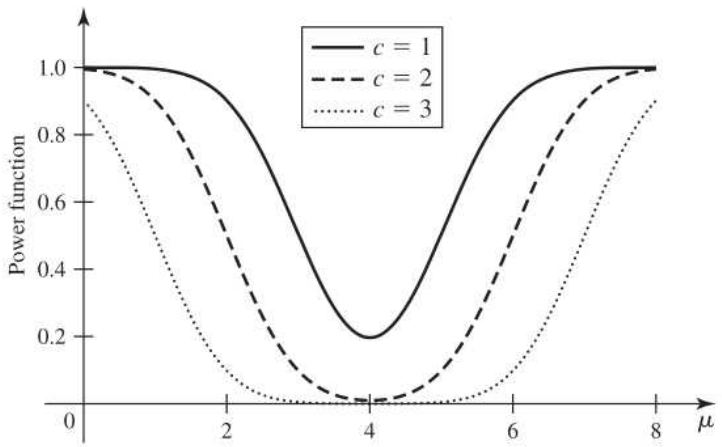


Figure 9.1. Power functions of three different tests in Example 9.1.5.

▲

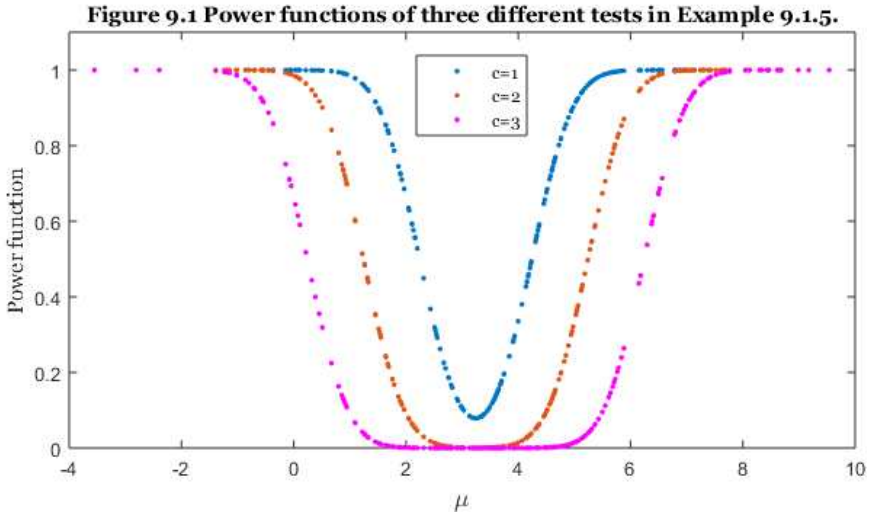


Figure 9.1. Power functions of three different tests using MATLAB.

MATLAB CODE - statistics2powerfunction.m

```
clear all
clc

% random samples & coefficients
samples=normrnd(4,3,1,250);
n=length(samples);
c=[1 2 3];
mu0 = 4;
sigma = 3;
% Pr(sample mean >= mu0+c | mean) and
in case of <=.
T1_c1=sqrt(n)*(mu0+c(1)-samples)./sig
ma;
T2_c1=sqrt(n)*(mu0-c(1)-samples)./sig
ma;
T1_c2=sqrt(n)*(mu0+c(2)-samples)./sig
ma;
T2_c2=sqrt(n)*(mu0-c(2)-samples)./sig
ma;
T1_c3=sqrt(n)*(mu0+c(3)-samples)./sig
ma;
T2_c3=sqrt(n)*(mu0-c(3)-samples)./sig
ma;
% c.d.f. of normal distribution (not
standard)
p1_1 = normcdf(T1_c1,mu0,sigma);
p2_1 = normcdf(T2_c1,mu0,sigma);
p1_2 = normcdf(T1_c2,mu0,sigma);
p2_2 = normcdf(T2_c2,mu0,sigma);
p1_3 = normcdf(T1_c3,mu0,sigma);
p2_3 = normcdf(T2_c3,mu0,sigma);
% Probability of T is  in R given mu
(Power func.)
P1 = 1-p1_1+p2_1;
P2 = 1-p1_2+p2_2;
P3 = 1-p1_3+p2_3;
% Graph of the power functions
plot(samples,P1,'.');
hold on
plot(samples,P2,'.');
plot(samples,P3,'.m');
title('Figure 9.1 Power functions of
three different tests in Example
9.1.5.');
xlabel('\mu');
ylabel('Power function');
legend('c=1','c=2','c=3');
xlim([-4 10]);
ylim([0 1.1]);
```

Definition 9.1.7 **제 1종 오류와 제 2종 오류**
귀무가설이 참일 때 이를 기각하는 것을 **제 1종 오류(type Ⅰ error)**라 말하고, 귀무가설이 거짓인데 이를 기각하지 않는 것을 **제 2종 오류(type Ⅱ error)**라고 말한다.

▷ **예제 9.1.6 : Egyptian Skulls.**
예제 9.1.2에서, 실험자들이 두개골 너비가 오랜 시간동안 증가해왔음을 언급하였다. 만약 μ가 기원전 4000년경의 두개골 너비의 평균이고, 140이 오늘날의 두개골 너비의 평균이라고 하면, 이론대로라면 $\mu < 140$이다. 실험자들은 사실 $\mu > 140$이거나 $\mu < 140$임에도 잘못된 주장으로 가지고 있는 데이터가 각각 그 반대의 경우로 주장할 수 있다. 이러한 경우를 각각 살펴보면, 첫 번째의 경우는 귀무가설이 참임에도 그와 반대를 주장함으로써 제 1종 오류를 범하게 되고, 두 번째의 경우는 귀무가설이 거짓인데도 불구하고 그 주장을 참이라 여기는 경우이므로 제 2종 오류를 범하게 된다. (예제 9.1.2에서 $H_0 : \mu \geq 140$, $H_1 : \mu < 140$ 이었다.)

▲

Definition 9.1.8 **수준 / 크기**
어떤 검정이 $\pi(\theta|\delta) \leq \alpha_0$ for all $\theta \in \Omega_0$, $\alpha_0 \in [0,1]$ 을 만족하면 이 검정을 **수준(level)** α_0 **검정**이라 말하고, 이러한 검정은 α_0 **유의수준(level of significance)**을 갖는다고 한다. 또한, **검정** δ**의 크기(size)** $\alpha(\delta)$는 다음과 같이 정의된다.

$$\alpha(\delta) = sup_{\theta \in \Omega_0} \pi(\theta|\delta)$$

▷ Corollary 9.1.1 ① 검정 δ가 수준 α_0 검정 $\Leftrightarrow$ $\alpha(\delta) \leq \alpha_0$. ② $H_0 : \theta = \theta_0 \Rightarrow \alpha(\delta) = \pi(\theta_0|\delta)$.

▷ **예제 9.1.7 : Testing Hypotheses about a Uniform Distribution.**
확률표본 $X_1, ..., X_n$이 구간 $[0,\theta]$인 균등분포로부터 취해진다고 가정하고, 모수 $\theta > 0$는 알려지지 않았다고 가정하자. 또한 다음의 가설을 검정한다고 가정하자.

$$H_0 : 3 \leq \theta \leq 4$$
$$H_1 : \theta < 3 \text{ or } \theta > 4$$

우린 이전의 예제로부터 θ의 최대우도추정량은 $Y_n = \max\{X_1, ..., X_n\}$임을 알고 있다. 비록 Y_n이 반드시 θ보다 작을지라도, Y_n은 표본크기 n이 상당히 커진다면 θ에 가까워질 높은 확률을 가지고 있다. 검정 δ가 $2.9 < Y_n < 4$인 경우 H_0을 기각하지 않는다고 하고, δ는 Y_n가 앞선 구간에서 있지 않을 때 H_0을 기각한다고 가정하자. 그러면 이 때 검정 δ의 임계역은 $Y_n \leq 2.9$ or $Y_n \geq 4$에 속하는 모든 $X_1, ..., X_n$을 포함한다. 이 때, 검정통계량 Y_n의 관점에서 기각역은 $(-\infty, 2.9] \cup [4, \infty)$이 된다.

δ의 검출력함수는 다음 관계로 규정된다.

$$\pi(\theta|\delta) = \Pr(Y_n \leq 2.9|\theta) + \Pr(Y_n \geq 4|\theta)$$

만약 $\theta \leq 2.9$이면, $\Pr(Y_n \leq 2.0|\theta) = 1$이고 $\Pr(Y_n \geq 4|\theta) = 0$이므로, $\theta \leq 2.9$인 경우 $\pi(\theta|\delta) = 1$이다.

만약 $2.9 < \theta \leq 4$이면, $\Pr(Y_n \leq 2.9|\theta) = (2.9/\theta)^n$이고 $\Pr(Y_n \geq 4|\theta) = 0$이다. 이 경우 검출력함수는 $\pi(\theta|\delta) = (2.9/\theta)^n$이 된다. 마지막으로, $\theta > 4$이면 $\Pr(Y_n \leq 2.9|\theta) = (2.9/\theta)^n$이고 $\Pr(Y_n \geq 4|\theta) = 1-(4/\theta)^n$이다. 이 경우, $\pi(\theta|\delta) = (2.9/\theta)^n + 1 - (4/\theta)^n$이 된다. 이러한 검출력함수의 그래프는 Figure 9.2에서 확인할 수 있다.

식 $\alpha(\delta) = sup_{\theta \in \Omega_0}\pi(\theta|\delta)$에 의하여, δ의 크기는 $\alpha(\delta) = sup_{3 \leq \theta \leq 4}\pi(\theta|\delta)$이 된다. 이는 Figure 9.2에서 확인할 수 있고, 그 계산은 단지 $\alpha(\delta) = \pi(3|\delta) = (29/30)^n$ 으로 주어진다. 특히, 만약 표본크기가 $n = 68$이면, δ의 크기는 $(29/30)^{68} = 0.0997$이 된다. 따라서, δ는 유의수준$\alpha_0 \geq 0.0997$인 수준 α_0 검정이다.

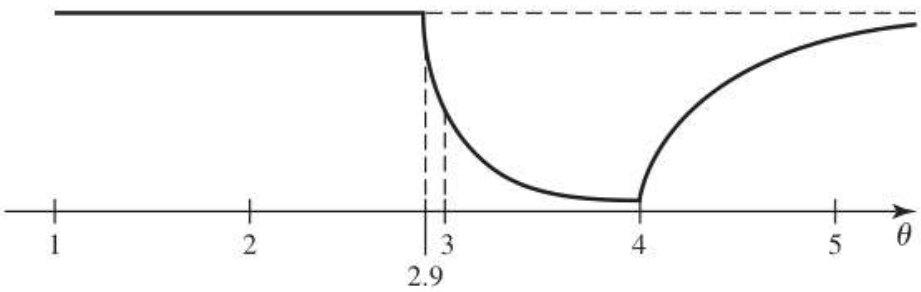


Figure 9.2. The power functions $\pi(\theta|\delta)$ in Example 9.1.7.

▲

▷ **예제 9.1.8 : Testing Hypotheses about Mean of a Normal Distribution with Known Variance.**
이전 예제 9.1.5에서 다룬 검정은, $\left|\overline{X_n}-\mu_0\right| \geq c$ 인 경우 $H_0 : \mu = \mu_0$ 을 기각하는 것이었다. $Y = \overline{X_n}-\mu_0$ 가 $\mu = \mu_0$인 경우 평균이 0이고 분산이 σ^2/n 인 정규분포를 가지므로, 우리는 각 α_0에 대하여 정확히 α_0의 크기를 가지도록 만들 값 c를 찾을 수 있다. Figure 9.3은 Y의 확률밀도함수(pdf)를 보여주고 있고, 검정의 크기도 밀도함수곡선 아래로 색칠된 영역으로 표시되어 있음을 확인할 수 있다.
이 정규확률밀도함수가 평균에서 대칭성을 가진다는 사실에 의하여, 두 개의 색칠된 영역들은 같은 넓이, 즉 $\alpha_0/2$를 갖는다. 이것은 c가 반드시 Y의 분포의 $1-\alpha_0/2$ 분위수에 위치한다는 것을 의미한다. 이러한 분위수는 $c = \Phi^{-1}(1-\alpha_0/2)\sigma n^{-1/2}$ 로 구해진다.

정규분포의 평균에 대한 가설 검정을 할 때, 이는 관습적으로 다음과 같이 통계량의 관점에서 이 검정을 재 서술할 수 있다.

$$Z = n^{1/2}\frac{\overline{X_n}-\mu_0}{\sigma}$$

그러면 이 검정은 $|Z| \geq \Phi^{-1}(1-\alpha_0/2)$ 일 때 H_0을 기각한다.

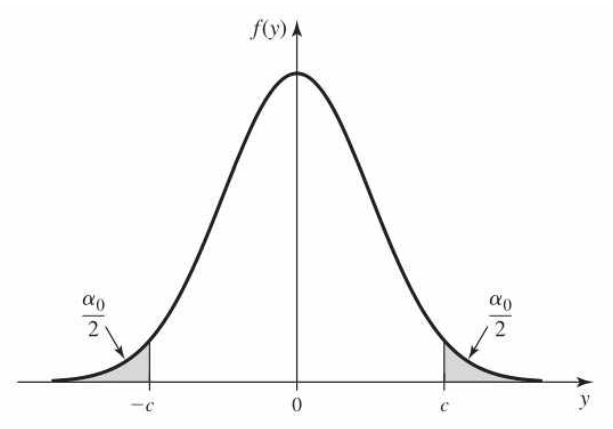


Figure 9.3. The p.d.f. of $Y = \overline{X_n}-\mu_0$ given $\mu = \mu_0$ for Example 9.1.8. The shaded areas represent the probility that $|Y| \geq c$.

▲

▷ **예제 9.1.9 : Testing Hypotheses about a Bernoulli Parameter.**

$X_1, \ldots, X_n$이 모수가 p인 베르누이 분포로부터 확률표본을 생성한다고 가정하자. 우리는 다음 가설에 대한 검정을 하길 원한다.

$$H_0 : p \le p_0$$
$$H_1 : p > p_0$$

$Y = \sum_{i=1}^{n} X_i$라 하자, 이는 $Y \sim Bin(n,p)$이다. 우리는 $p > p_0$임을 주장하고자 하므로 우린 $Y \ge c$인 경우 H_0을 기각하는 것을 택한다. 또한 우리는 가능한 한 α_0을 넘지 않고 α_0에 가까워지도록 하는 검정의 크기를 원한다고 하자. $\Pr(Y \ge c|p)$가 p의 함수로써 증가함수임은 쉽게 알 수 있고, 이런 이유로 검정의 크기는 $\Pr(Y \ge c|p = p_0)$ 이 된다. 그래서 c는 $\Pr(Y \ge c|p = p_0) \le \alpha_0$을 만족하는 가장 작은 수가 되어야 한다.

예를 들어, 만약 $n = 10$, $p_0 = 0.3$, $\alpha_0 = 0.1$이라 하면, 우린 c를 결정하기 위해 교재의 부록의 이항 확률표를 사용할 수 있다. 또한 $\sum_{y=6}^{10} \Pr(Y = y|p = 0.3) = 0.0473$, $\sum_{y=5}^{10} \Pr(Y = y|p = 0.3) = 0.1503$ 으로 계산할 수 있다. 검정의 크기를 많아야 0.1로 유지하기 위해서, 우린 반드시 $c > 5$가 되도록 해야 한다. 구간 $(5, 6]$의 모든 c값은, Y가 오직 정수값만 취하므로 같은 검정을 만들어내게 된다.

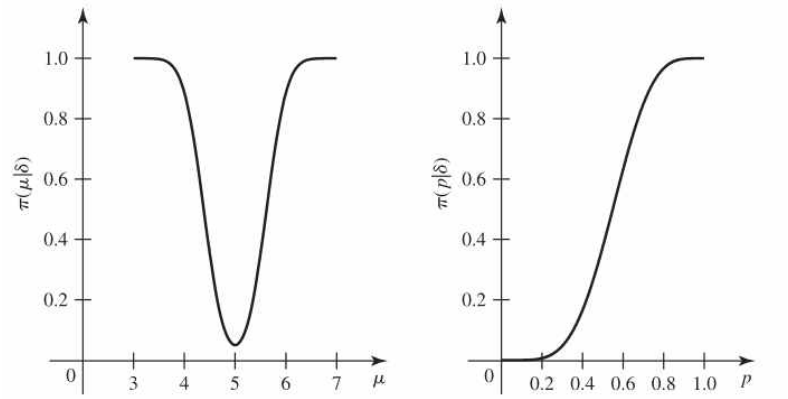


Figure 9.4. Power functions of two tests. The plot on the left is the power function of the test from Example 9.1.8 with $n = 10$, $\mu_0 = 5$, $\sigma = 1$, and $\alpha_0 = 0.05$. The plot on the right is the power function of the test from Example 9.1.9 with $n = 10$, $p_0 = 0.3$, and $\alpha_0 = 0.1$

▲

§. p-값 (The p-value)

Definition 9.1.9 **p-값 (유의확률, significance probability / p-value)**
일반적으로, **p-값 (p-value)**은 주어진 검정통계량의 값으로부터 H_0을 기각하게 하는 최소의 유의수준 α_0을 말한다. 이는 귀무가설이 참일 때 표본이 대립가설 방향으로 검정통계량의 값보다 더 어긋나게 될 확률을 의미한다.

▷ **[NOTE]** p-value $\leq \alpha_0$: 유의수준 α_0에서 H_0을 기각한다. (통계적으로 유의하다)
p-value $> \alpha_0$: 유의수준 α_0에서 H_0을 기각할 수 없다. (통계적으로 유의하지 않다)

▷ **[NOTE] 가설검정의 절차**

1. 귀무가설과 대립가설을 설정한다.
2. 주어진 문제의 특성에 따라 유의수준 α_0을 결정한다.
3. 표본 데이터로부터 검정통계량을 계산한다.
4. 가설검정의 결론 도출
   (i) 기각역 사용
      - 유의수준에 따라 기각역을 구한다.
      - 검정통계량이 기각역에 속하면 H_0을 기각한다.
   (ii) p-value 사용
      - 검정통계량을 이용하여 p-value를 구한다.
      - $p-value \leq \alpha_0$이면 H_0을 기각한다.

▷ **예제 9.1.10 : Testing Hypotheses about Mean of a Normal Distribution with Known Variance.**
앞선 예제 9.1.8에서, 우린 수준 $\alpha_0 = 0.05$ 에서의 귀무가설을 선택하였다. 이제 우린 검정통계량 $Z = n^{1/2}\dfrac{\overline{X_n} - \mu_0}{\sigma}$ 을 계산하고 $Z \geq \Phi^{-1}(1-0.05/2) = 1.96$인 경우 H_0을 기각할 것이다. 예를 들면, $Z = 2.78$로 구해졌다고 가정하자. 그러면 우린 H_0을 기각할 수 있다. 이러한 결과를 우리는 수준 0.05에서 H_0을 기각했다고 보고하기로 가정하자. 다른 유의수준에서 귀무가설을 좀 더 적절하게 검정하고자 하는 어떤 또 다른 통계학자는 이러한 보고를 그대로 참조하여도 되는가?

▲

▷ **예제 9.1.11 : Testing Hypotheses about Mean of a Normal Distribution with Known Variance.**
이전 예제에서 검정통계량은 $Z = 2.78$로 계산되었다. 이는 $2.78 \geq \Phi^{-1}(1-\alpha_0/2)$를 만족하는 모든 유의수준 α_0에 대하여 귀무가설을 기각할 수 있는 것으로 알려졌다. 본 교재의 부록에 있는 정규분포표를 살펴보면, 이러한 부등식은 $\alpha_0 \geq 0.0054$로 변환될 수 있다. 여기서 값 0.0054는 관측된 데이터와 검정된 가설에 대한 $p-value$라고 부른다. 0.01>0.0054 이므로, 수준 0.01에서의 가설을 검정하길 원하는 통계학자들 또한 H_0을 기각할 수 있다.

▲

▷ **[NOTE] 유의확률(p-value) 계산하기**

통계량 T에 대한 단순검정에서 "$T \geq c$일 때 귀무가설을 기각한다."의 형태를 가지면, 직관적인 방법으로 유의확률을 계산할 수 있다.

각 t에 대하여, δ_t를 $T \geq t$이면 H_0을 기각하는 검정(test)라고 하자. 그러면 $T = t$일 때의 유의확률은 검정 δ_t의 크기로 구해진다. 즉, 유의확률은 다음과 같다.

$$sup_{\theta \in \Omega_0} \pi(\theta | \delta_t) = sup_{\theta \in \Omega_0} \Pr(T \geq t | \theta)$$

일반적으로, $\pi(\theta|\delta_t)$는 Ω_0와 Ω_1 사이의 경계에 놓인 어떤 θ_0에서 최대화된다. 유의확률은 T의 분포의 위쪽 꼬리에서의 확률로 계산되기 때문에, 이는 때때로 꼬리영역(a tail area) 라고 불린다.

▷ **예제 9.1.12 : Testing Hypotheses about a Bernoulli Parameter.**

예제 9.1.9에서의 가설 검정을 고려하자. 우린 $Y \geq c$이면 H_0을 기각하는 검정을 사용할 것이다. $Y = y$가 구해졌을 때의 p-값은 $sup_{p \leq p_0} \Pr(Y \geq y | p)$ 로 구해진다. 이 예제에서, $\Pr(Y \geq y|p)$는 p의 함수로, 증가함수임을 쉽게 알 수 있다. 이런 이유로 p-값은 $\Pr(Y \geq y | p = p_0)$이다. 예를 들어, $p_0 = 0.3$이고 $n = 19$이라 하자. 만약 $Y = 6$으로 구해졌다면, $\Pr(Y \geq 6 | p = 0.3) = 0.0473$으로 구해진다.

*** p-값의 계산은 "$T \geq c$일 때 귀무가설을 기각한다."인 검정이 아닌 경우에 계산이 더욱 복잡해진다. 우리는 본 교재에서 p-값은 오직 이러한 형태의 검정에서만 계산하는 것을 권장한다.

▲

§. 동등성 검정과 신뢰 집합 (Equivalence of Tests and Confidence Sets)

Definition 9.1.10. 신뢰집합
만약 확률 집합(random set) $\omega(\vec{X})$이 $\Pr\left[g(\theta_0)\in\omega(\vec{X})|\theta=\theta_0\right] \geq \gamma$ for $\forall\,\theta_0\in\Omega$을 만족하면 이를 $g(\theta)$ **에 대한 계수** γ **의 신뢰집합(coefficient** γ **confidence set for** $g(\theta)$**)** 라고 부른다. 만약 위의 부등식이 모든 θ_0에 대하여 등식이 성립하면 우리는 이를 신뢰집합이 **정확하다(exact)**고 부른다.

▷ **예제 9.1.15 : Constructing a Test from a Confidence Interval.**
정규분포에서의 알려지지 않은 평균과 분산에 대하여 신뢰구간을 구하는 방법은 이전 Section에서 논의하였다. 그러한 논의들에 이어서, 다음을 가정해보고자 한다.

$$X_1,\ ...,X_n \sim N(\mu,\sigma^2),\ \mu,\sigma^2 : \text{unknown},\ \theta=(\mu,\sigma^2),\ g(\theta)=\mu.$$

이전 Section의 방법으로, 우리는 다음의 통계량들을 사용할 것이다.

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i\ ,\ \sigma'=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2}$$

그러면, $g(\theta)$에 대한 계수 γ의 신뢰구간은 다음과 같다.

$$I=\left(\overline{X}_n-T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{\sqrt{n}},\overline{X}_n+T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{\sqrt{n}}\right)\ ,\ \text{where}$$

$T_{n-1}^{-1}(\bullet)$: the quantile function of the t-distribution with $n-1$ degrees of freedom.

각 μ_0에 대하여, 우리는 다음 가설에 대한 수준 $\alpha_0=1-\gamma$ 검정을 구하기 위해 위 구간을 사용한다.

$$H_0:\mu=\mu_0$$
$$H_1:\mu\neq\mu_0$$

이 가설에 대한 검정은 $\mu_0\in I$ 인 경우 H_0을 기각한다. 약간의 대수적 조작을 가하면, 다음의 명제가 동치임을 보일 수 있다.

$$\mu_0\notin I \Leftrightarrow \left|\sqrt{n}\frac{\overline{X}_n-\mu_0}{\sigma'}\right| \geq T^{-1}\left(\frac{1+\gamma}{2}\right)$$

이 검정은 사실 t 검정과 동일하다는 것을 Section 9.5에서 좀 더 면밀하게 다룰 것이다.

▲

§. 우도비 검정 (Likelihood Ratio Tests)

Definition 9.1.11 우도비 검정

$$\text{통계량 } \Lambda(\vec{x}) = \frac{sup_{\theta \in \Omega_0} f_n(\vec{x}|\theta)}{sup_{\theta \in \Omega_0} f_n(\vec{x}|\theta)}$$

을 **우도비 통계량(likelihood ratio statistic)** 이라 하고, 가설에 대한 우도비 검정은 아래와 같다.

$$\Lambda(\vec{x}) \leq k \text{ for some constant } k \Rightarrow H_0 \text{ 을 기각한다.}$$

▷ **예제 9.1.18 : Likelihood Ratio Test of Two-Sided Hypotheses about a Bernoulli Parameter.**
Y : 알려지지 않은 모수 θ에 대한 n개의 독립적인 베르누이 시행에서 성공(successes)의 개수, 라고 하고 이를 관측한다고 가정하자. 가설 $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ 을 고려하자. $Y = y$로 관측되었을 때, 우도함수는 다음과 같이 구해진다.

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

이 경우, $\Omega_0 = \{\theta_0\}$, $\Omega = [0,1]$ 이다. 우도비 통계량은 아래와 같이 구할 수 있다.

$$\Lambda(y) = \frac{\theta_0^y(1-\theta_0)^{n-y}}{sup_{\theta \in [0,1]}\theta^y(1-\theta)^{n-y}}$$

이 우도비 통계량의 분모에서, 상한(supremum)값은 최대우도추정량을 찾는 방법을 고려하여 구하면 찾을 수 있다. 그러한 상한값, 즉 최댓값은 θ가 최대우도추정량 $\hat{\theta} = y/n$와 같을 때 구해지며, 자세한 과정은 생략한다. 따라서 우도비 통계량은 다음과 같이 구해진다.

$$\Lambda(y) = \left(\frac{n\theta_0}{y}\right)^y\left(\frac{n(1-\theta_0)}{n-y}\right)^{n-y}.$$

이 우도비 통계량은 y가 0과 n에 가까워질 때 작아지고, $y = n\theta_0$ 에 가까워질 때 그 값이 커진다. 특정한 예로써, $n = 10$이고 $\theta_0 = 0.3$이라 가정하자. 그러면 $y = 0,...,10$일 때의 우도비 통계량을 구할 수 있다.

y	0	1	2	3	4	5	6	7	8	9	10
$\Lambda(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	0.034	0.005	3×10^{-4}	6×10^{-6}
$\Pr(Y=y\|\theta=0.3)$	0.028	·0.121	0.233	0.267	0.200	0.103	0.037	0.009	0.001	1×10^{-4}	6×10^{-6}

▲

9.5 T 검정(The t-test)

§. T 검정의 여러 성질(Properties of the t-test)

▷ **예제 9.5.1~3 : Nursing Homes in New Mexico.**

뉴 멕시코 요양원에서의 입원기간에 대한 연구를 실시하였다고 가정하자. 우리는 표본의 크기가 $n=18$인 알려지지 않은 모수 μ, σ^2를 갖는 정규확률변수로써 입원기간 수에 대한 모형화를 시행하였다. 우린 다음과 같은 가설에 관심이 있다.

$$H_0 : \mu \geq 200 \text{ vs } H_1 : \mu < 200$$

Introduction of T test

9.5장에서 우리는, 정규분포의 모평균과 모분산을 모를 때, 평균에 대한 가설검정의 문제를 고려하고자 한다. 즉, 모평균과 모분산이 알려지지 않았을 때의 정규확률변수 $X_1, ..., X_n$에 대하여 다음과 같은 가설을 검정하고자 한다.

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

이 때, 모수공간 Ω 는 $-\infty < \mu < \infty$, $\sigma^2 > 0$인 모든 2차원 벡터 (μ, σ^2)로 구성되어 있다. 귀무가설 H_0은 $\mu \leq \mu_0$를 만족하는 벡터가 $(\mu, \sigma^2) \in \Omega_0 \subset \Omega$ 임을 의미하고, 대립가설 H_1은 $\mu > \mu_0$을 만족하는 벡터가 $(\mu, \sigma^2) \in \Omega_1 = \Omega_1^C \subset \Omega$임을 의미한다. 이전 예제에서 보였듯이, 이 가설은 μ에 대한 단측검정(one-sided)에 해당한다. 검정통계량은 다음의 식을 사용한다.(이 식들은 정규분포에 대한 최대우도추정량에서 유래한다.)

$$\text{표본평균 } \overline{X}_n = \sum_{i=1}^{n} \frac{X_n}{n}, \text{ 표본표준편차 } \sigma' = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}{\sqrt{n-1}}, \text{ 검정통계량 } U = \frac{\overline{X}_n - \mu_0}{\frac{\sigma'}{\sqrt{n}}}.$$

검정은 $U \geq c$일 때 H_0를 기각한다. $\mu = \mu_0$ 일 때, 정리 8.4.2 에 의하면, 통계량 U의 분포는 σ^2의 값과 관계없이 자유도 $n-1$인 t-분포를 따른다. 이런 이유로, U에 대한 이러한 검정은 "t-검정"이라고 부른다. 앞서 소개한 가설과 달리 가설이 $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$ 이면 이는 마찬가지의 논의에 의해, 검정은 $U \leq c$일 때 H_0를 기각하면 된다.

α_0 검정을 통해 이를 검정하고자 한다면, 우린 검정통계량 U 가 검정의 크기가 α_0와 같도록 만드는 c를 택하여 이와 많아야 같을 때, H_0을 기각하도록 하는 T 검정을 사용할 수 있다. 유의수준 $\alpha_0 = 0.1$로 두자. 그러면 우린 $U \leq c$, c : 자유도 17인 t-분포의 0.1 분계선(quantile)≡-1.333 일 때 H_0을 기각한다. 교재의 예제 8.6.3의 데이터를 사용하여 표본 평균을 구하면 $\overline{X}_{18} = 182.17$이고, 표본표준편차는 $\sigma' = 72.22$이다. 따라서 $U = \frac{182.17 - 200}{\frac{72.22}{\sqrt{17}}} = -1.018$ 로 구해진다. 그러면 $U = -1.018 > -1.333$ 이므로 우리는 유의수준 0.1에서 귀무가설 $H_0 : \mu_0 \geq 200$을 기각할 수 없다.

▲

Theorem 9.5.2 t-검정의 p-값

가설 $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ 또는 $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$ 을 검정한다고 가정하자. u를 통계량 U의 관측값이라 하고, $T_{n-1}(\bullet)$을 $n-1$ 자유도인 t-분포의 누적분포함수(c.d.f.)라고 하자. 그러면 각 가설에 대한 p-값은 $1 - T_{n-1}(u)$, $T_{n-1}(u)$ 이다.

▷ **예제 9.5.4 : Length of Fibers.**

밀리미터인 금속섬유의 길이가 알려지지 않은 모수 μ, σ^2를 갖는 정규분포를 따른다고 가정하자. 우리는 다음 가설을 검정할 것이다.

$$H_0 : \mu \leq 5.2 \text{ vs } H_1 : \mu > 5.2$$

15개의 금속섬유의 길이가 무작위하게 선택되고 측정되었다고 가정하고, 표본평균이 $\overline{X}_{15} = 5.4$, $\sigma' = 0.4226$으로 측정되었다고 가정하자. 이러한 값들에 기초하여, 우리는 유의수준 $\alpha_0 = 0.05$에서의 T검정을 실시할 것이다.

$n = 15$, $\mu_0 = 5.2$ 이므로, 검정통계량 U는 자유도가 14이고 $\mu = 5.2$인 t-분포를 따른다. 이는 t-분포표를 참조하면, $T_{14}^{-1}(0.95) = 1.761$로 구해진다. 이런 이유로, 귀무가설은 $U > 1.761$일 때 기각될 수 있다. U에 대한 수치적 값은 1.833으로 계산되므로, H_0은 유의수준 0.05에서 기각될 수 있다.
통계량 U에 대한 관측값 $u = 1.833$과 $n = 15$에서 우리는 컴퓨터 소프트웨어를 이용하여 t-분포의 누적분포함수인 p-값을 계산할 수 있다. 이 가설에 대한 p-값은 $1 - T_{14}(1.833) = 0.0441$로 구해진다.

▲

▷ **예제 9.5.7 : Crash Test Dummies.**

국가 교통안전위원회는 자동차 충돌에 대한 더미의 손상량과 위치에 관한 충돌 테스트 데이터를 수집한다. 더미를 운전자 좌석, 탑승자 좌석에 배치하고 각 경우 충돌 시 더미의 머리 손상량을 측정하였다. 이들에 대한 측정값과 그 관계는 아래의 Figure 9.13에 제시하였다.

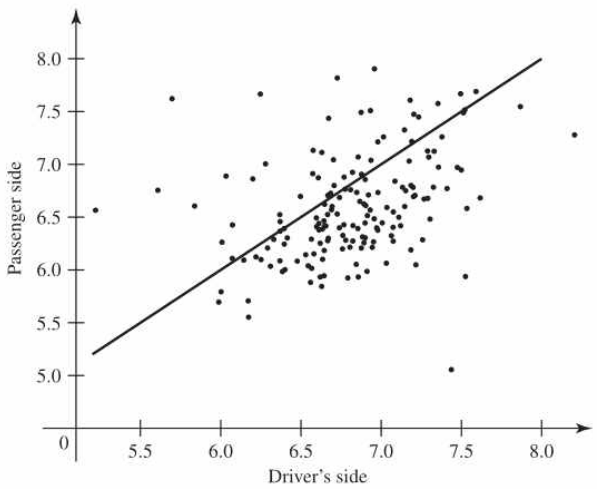


Figure 9.13 Plot of logarithms of head injury measures for dummies on driver's side and passenger's side. The line indicates where the two measures are equal.

확률변수 $X_1, \ldots, X_n$을 각 경우의 더미의 머리 부상 측정치 사이의 차이라고 하자. 우린 $X_1, \ldots, X_n$를 평균μ와 분산 σ^2를 갖는 정규분포로부터 취해지는 확률표본으로 모형화할 수 있다. 우린 다음과 같은 가설을 세우고 유의수준 $\alpha_0 = 0.01$에서 검정하고자 한다.

$$H_0 : \mu \leq 0 \ \text{ vs } \ H_1 : \mu > 0$$

표본크기가 $n = 164$개로 정해졌을 때, 이 가설은 $U \geq T_{163}^{-1}(0.99) = 2.35$ 일 때 귀무가설을 기각할 수 있다.

Figure 9.13에서 각 좌표의 차이에 대한 평균은 $\overline{x_n} = 0.2199$, 표본표준편차는 $\sigma' = 0.5342$로 구해졌다. 그러면 검정통계량 $U = 5.271$로 구할 수 있다. 이는 명백히 2.35보다 큰 값이므로, 귀무가설은 유의수준 0.01에서 기각할 수 있다. 추가적으로 사실, p-값은 1.0×10^{-6}보다 작음을 확인할 수 있다.

▲

▷ **예제 9.5.9 : Egyptian Skulls.**

유의수준 $\alpha_0 = 0.05$에서 가설 $H_0 : \mu = 140$ vs $\mu \neq 140$에 대한 검정을 실시하고자 한다. 이는 양측검정(Two-sided)에 해당하는 가설이므로, 앞선 논의에 의한 단측검정을 각각 다른 방향의 부등호로 실시하여야 한다. 즉 c_1과 c_2를 구하여 이에 맞는 기각역을 설정해주어야 한다. **T검정에서 양측검정인 경우, $|U| \geq T_{n-1}^{-1}(1-\alpha_0/2)$ 일 때 H_0을 기각함을 반드시 기억하길 바란다.** $c_1 = -T_{29}^{-1}(0.975) = -2.045$이고 $c_2 = 2.045$로 구해진다. $n = 30$에 대한 표본평균은 $\overline{X}_{30} = 131.37$로, 표본표준편차는 $\sigma' = 5.129$로 구해진다. 따라서 검정통계량 U의 관측값 $u = \dfrac{131.37 - 140}{\dfrac{5.129}{\sqrt{30}}} = -9.219$로 구해진다. 이는 -2.045보다 작으므로 유의수준 0.05에서 귀무가설 H_0을 기각할 수 있다.

▲

9.6 두 정규분포의 평균 비교 (Comparing the Means of Two Normal Distributions)

§. 이표본 T 검정(The Two-Sample t-test)

Theorem 9.6.1 이표본 t-통계량

두 표본공간에 대한 표본평균과 표본분산을 다음과 같이 가정하자.

$$\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i \ , \ \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S_X^2 = \sum_{i=1}^{m}(X_i - \overline{X}_m)^2, \ S_Y^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

그러면 검정통계량은 다음과 같이 정의된다.

$$U = \frac{(\overline{X}_m - \overline{Y}_n)}{\sqrt{\frac{\left(\frac{1}{m}+\frac{1}{n}\right)(S_X^2 + S_Y^2)}{m+n-2}}}$$

$\mu_1 = \mu_2$를 만족하는 모든 $\theta = (\mu_1, \mu_2, \sigma^2)$에 대하여, U의 분포는 자유도 $m+n-2$를 갖는 t-분포이다.

▷ **Theorem 9.6.3 이표본 T 검정에서의 p-값 $\Leftrightarrow$ $df = m+n-2$임을 제외하고 정리 9.5.2와 같다.**

▷ **예제 9.6.3 : Roman Pottery in Britain.**

과거 Great Britain의 다양한 지역에서 발견된 로마시대 양식 도기 표본에 대한 연구에 대하여 이야기하고자 한다. 각 도자기 표본에 대한 측정 수단 중 하나로 표본의 알루미늄 산화물에 대한 백분율을 고려하자.

우린 두 개의 다른 지역에서 발굴된 도기 표본에 대한 알류미늄 산화물의 백분율에 관심이 있다고 가정하자. 각각의 표본 크기는 $m = 14$, $n = 5$로, 표본평균과 표본분산은 각각 $\overline{X}_m = 12.56$, $S_X^2 = 24.65$, $\overline{Y}_n = 17.32$, $S_Y^2 = 11.01$로 구해졌다. 우린 두 개의 다른 평균 μ_1, μ_2와 같은 분산 σ^2을 갖는 정규확률변수로써의 데이터를 모형화한다고 가정하자. 그러면 U의 관측값 $u = -6.302$로 구해진다.

t-분포표를 참조하면, 자유도는 $m+n-2 = 17$이고, $T_{17}^{-1}(0.995) = 2.898$, $U < -2.898$임을 알 수 있다. 따라서, 우린 임의의 유의수준 $\alpha_0 \geq 0.0005$에서 H_0을 기각할 수 있다. 이 때, U에 대한 p-값은 $T_{17}(-6.302) = 4 \times 10^{-6}$로 구해짐을 확인할 수 있다.

▲

§. T검정에서의 양측검정 (Two-sided Alternatives)

t-검정에서의 양측검정 (One sample & Two sample)
유의수준 α_0에서 가설이 양측검정의 형태일 때, 즉 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$일 때, 기각역은 다음과 같이 구해진다.

$$|U| \geq c \text{ where } c = T_k^{-1}(1-\alpha_0/2),\ U : \text{검정통계량},\ k = \begin{cases} n-1 & One-sample \\ m+n-2 & Two-sample \end{cases}.$$

검정통계량 U가 $U = u$로 관측되었을 때, p-값은 다음과 같이 구해진다.

$$p-value = 2\left[1 - T_k(|u|)\right]\ ,\ k = \begin{cases} n-1 & One-sample \\ m+n-2 & Two-sample \end{cases}$$

▷ **예제 9.6.5 : Comparing Copper Ores.**
어떤 특정한 지역의 구리광산에서 얻어지는 8개의 광석 견본에 대한 확률 표본을 가정하자, 여기서 각 견본에서의 구리의 양은 gram 단위로 측정되었다. 이러한 8개의 견본을 $X_1, \ldots, X_8$로 두고, 이에 대한 표본평균과 표본분산은 $\overline{X}_8 = 2.6$과 $S_X^2 = 0.32$로 구하였다.

앞서 설명한 지역이 아닌 다른 지역의 광산에서 얻어지는 10개의 광석 견본에 대한 확률 표본을 가정하자, 여기서 각 견본에서의 구리의 양은 마찬가지로 gram단위로 측정되었고, 이러한 10개의 견본을 $Y_1, \ldots, Y_{10}$으로 두어 $\overline{Y}_{10} = 2.3$, $S_Y^2 = 0.22$로 구하였다. μ_1을 첫 번째 광산에서의 구리의 양에 대한 평균, μ_2를 두 번째 광산에서의 구리의 양에 대한 평균으로 두자. 우리가 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

우린 모든 이러한 모든 관측값들이 정규분포를 따르고, 만약 평균의 값이 서로 다르더라도, 두 광산 모두 같은 분산을 가진다고 가정할 것이다. 앞선 논의에 의하면, 표본크기는 $m = 8$, $n = 10$이고, 검정통계량 U는 $U = 3.442$로 구해진다. 또한 자유도 16인 t-분포표를 참조하면, $T_{16}^{-1}(0.995) = 2.921$임을 알 수 있고, 이는 검정통계량의 관측값과 연관된 꼬리 영역이 2×0.005보다 작음을 확인할 수 있다. 이런 이유로, 귀무가설은 임의의 유의수준 $\alpha_0 \geq 0.01$에서 기각될 수 있다. (사실, $U = 3.442$의 양측검정 꼬리영역은 0.003이다.)

▲

9.7 F 분포(The F Distributions)

§. F 분포의 정의 (Definition of the F Distribution)

Definition 9.7.1 F 분포

Y와 W가 독립적인 확률변수로써 $Y \sim \chi^2_m$, $W \sim \chi^2_n$, $m,n \in Z_+$ 을 만족한다고 하자. 그러면 새로운 확률변수 X는 다음과 같이 정의할 수 있다.

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW}$$

그러면 이러한 X의 분포를 **자유도 m,n을 갖는 F 분포(F distribution)**라고 부른다.

▷**Theorem 9.7.1 : p.d.f. of F distribution**

For $X = x$, p.d.f. $f(x)$ is as follows : $f(x) = 0$ for $x \leq 0$ and

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \left(\frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} \right) \text{ for } x > 0$$

Theorem 9.7.2 F 분포의 자유도

확률변수 X와 Y에 대하여 다음이 성립한다.

$$X \sim F_{m,n} \Rightarrow \frac{1}{X} \sim F_{n,m}$$

$$Y \sim T_n \Rightarrow Y^2 \sim F_{1,n}$$

▷ **예제 9.7.2 : Determing the 0.05 Quantile of an F Distribution.**

확률변수 X가 자유도를 6과 12로 갖는 F 분포를 따른다고 가정하자. 우린 X의 0.05 분계선을 결정하고자 한다. 즉, 값 x에 대하여 $\Pr(X < x) = 0.05$를 결정하고자 함이다.

만약 $Y = 1/X$로 두면, Y는 정리 9.7.2에 의해 자유도 12와 6을 갖는 F 분포를 따른다. 이는 교재의 F 분포표를 참조하면 $\Pr(Y \leq 4.00) = 0.95$임을 확인할 수 있다. 이런 이유로, $\Pr(Y > 4.00) = 0.05$이다.
$Y > 4.00 \Leftrightarrow X < 0.25$이므로, 이는 $\Pr(X < 0.25) = 0.05$임을 얻는다. 이 때, F 분포는 연속분포이므로, $\Pr(X \leq 0.25) = 0.05$이고, 0.25는 X의 0.05분계선에 위치한다.

▲

§. 두 정규분포의 분산의 비교 (Comparing the Variances of Two Normal Distributions)

Definition 9.7.2 F 검정

확률변수 $X_1,...,X_m$이 알려지지 않은 모수 μ_1, σ_1^2을 갖는 정규분포로부터 m개의 관측값인 확률표본을 생성한다고 가정하자. 그리고 확률변수 $Y_1,...,Y_n$ 또한 마찬가지로, 독립적인 확률표본으로써 알려지지 않은 모수 μ_2, σ_2^2를 갖는 정규분포로부터 n개의 관측값인 확률표본을 생성한다고 가정하자. 마지막으로, 다음 가설을 유의수준 $0<\alpha_0<1$에서 검정하고자 한다고 가정하자.

$$H_0 : \sigma_1^2 \le \sigma_2^2 \text{ vs } H_1 : \sigma_1^2 > \sigma_2^2$$

각 검정 절차 δ에 대하여, $\pi(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2 \mid \delta)$를 δ의 검출력함수(power function)이라 두자. 각 확률변수에 대한 표본분산 S_X^2와 S_Y^2를 다음과 같이 정의하자.

$$S_X^2 = \sum_{i=1}^{m}(X_i - \overline{X}_m)^2,\ S_Y^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

그러면 $S_X^2/(m-1)$과 $S_Y^2/(n-1)$은 각각 σ_1^2과 σ_2^2의 추정량(estimator)이 된다. 이제 여기서 V를 다음과 같이 정의하고, c에 대한 조건을 다음과 같이 주면 F 검정의 귀무가설 H_0을 기각할 수 있다.

$$\text{검정통계량 } V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}$$

$V \ge c$, c is chosen to make the test have a desired level of significance.

F 검정에서의 양측검정

유의수준 α_0에서 가설이 양측검정의 형태일 때, 즉 $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \ne \sigma_2^2$일 때, 기각역은 다음과 같이 구해진다.

$$V \le c_1 \text{ or } V \ge c_2 \text{ s.t. } \Pr(V \le c_1) + \Pr(V \ge c_2) = \alpha_0$$
$$\Rightarrow \Pr(V \le c_1) = \Pr(V \le c_2) = \alpha_0/2 \text{ when } \sigma_1^2 = \sigma_2^2$$

c_1 : $\alpha_0/2$ quantiles of the appropriate F distribution
c_2 : $1-\alpha_0/2$ quantiles of the appropriate F distribution

Theorem 9.7.5 p-value of Equal-Tailed Two-Sided F Test.

가설이 양측검정, 즉 $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \ne \sigma_2^2$일 때, δ_{α_0}을 F 검정에서의 양측검정이라 하자. 그러면 $V=v$로 관측되었을 때 H_0을 기각하는 δ_{α_0}에 대한 가장 작은 유의수준 α_0는 다음과 같다.

$$p-value = 2\min\{1 - G_{m-1,n-1}(v),\ G_{m-1,n-1}(v)\},\ G : \text{c.d.f. of } F \text{ distribution.}$$

▷ **예제 9.7.4 : Rain from Seeded Clouds.**

이전 예제 9.6.2에서, 우리는 seeded와 unseeded clouds로 나뉘는 log-rainfalls의 평균을 각 경우의 분산이 같다는 가정하에 비교하였다. 이제 우리는 이러한 두 분산이 서로 다른 경우에 대한 검정을 고려하고자 한다. 즉, 가설을 다음과 같이 설정하고자 한다. 이 때, 유의수준 $\alpha_0 = 0.05$로 둔다.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs } H_1 : \sigma_1^2 \neq \sigma_2^2$$

F 검정의 정의에 따르면, $m = n$이므로, 검정통계량 $V = \frac{63.96}{67.39} = 0.9491$로 구해진다. 우린 이를 자유도를 25와 25로 갖는 F 분포의 0.025와 0.975 분계수를 비교할 필요가 있다. 우리가 가진 F 분포표의 분계수는 자유도 25일 때의 행과 열을 갖지 않으므로, 우린 20과 30의 자유도 사이를 보간하여 구하거나 컴퓨터 프로그램을 이용하여 이러한 분계수를 구해야만 한다. 그러한 분계수는 0.4484와 2.2303으로 구해진다. V는 이러한 두 수 사이의 값이므로, 우린 유의수준 $\alpha_0 = 0.05$에서 귀무가설을 기각할 수 없다.

▲

10.1 적합도 검정 (Test of Goodness-of-Fit)

§. 비모수적 문제의 서술 (Description of Nonparametric Problems)

Definition 비모수적 문제와 비모수적 방법
추정, 검정에 대한 문제 또는 이를 아우르는 통계적 문제들에서, 관측값들(observations)에 대한 가능한 분포가 특정 분포로 제한되지 아니한 경우 이를 **비모수적 문제(Nonparametric problem)** 라고 부르고, 이러한 문제에 적용할 수 있는 방법을 **비모수적 방법(Nonparametric method)**라고 부른다.

▷ **(예제 1)** 어떤 관측값들이 포아송 분포를 따르는 확률표본을 생성한다고 가정하자. 그러면 우리는 포아송 분포의 모수 λ를 모르더라도 모수 λ가 포아송 분포, 즉 특정 분포를 따르는, 다시말하면 어떤 특정 모수 집합을 따른다는 사실을 자연스럽게 알 수 있다. 그러므로 이는 비모수적 문제에 해당하지 않고, 비모수적 방법을 적용하지 않고 모수를 추정하거나 검정을 할 수 있다.

▷ **(예제 2)** 어떤 관측값들이 정규분포를 따르는 확률표본을 생성한다고 가정하면, 정규분포의 모수는 평균과 분산으로 2개를 갖는다. 그런데 만약 이러한 모수들이 알려져 있지 않다고 가정하자. 그렇다면 이는 비모수적 문제에 해당하는가?. 결과적으로, 이는 비모수적 문제가 아니다. 모수를 모르더라도 모수가 특정 분포인 정규분포를 따르는 일종의 확률변수임을 알고 있기 때문에, 평균 μ와 분산 σ^2는 정규분포의 성질을 갖는 어떤 특정 모수집합 Ω에 속할 것이다. 그러므로 이는 비모수적 방법이 아니다. 앞선 여러 챕터에서 우리가 배운 여러 추정 및 검정의 내용도 모수적 방법이었다는 것을 상기하자.

§. 범주형 데이터 (Categorical Data)

Definition χ^2 **검정 (Pearson's chi-square test)**

k개의 서로 다른 유형의 개체들로 구성된 큰 모집단을 가정하고, p_i를 유형 $i=1,...,k$에 대해 무작위로 선택되는 개체의 확률이라 하자. 물론, p_i는 확률이므로 확률의 공리 $p_i \geq 0$, $\Sigma_{i=1}^{k} p_i = 1$를 만족한다. p_1^0, ..., p_k^0 를 $p_i^0 > 0$을 만족하는 특정한 수라 가정하고, $\Sigma_{i=1}^{k} p_i^0 = 1$을 만족하며 다음 가설을 검정하려한다고 가정하자.

$$H_0 : p_i = p_i^0 \text{ for } i = 1, ..., k,$$
$$H_1 : p_i \neq p_i^0 \text{ for at least one value of } i.$$

모집단에서 취해지는 크기가 n인 확률표본을 가정하자, 즉 n개의 독립적인 관측값들이 취해지고, 각 관측값이 $i=1,...,k$인 유형이 될 확률 p_i이 존재한다고 하자. 이러한 n개의 관측값을 기저로 하여, 위 가설은 검정할 수 있다.

$i=1,...,k$에 대하여, N_i 를 유형 i에 대한 확률표본의 관측값의 수라고 하자. 그러면 N_1, ..., N_k는 음수가 아닌 정수로써 $\Sigma_{i=1}^{k} N_i = n$을 만족한다(사실, $(N_1,...,N_n)$는 모수가 n과 $\vec{p} = (p_1, ..., p_k)$인 다항분포를 갖는다.).

귀무가설 H_0이 참이면, 유형 i의 관측값들의 예측 수는 np_i^0 , $i=1,...,k$ 이다. 관측값 N_i의 실체 수와 예측 수 np_i^0 사이의 차이는 H_0이 거짓일 때보다 H_0이 참일 때 작아지는 경향이 있다. 따라서 가설검정은 결국 $N_i - np_i^0$ for $i=1,...,k$를 가설 검정으로 하는 것과 같은 문제가 된다. 그리고 $N_i - np_i^0$이 상대적으로 커질 때 H_0을 기각하게 된다.

아래 정리에 제시된 χ^2 검정의 검정통계량은 Karl Pearson 에 의해 1990년도에 증명되었다.

Theorem 10.1.1 χ^2 **검정의 검정통계량**

다음 통계량 Q를 χ^2 검정의 검정통계량으로 택한다.

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} = \sum \frac{(O-E)^2}{E}, \quad O : \text{Observed}, \quad E : \text{Expected}.$$

이는 H_0이 참이고 표본 크기가 $n \to \infty$ 일 때 자유도가 $k-1$인 χ^2 분포로 수렴하는 성질을 갖는다.

▷ **[Definition]** 위 검정통계량을 이용하여 가설검정을 시행하는 것을 **카이제곱 적합도 검정**이라 부른다.

▷ **예제 10.1.3 : Blood Types.**

우리는 아래에 제시된 Table 10.1, 10.2과 같은 범주형 데이터(숫자가 아닌 명목형 자료)에 대한 가설검정을 실시하고자 한다. 특히, 이러한 범주형 데이터에서 확률에 대한 추론을 하고 싶기 때문에, 그에 적합한 방법을 사용해야만 한다. 이러한 범주형 데이터에 대한 가설검정은 χ^2 검정이 그 예로 적합하다. χ^2 검정은 모수 또는 비모수적 문제 둘 다 적용할 수 있는 방법이다. 주어진 Table은 다음과 같다.

Table 10.1 Counts of blood types for white Californians

A	B	AB	O
2162	738	228	2876

Table 10.2 Theoretical probabilities of blood types for white Californians

A	B	AB	O
1/3	1/8	1/24	1/2

이 범주형 데이터는 어떤 특정 분포를 따른다거나 모수집합을 특정할 수 있는 문제가 아니므로, 이는 비모수적 문제가 된다. 그러나 Table 10.1을 잘 살펴보면, 이는 각 혈액형 유형(type)의 빈도수이므로, 각 유형은 전체 빈도수 6004에 대한 수학적 확률, 즉 어떤 비율로 나타낼 수 있다. 실제로, 예를 들어 B형의 상대도수를 컴퓨터를 이용하여 구하면 $738/6004 = 0.1229181$인데, 제시된 Table 10.2를 참고하면 $1/8 = 0.125$로 거의 근사한 값을 가진다. 문제에서는 수학적 확률(Theoredical probabilities)로 Table 10.2를 제시했으므로, 자료의 출처에 관한 별 다른 의심 없이 이를 이용하기로 하자.

이제 우리는 혈액형 타입에 대한 데이터의 각 범주의 수학적 확률을 알고 있다. 그렇다면 이를 가지고 어떻게 가설을 세우고 검정할 것인가?, 이는 앞서 언급한 χ^2 검정이 적합하다. χ^2 검정 방법을 살펴보면, 비율에 대한 가설을 세우고, 이를 검정하는 것임을 쉽게 알 수 있다. 즉, 어떤 모집단의 한 범주가 어떤 비율을 따를 때, 그 비율을 추정하는 것과 같다고 생각할 수 있다. 그러면 우리는 가설을 세울 때 우리가 제시한 특정 비율이 모비율과 같음을 귀무가설로 채택하는 것은 자연스럽다. 그러면 이제 문제는 비율에 대한 가설검정의 문제로 넘어가게 되고, 우리는 이제 χ^2 검정 절차를 잘 따라서 결과를 도출해내면 될 것이다.

이제 가설을 세우자. 우선 범주의 개수는 4이므로, $i = 1, 2, 3, 4$, $k = 4$, Table 10.2에 제시된 확률은 모두 양수이고 $1/3 + 1/8 + 1/24 + 1/2 = 1$로 확률의 공리를 만족한다. 이들 확률을 각각 $p_1^0, p_2^0, p_3^0, p_4^0$로 두고, 표본 크기는 $n = 6004$, 각 혈액형에 대한 빈도수를 N_1, N_2, N_3, N_4로 두면, $\Sigma N_i = n$임은 쉽게 확인할 수 있다. 검정 절차에 따르면, 가설은 다음과 같이 세울 수 있다.

$$H_0 : p_i = p_i^0 \text{ for } i = 1, 2, 3, 4 \text{ vs } H_1 : p_i \neq p_i^0 \text{ for at least one value of } i.$$

가설을 세웠으니, 이제 본격적인 검정을 시행하기 위해 검정통계량을 구해보자. 검정통계량은 Pearson의 검정 통계량을 이용하면 된다.

$$Q=\sum_{i=1}^{k}\frac{(N_i-np_i^0)^2}{np_i^0}=\sum\frac{(O-E)^2}{E},\ O:\text{Observed},\ E:\text{Expected}.$$

우리는 검정통계량 Q의 각 성분들을 모두 구하였으므로, 이를 계산하면 아래와 같이 얻을 수 있다.

$$Q=\sum_{i=1}^{4}\frac{(N_i-np_i^0)^2}{np_i^0}=\left(\frac{(2162-6004\times1/3)^2}{6004\times1/3}\right)+\cdots\ \cdots+\left(\frac{(2876-6004\times1/2)^2}{6004\times1/2}\right)=20.37$$

유의수준 α_0에 대하여 H_0를 검정하기 위해, 자유도가 3인 χ^2 분포의 $1-\alpha_0$ 분계선과 Q를 비교해야 한다. 먼저, p-값을 계산해보자. 이는 H_0을 기각할 수 있는 가장 작은 유의수준을 찾으면 된다. 카이제곱 적합도 검정의 경우, p-값은 $1-X_{k-1}^2(Q)$로 구할 수 있다. 이 때, X : $k-1$ 자유도를 갖는 카이제곱분포의 누적분포함수(c.d.f.)이다. 여기서 $k=4$이므로, $p-value=1-X_3^2(Q)=1.42\times10^{-4}$ 이다.

만약 유의수준 $\alpha_0=0.05$로 두면, 꼬리영역이 되는 α_0보다 p-값이 훨씬 작으므로($1.42\times10^{-4}=0.000142$) $\alpha_0=0.05$에서 귀무가설 H_0을 기각할 수 있다. (α_0는 0.01인 경우도 p-값이 훨씬 작음을 알 수 있다.)

(** $Q\geq(X_{k-1}^2)^{-1}(1-\alpha_0)=c\ \Leftrightarrow\ X_{k-1}^2(Q)\geq1-\alpha_0\ \Leftrightarrow\ \alpha_0\geq1-X_{k-1}^2(Q)=p-value$)

▲

R Code in example 10.1.3

Results

```
> df_blood_freq
     A   B  AB    O
1 2162 738 228 2876

> chisq.test(df_blood_freq,p=df_blood_prob)

        Chi-squared test for given probabilities

data:  df_blood_freq
X-squared = 20.359, df = 3, p-value = 0.000143
```

R Code

```
A=2162
B=738
AB=228
O=2876

df_blood_freq=data.frame(A,B,AB,O)
df_blood_freq
df_blood_prob=c(1/3,1/8,1/24,1/2)

chisq.test(df_blood_freq,p=df_blood_prob)
```

§. 연속분포에 대한 가설 검정 (Testing Hypotheses about a Continuous Distribution)

연속분포에 대한 가설 검정 (Testing Hypotheses about a Continuous Distribution)

특정 분포(in continuous distribution)로부터 취해지는 관측값들에 대한 확률 표본에 대하여 검정하기 위해, 다음의 과정을 적용한다.

1. 실직선(real line) 전체를 분할하거나, k개의 서로소인 부분구간으로 이루어지며 구간 전체의 확률은 1을 만족하는 임의의 특정 구간을 설정한다. 일반적으로, 귀무가설 H_0이 참이면, k는 선택되며 각 부분구간의 관측값의 예상 개수는 적어도 5 이상이다.

2. 특정한 가정된 분포로써 i번째 부분구간에 배당되는 확률 p_i^0 를 결정한다, 그리고 각 $i=1,...,k$에 대하여 i번째 부분구간의 관측값들의 예상 갯수 np_i^0를 계산한다.

3. i번째 부분구간에 해당하는 표본의 관측값들의 개수 N_i를 센다.

4. 검정통계량 Q를 계산한다. 만약 가정된 분포가 정확하면, Q는 근사적으로 $k-1$의 자유도를 갖는 χ^2 분포를 따른다.

▷ **예제 10.1.6 : Failure Times of Ball Bearings.**
이전 예제 10.1.1로 돌아가자. 수명에 대한 로그값이 $i.i.d.$한 정규분포를 따르는 확률 표본이라 가정하고, 귀무가설에 대한 χ^2검정을 실시하고자 한다. 이 때, 평균 $\mu=\log(50)=3.912$, 분산 $\sigma^2=0.25$라 하자.

각 구간에 대한 기대 개수가 적어도 5개 이상이기 위해서, 우린 많아야 $k=4$개의 구간들을 사용할 수 있다. 또한 우린 이러한 구간들이 귀무가설의 가정 아래에서 각각 0.25의 확률을 가진다고 가정할 것이다. 즉, 우린 총 확률이 1인 구간을 가정한 정규분포에서의 0.25, 0.5, 0.75 분계수를 갖는 구간으로 나눌 것이다. 이러한 분계수는 다음과 같이 정해진다.

$$\begin{aligned} 3.912+0.5\Phi^{-1}(0.25) &= 3.192+0.5\times(-0.674) = 3.575 \\ 3.912+0.5\Phi^{-1}(0.5) &= 3.192+0.5\times 0 \quad\quad\;\; = 3.912 \\ 3.912+0.5\Phi^{-1}(0.75) &= 3.192+0.5\times 0.674 \quad = 4.249 \end{aligned}$$

($\because$ 표준정규분포의 0.25와 0.75 분위수 : ± 0.674)

관측된 로그값들은 다음과 같이 주어진다.

2.88	3.36	3.50	3.73	3.74	3.82	3.88	3.95
3.95	3.99	4.02	4.22	4.23	4.23	4.23	4.43
4.53	4.59	4.66	4.66	4.85	4.85	5.16	

네 개의 구간 각각의 관측값들의 개수를 구하면 3, 4, 8, 8로 구해진다. 우린 이제 검정통계량 Q를 계산할 수 있다. 검정통계량 Q는 다음과 같이 구해진다.

$$Q=\frac{(3-23\times0.25)^2}{23\times0.25}+\frac{(4-23\times0.25)^2}{23\times0.25}+\frac{(8-23\times0.25)^2}{23\times0.25}+\frac{(8-23\times0.25)^2}{23\times0.25}=3.609$$

χ^2분포표에 의하면, $Q=3.609$에 대한 자유도가 3인 χ^2분포의 분계수는 0.6과 0.7 사이에 위치한다(2.946과 3.665 사이). 따라서 우린 유의수준 0.3 이하에서 귀무가설을 기각할 수 없고, 유의수준 0.4 이상부터는 귀무가설을 기각할 수 있다. (실제로, p-값은 0.307이다.)

$(\because\ \alpha_0 \geq 1-X^2_{k-1}(Q)=p-value \Rightarrow \alpha_0 \geq p-value=\begin{cases}1-0.6 & Let\ Q=2.946,\ Q\in(2.946,3.665)\\ 1-0.7 & Let\ Q=3.665,\ Q\in(2.946,3.665)\end{cases}$

$\Rightarrow \alpha_0 \geq 0.4 \text{ or } 0.3 \Rightarrow \alpha_0 \in (0.3,\infty))$

▲

R Code in example 10.1.6

Results

```
> df_ball_freq
  log1 log2 log3 log4
     3    4    8    8

> chisq.test(df_ball_freq,p=df_ball_prob)

        Chi-squared test for given probabilities

data:  df_ball_freq
X-squared = 3.6087, df = 3, p-value = 0.3069
```

R Code

```
log1=3
log2=4
log3=8
log4=8

df_ball_freq=data.frame(log1,log2,log3,log4)
df_ball_freq
df_ball_prob=c(0.25,0.25,0.25,0.25)

chisq.test(df_ball_freq,p=df_ball_prob)
```

▷ **Additional Problem of χ^2 test**

The data on the number of arrivals of cars at an intersection in 360 10s intervals are as shown in Table. (Mean = 1.013, Variance = 1.1314)

Cars per interval	Number of observations
0	139
1	128
2	55
3	25
4	13

Three models are proposed :

model 1 : $p_X(x) = \frac{e^{-1}}{x!}, x = 0, 1, \cdots$

model 2 : $p_X(x) = \frac{e^{-1}\lambda^x}{x!}, x = 0, 1, \cdots$

model 3 : $p_X(x) = \binom{x+k-1}{k-1} p^k (1-p)^k, x = 0, 1, \cdots$

(a) Use the χ^2 test; are these models acceptable at the 5% significance level?

(b) In your opinion, which is a better model? Explain your answer.

(Solution)

(a) 우선, 주어진 Table의 관측값들의 총합은 $n=360$으로 구해진다. 각 범주별 확률을 구하자.

① model 1 : $p_X(x)=\frac{e^{-1}}{x!}, x=0,1,\cdots$

Cars per interval	Number of observations	Probability of occurrence	Expected frequency
0	139	0.3679	132.444
1	128	0.3679	132.444
2	55	0.184	66.24
3	25	0.0613	22.068
4	13	0.01533	5.5188

χ^2 검정의 검정통계량 Q는 다음과 같이 구해진다.

$$Q=\frac{(132.444-139)^2}{132.444}+\frac{(128-132.444)^2}{132.444}+\frac{(55-66.24)^2}{66.24}+\frac{(25-22.068)^2}{22.068}+\frac{(13-5.5188)^2}{5.5188}=13.088$$

$k=5$이므로, 0.05 유의수준에서 기각값을 구하면 $\chi^2_{0.05}(5-1)=9.49$이다.

∴ 통계량 13.088이 0.05 유의수준에서 기각값 9.49보다 크기 때문에 표본의 분포와 model 1의 분포간에 차이가 없다는 귀무가설을 기각한다. 따라서 model 1의 분포는 표본분포를 반영하지 못한다.

② model 2 : $p_X(x)=\frac{e^{-1}\lambda^x}{x!}, x=0,1,\cdots$ (10초에 평균적으로 1.013대 관측된다)

Cars per interval	Number of observations	Probability of occurrence	Expected frequency
0	139	$\frac{e^{-1.013}1.013^0}{0!}=0.3631$	130.716
1	128	$\frac{e^{-1.013}1.013^1}{1!}=0.3678$	132.408
2	55	$\frac{e^{-1.013}1.013^2}{2!}=0.1863$	67.088
3	25	$\frac{e^{-1.013}1.013^3}{3!}=0.0629$	22.644
4	13	$\frac{e^{-1.013}1.013^4}{4!}=0.0159$	5.724

χ^2 검정의 검정통계량 Q는 다음과 같이 구해진다.

$$Q=\frac{(130.716-139)^2}{130.716}+\frac{(128-132.408)^2}{132.404}+\frac{(55-67.068)^2}{67.068}+\frac{(25-22.644)^2}{22.644}+\frac{(13-5.724)^2}{5.724}=12.333$$

$k=5$이므로, 0.05 유의수준에서 기각값을 구하면 $\chi^2_{0.05}(5-1)=9.49$이다.

∴ 통계량 12.333이 0.05 유의수준에서 기각값 9.49보다 크기 때문에 표본의 분포와 model 2의 분포간에 차이가 없다는 귀무가설을 기각한다. 따라서 model 2의 분포는 표본분포를 반영하지 못한다.

③ model 3 : $p_X(x)=\binom{x+k-1}{k-1}p^k(1-p)^k, x=0,1,\cdots$

주어진 model 3을 구체적으로 정리해보자. 우선, k, p는 주어진 데이터로부터 구할 수 있다. model 3의 기댓값(평균) 과 분산의 값을 알고 있으므로, 즉 $E(X)=\frac{k(1-p)}{p}=1.013$, $Var(X)=\frac{k(1-p)}{p^2}=1.1314$ 이므로 방정식을 풀면 $k=8.635, p=0.895$로 구해진다.
따라서 발생확률은 $p_X(x)=\binom{x+9-1}{9-1}(0.895)^9(1-0.895)^x$, $x=0,1,\cdots$로 구해진다.

Cars per interval	Number of observations	Probability of occurrence	Expected frequency
0	139	0.3685	132.66
1	128	0.3482	125.352
2	55	0.183	65.88
3	25	0.0704	25.344
4	13	0.0222	7.92

χ^2 검정의 검정통계량 Q는 다음과 같이 구해진다.

$$Q=\frac{(132.66-139)^2}{132.66}+\frac{(128-125.352)^2}{125.352}+\frac{(55-65.88)^2}{65.88}+\frac{(25-25.344)^2}{25.344}+\frac{(13-7.92)^2}{7.92}=5.423$$

$k=5$이므로, 0.05 유의수준에서 기각값을 구하면 $\chi^2_{0.05}(5-1)=9.49$이다.

∴ 통계량 5.423이 0.05 유의수준에서 기각값 9.49보다 작기 때문에 표본의 분포와 model 3의 분포간에 차이가 없다는 귀무가설을 기각하지 못한다. 따라서 model 3의 분포는 표본분포를 반영한다.

(b) (a)의 논의에 의해, model 3이 5% 유의수준에서 표본분포를 반영하므로, model 3이 적합하다.

▲

R Code in Additional Problem of χ^2 test.

Results

```
> df_model
    car_0 car_1 car_2 car_3 car_4
1   139   128   55    25    13

> chisq.test(df_model,p=df_model_1_prob)

        Chi-squared test for given probabilities

data:  df_model
X-squared = 8.4128, df = 4, p-value = 0.07758

> chisq.test(df_model,p=df_model_2_prob)

        Chi-squared test for given probabilities

data:  df_model
X-squared = 7.8425, df = 4, p-value = 0.09752

> chisq.test(df_model,p=df_model_3_prob)

        Chi-squared test for given probabilities

data:  df_model
X-squared = 2.6249, df = 4, p-value = 0.6224
```

R Code

```
#Variables of Table
car_0=139
car_1=128
car_2=55
car_3=25
car_4=13

#Table
df_model=data.frame(car_0,car_1,car_2,car_3,car_4)
df_model

#Probabilites of model 1,2,3.
df_model_1_prob=c(0.3679,0.3679,0.184,0.0613,0.0189)
df_model_2_prob=c(0.3631,0.3678,0.1863,0.0629,0.0199)
df_model_3_prob=c(0.3685,0.3482,0.183,0.0704,0.0299)

#Confirm the sum of the probability
sum(df_model_1_prob)
sum(df_model_2_prob)
sum(df_model_3_prob)

#Chi^2 - test of three models
chisq.test(df_model,p=df_model_1_prob)
chisq.test(df_model,p=df_model_2_prob)
chisq.test(df_model,p=df_model_3_prob)
```

※ 문제에서 직접 구했던 발생확률의 총합이 1이 되지 않게 구하였으므로(반올림오차, 절단오차 등의 이유로 인한)이를 맞추기 위해 범주 4(car per interval)에 부족한 확률을 더하여 구하였다. 이는 검정통계량 값이 달라질 수 있으나 결과적으로 검정 결과는 본문과 같다. 되도록 발생확률을 정확히 구하는 것이 중요하다. (발생확률은 직접 소프트웨어를 이용하여 구해보자, 모두 특정 분포의 pdf가 되므로 이를 유의하여 함수를 사용하거나 해석적으로 풀이 후 값만 계산하는 방법 모두 활용하여 정확히 구하도록 하자.)

10.2 복합 가설에 대한 적합도 검정 (Goodness-of-Fit for Composite Hypotheses)

§. 복합 귀무가설 (Composite Null Hypotheses)

▷ **예제 10.2.1 : Failure Times of Ball Bearings.**

예제 10.1.6에서, 평균이 3.912, 분산이 0.25인 정규분포를 갖는 볼 베어링 수명의 로그 값에 대한 귀무가설을 검정하였다. 이제 우린 이러한 정규분포가 수명의 로그값에 대해 좋은 모델임을 보장할 수 없다고 가정하자. 혼합 귀무가설 : 로그-수명 값에 대한 분포가 정규-집합족의 원소를 검정할 방법이 있는가?

▲

▷ **예제 10.2.2 : Genetics.**

이전 예제 1.6.4에서, 유전자가 두 개의 다른 대립유전자를 가진다는 것을 알아두자. 주어진 모집단에서 각 개체들은 반드시 세 개 중 하나의 유전형질을 갖는다. 만약 대립유전자가 두 부모들로부터 독립적으로 받아지고, 모든 부모들이 동일한 첫 번째 대립유전자를 통과할 확률 θ를 갖는다면, 세 개의 서로 다른 유전형질의 확률 p_1, p_2, p_3은 다음과 같은 형태로 표현될 수 있다.

$$p_1 = \theta^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = (1-\theta)^2$$

여기서, 모수 θ의 값은 알려지지 않고 구간 $0 < \theta < 1$ 내의 어디든 존재할 수 있다. 이 구간에서 각 θ의 값에 대하여, 이는 $p_i > 0$ for $i = 1, 2, 3$이고, $p_1 + p_2 + p_3 = 1$을 만족한다. 이 문제에서, 확률표본은 모집단으로부터 취해지고, 통계학자들은 반드시 p_1, p_2, p_3이 위와 같이 가정된 형태로 표현될 수 있는 구간 $0 < \theta < 1$에서 어떤 θ값이 존재한다는 믿음이 합리적인지 결정하고자 각 세 개의 유전형질에 대한 개개인의 관측된 수를 사용해야한다. 만약 유전자가 세 개의 다른 대립유전자를 갖는다면, 모집단의 개개인들은 반드시 6개의 가능한 유전형질 중 하나를 가져야한다. 다시 한 번, 만약 대립유전자가 부모로부터 독립적으로 통과한다면, 그리고 각 부모는 첫 번째와 두 번째 대립유전자가 자손을 통과할 확률 θ_1, θ_2을 갖는다고 하면, 확률 서로 다른 유전형질에 대한 확률 $p_1, \ldots, p_6$는 $\theta_1 > 0$, $\theta_1 + \theta_2 < 1$을 만족하는 θ_1, θ_2 로써 다음과 같이 표현할 수 있다.

$$p_1 = \theta_1^2,\ p_2 = \theta_2^2,\ p_3 = (1-\theta_1-\theta_2)^2,\ p_4 = 2\theta_1\theta_2,\ p_5 = 2\theta_1(1-\theta_1-\theta_2),\ p_6 = 2\theta_2(1-\theta_1-\theta_2)$$

다시, 앞서 서술된 조건을 만족하는 모든 θ_1과 θ_2에 대하여, $p_i > 0$ $i = 1, \ldots, 6$과 $\Sigma_{i=1}^{6} p_i = 1$임은 증명될 수 있다. 확률표본에서 각 유전형질을 갖는 개개인의 기본적인 관측 수 $N_1, \ldots, N_6$에서, 통계학자들은 어떤 θ_1, θ_2에 대하여 확률 $p_1, \ldots, p_6$이 $p_1 = \theta_1^2$, $p_2 = \theta_2^2, \ldots$ 와 같이 표현될 수 있다는 귀무가설을 기각할지 기각할 수 없는지 반드시 결정해야만 한다.

▲

Definition 복합가설 (Composite Hypotheses)
만일 가설이 완전히 확률밀도함수 $f(x|\theta)$를 명확히 결정하면 그 가설을 **단순가설(simple hypothesis)**이라 하고, 그렇지 못할 경우 **복합가설(composite hypothesis)** 이라 한다.

일반적으로, Ω가 모수공간이고 $\Omega_0 \subset \Omega$일 때 귀무가설을 $H_0 : \theta \in \Omega_0$ 으로 표시할 수 있으며 대립가설은 $H_1 : \theta \in \Omega_1 = \Omega - \Omega_0$으로 표시할 수 있다. 이 경우, θ가 두 개 이상의 값이거나 특정할 수 없고, $\theta \in \Omega_0$이면 H_0은 **복합 귀무가설(Composite Null Hypotheses)**이다. 반면에 $\theta = \theta_0$인 하나의 값으로 특정될 수 있다면 이는 단순 귀무가설이 된다.

▷ **(예제1)** $X \sim N(\mu, \sigma_0^2)$이라 하자. σ_0^2를 알고 있다면 모수공간은 $\Omega = (-\infty, \infty)$이다. 이 때 귀무가설은 $H_0 : \mu \leq \mu_0$이고 대립가설은 $H_1 : \mu > \mu_0$이라 하자. $\Omega_0 = (-\infty, \mu_0]$, $\Omega_1 = (\mu_0, \infty)$이므로 이는 복합가설이다.

▷ **(예제2)** 예제1에서 귀무가설은 $H_0 : \mu = \mu_0$이고 대립가설은 $H_1 : \mu \neq \mu_0$이라 하자. 그러면 $\Omega_0 = \mu_0$, $\Omega_1 = (-\infty, \mu_0) \cup (\mu_0, \infty)$이므로 귀무가설은 단순가설이고 대립가설은 복합가설이다.

▷ **(예제3)** $X \sim N(\mu, \sigma_0^2)$이라 하자. 이 때 σ^2는 알려지지 않은 모수이다. 가설 $H : \mu = \mu_0$, μ_0은 알고 있는 모수라 하면 이는 복합가설이다.

▷ **(예제4)** $X \sim N(\mu, \sigma_0^2)$이라 하자. 가설 $H : \mu = \mu_0$, $\sigma^2 = \sigma_0^2$, μ_0, σ_0^2은 알고 있는 모수이면 이는 단순가설이 된다.

▷ **(예제5)** 가설 H : X가 모수 $\lambda < 3$인 포아송 확률변수이다. → 복합가설
▷ **(예제6)** 가설 H : X가 모수 $\theta = 5$인 지수 확률변수이다. → 단순가설
▷ **(예제7)** 가설 H : $F_X(x|\theta) = F_X(x|\theta_0)$, θ_0은 알고 있는 모수(상수)이다. → 단순가설

Theorem 10.2.1 복합 귀무가설에서의 χ^2 검정
가설검정에서 다음과 같은 복합가설이 주어진다고 하자, 이 때, $\theta = (\theta_1, \ldots, \theta_s)$, $\pi_i(\theta)$는 확률 p_i의 특정 함수이고, 모수공간의 차원은 $\dim(\Omega) = s$이며 $\Sigma_{i=1}^{k} \pi_i(\theta) = 1$을 만족한다.

$$H_0 : \exists \text{ a value of } \theta \in \Omega \text{ s.t. } p_i = \pi_i(\theta) \text{ for } i = 1, \ldots, k,$$
$$H_1 : \text{The hypothesis } H_0 \text{ is not true.}$$

이 귀무가설이 참이고 특정 정칙 조건이 만족된다고 가정하자. 그러면 표본 크기가 $n \to \infty$일 때, 다음의 검정통계량 Q의 c.d.f.가 자유도가 $k-1-s$인 χ^2-분포의 c.d.f.로 수렴한다.

$$Q = \sum_{i=1}^{k} \frac{\left[N_i - n\pi_i(\hat{\theta})\right]^2}{n\pi_i(\hat{\theta})}, \quad N_i : \text{observed number}, \quad \hat{\theta} : \text{M.L.E. of } \theta.$$

▷ **예제 10.2.3 : Genetics.**
모든 유전자는 두 개의 대립유전자를 갖고, 모집단에서의 각 개체들은 반드시 세 가지 유전형질 중 하나를 가져야 한다고 가정하였다.

이 때, 첫 번째 대립유전자를 가진다고 가정했을 때 나타날 유전형질의 확률 세 가지를 제시하고, 두 번째 대립유전자에 대해서도 마찬가지로 정의하였다. 따라서 첫 번째 대립유전자에 대해서는 $k=3$이고, 귀무가설은 $H_0 : p_1=\theta^2,\ \ p_2=2\theta(1-\theta),\ \ p_3=(1-\theta)^2$이며, 대립가설 H_1은 H_0이 참이 아닌 것으로 정하면 된다. 이 때, $s=1$로 할 수 있고(모수가 θ 하나이므로), 그래서 귀무가설이 참이면 정리 10.2.1의 검정통계량 Q는 근사적으로 자유도가 $k-1-s=3-1-1=1$인 χ^2-분포를 따른다.

두 번째 대립유전자를 가진다고 가정했을 때 나타날 유전형질의 확률은 6개로 정의했으므로, $k=6$이 된다. 귀무가설과 대립가설은 이전과 마찬가지로,

$$H_0 : p_1=\theta_1^2,\ p_2=\theta_2^2,\ p_3=(1-\theta_1-\theta_2)^2,\ p_4=2\theta_1\theta_2,\ p_5=2\theta_1(1-\theta_1-\theta_2),\ p_6=2\theta_2(1-\theta_1-\theta_2)$$
$$H_1 : \text{귀무가설 } H_0\text{이 참이 아니다.}$$

따라서 이 경우에서 $s=2$가 된다(모수가 θ_1,θ_2로 2개). 그러므로 이 또한 마찬가지로 귀무가설이 참일 때 정리 10.2.1의 검정통계량 Q는 근사적으로 자유도가 $k-1-s=6-1-2=3$인 χ^2-분포를 따른다.

▲

§. 최대우도추정량의 결정 (Determining the Maximum Likelihood Estimates)

최대우도추정량의 결정 (Determining the Maximum Likelihood Estimates)
귀무가설 H_0이 참일 때, 관측값의 개수 $N_1, \ldots, N_k$ 에 대한 우도함수 $L(\theta)$는 다음과 같다.

$$L(\theta) = \binom{n}{N_1, \ldots, N_k} [\pi_i(\theta)]^{N_1} \cdots [\pi_k(\theta)]^{N_k}$$

이는 곧, 다음과 같이 덧셈에 대한 식으로 분리할 수 있다.

$$\log L(\theta) = \log \binom{n}{N_1, \ldots, N_k} + \sum_{i=1}^{k} N_i \log \pi_i(\theta)$$

따라서, 최대우도추정량 $\hat{\theta}$은 $\log L(\theta)$가 최대가 되는 θ의 값을 찾으면 된다. 또한, 여기서 $\binom{n}{N_1, \ldots, N_k}$는 최대값을 찾을 때 어떤 영향도 끼치지 못하므로, 계산 시 이를 무시하여도 좋다.

▷ **예제 10.2.4 : Genetics.**
이전 예제에 이어서 실질적인 검정통계량을 구해보자. 우선 $k=3$이고 각각의 확률 p_1, p_2, p_3은 우도함수를 이용하여 다음과 같이 구할 수 있다.

$$\begin{aligned} \log L(\theta) &= N_1 \log(\theta^2) + N_2 \log[2\theta(1-\theta)] + N_3 \log[(1-\theta)^2] \\ &= (2N_1 + N_2)\log\theta + (2N_3 + N_2)\log(1-\theta) + N_2 \log 2 \end{aligned}$$

이제 이를 미분하여 $L(\theta)$가 최대가 되는 θ값을 찾으면, 최대우도추정량 $\hat{\theta}$는 아래와 같이 구해질 수 있다.

$$\hat{\theta} = \frac{2N_1 + N_2}{2(N_1 + N_2 + N_3)} = \frac{2N_1 + N_2}{2n}$$

이를 이용하여 정리 10.2.1.의 검정통계량 Q를 계산할 수 있다. 여기서 Q는 위의 M.L.E.에 의하여 오직 관측값의 개수 N_1, N_2, N_3로 구해질 수 있음에 주목하자. 앞서 언급하였듯이, 귀무가설이 참이고 표본크기 n이 크다면 Q의 분포는 근사적으로 자유도가 1인 χ^2-분포를 따른다. 이런 이유로, 검정통계량 Q값에 대응하는 꼬리 쪽 영역은 χ^2-분포에서 찾아질 수 있다.

▲

§. 분포가 정규분포인 경우의 가설 검정 (Testing Whether a Distribution Is Normal)

Theorem 10.2.2 복합 귀무가설, 분포가 정규분포인 경우의 χ^2 검정

$X_1, ..., X_n$이 p차원 모수 θ를 갖는 분포로부터 취해지는 확률 표본이라고 하고, $\hat{\theta}_n$을 최대우도추정량이라 하자. 모수공간의 실직선을 $k > p+1$인 k개의 서로소인 구간들 $I_1, ..., I_k$로 분할하자. N_i을 I_i $i = 1,...,k$ 각각의 관측값들의 개수라고 하고, $\pi_i(\theta) = \Pr(X_i \in I_i|\theta)$로 정의하자. 검정통계량 Q'는 다음과 같이 정의하자.

$$Q' = \sum_{i=1}^{k} \frac{\left[N_i - n\pi_i(\hat{\theta}_n)\right]^2}{n\pi_i(\hat{\theta}_n)}, \quad N_i : \text{observed number}, \quad \hat{\theta}_n : \text{M.L.E. of } \theta.$$

최대우도추정량의 점근적 정규성에 대하여 정칙 조건을 만족한다고 가정하자. 그러면 $n \to \infty$일 때 Q'의 c.d.f.는 자유도가 $k-p-1$인 χ^2-분포의 c.d.f.와 자유도 $k-1$인 χ^2-분포의 c.d.f. 사이에서 수렴한다.

▷ **예제 10.2.6 : Prussian Army Deaths.**

이전 예제 7.3.14를 참조하자. 우린 말 발길질에 의해 사망한 프로이센 군인들의 수가 포아송 확률변수로써 모형화 되었다고 가정하였다. 여기서 우린 이러한 사망 수가 포아송 분포를 따른다는 귀무가설과 그를 부정하는 대립가설을 세워서 검정하고 싶다. 그러한 사망 수는 다음과 같이 주어진다.

Count	0	1	2	3	≥ 4
Number of Observations	144	91	32	11	2

주어진 데이터가 포아송 분포로부터 확률표본을 생성한다고 가정하였을 때의 우도함수는 θ의 함수로써 $\exp(-280\theta)\theta^{196}$ 에 비례한다. 최대우도추정량은 $\hat{\theta}_n = 196/280 = 0.7$로 구해진다. 우린 검정통계량 Q'를 계산하기 위해 $k = 5$로 분류할 것이다. 이러한 다섯 개의 확률은 다음과 같이 결정된다.

Count	0	1	2	3	≥ 4
$\pi_i(\hat{\theta}_n)$	0.4966	0.3476	0.1217	0.0283	0.0058

따라서 검정통계량 Q'는 다음과 같이 구할 수 있다.

$$Q' = \frac{(144 - 280 \times 0.4966)^2}{280 \times 0.4966} + \frac{(91 - 280 \times 0.3476)^2}{280 \times 0.3476} + \frac{(32 - 280 \times 0.1217)^2}{280 \times 0.1217}$$

$$+ \frac{(11 - 280 \times 0.0283)^2}{280 \times 0.0283} + \frac{(2 - 208 \times 0.0058)^2}{208 \times 0.0005} = 1.979$$

구한 Q'값과 대응하는 꼬리쪽 영역과, 자유도가 각각 4와 3일 때 χ^2분포의 누적분포함수는 0.7396, 0.5768로 구해진다. 그러므로 우린 유의수준 $\alpha_0 < 0.5768$ 일 때 H_0을 기각할 수 없다.

▲

10.3 분할표 (Contingency Tables)

§. 분할표에서의 독립성 (Independence in Contingency Tables)

Definition 10.3.1 분할표
각 관측값들이 두 가지 또는 그 이상으로 분류되는 표를 **분할표(Contingency Tables)**라고 부른다.

▷ **예제 10.3.1 : College Survey.**
어떤 대학의 전체 재학 중인 학생들 중 200명의 학생들이 확률표본으로 선택되었다고 가정하자, 그리고 표본의 각 학생들은 각 커리큘럼에 따라 분류되고, 다가올 선거에서의 두 정당 A, B의 후보자들 중 누구를 선호하는지에 따라 분류하였다.

Table 10.12 Classification of students by curriculum and candidate preference

	Candidate preferred			
Curriculum	A	B	Undecided	Totals
Engineering and science	24	23	12	59
Humanities and social sciences	24	14	10	48
Fine arts	17	8	13	38
Industrial and public administration	27	19	9	55
Totals	92	64	44	200

우리는 이러한 Table에서, 커리큘럼과 후보자의 선택이 서로 독립적인지 알고 싶어 한다. 좀 더 자세하게 말하면, 대학의 전체 재적자 수로부터 무작위적으로 학생들을 뽑는다고 가정하자. 이 때, 독립성은 각 i, j에 대하여, $\Pr(candidate\ j \cap curriculum\ i) = \Pr(candidata\ j)\Pr(curriculum\ i)$ 를 만족하는지를 확인하고 싶은 것이다.

▲

χ^2 독립성 검정 (The χ^2 Test of Independence)

분할표로 표현된 데이터 셋에서 for $i=1,...,R$ (rows) and $j=1,...,C$ (columns), p_{i+}, p_{+j} : 주변확률(marginal probabilities) 라 할 때, 다음과 같은 가설이 주어진다고 가정하자.

$$H_0 : p_{ij} = p_{i+}p_{+j}$$
$$H_1 : H_0 \text{ is not true.}$$

이러한 귀무가설에서, 각 셀(cell)의 알려지지 않은 확률 p_{ij}은 함수로써 표현되는 알려지지 않은 모수 p_{i+}, p_{+j}로 표현된다. 각각의 확률은 $\sum_{i=1}^{R} p_{i+} = 1$, $\sum_{j=1}^{C} p_{+j} = 1$이므로, 귀무가설이 참일 때 추정될 수 있는 알려지지 않은 모수의 실질적인 개수는 $s=(R-1)+(C-1)$ 또는 $R+C-2$로 구해진다.

$\hat{E}_{ij}$를 귀무가설이 참인 경우의 최대우도추정량(M.L.E.)라고 하자. 이 때, 검정통계량 Q는 다음과 같다.

$$Q = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

여기서 $n \to \infty$일 때 Q의 c.d.f.는 자유도가 $RC-1-s=(R-1)(C-1)$인 χ^2-분포의 c.d.f.로 수렴한다.

이제 최대우도추정량의 식을 고려하자. 각각의 i행과 j행의 기댓값은 np_{ij}로 구해진다. 귀무가설이 참일 때, $p_{ij} = p_{i+}p_{+j}$를 만족하므로 $\hat{p}_{i+}$와 $\hat{p}_{+j}$를 p_{i+}, p_{+j}의 최대우도추정량이라 하면 $\hat{E}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$ 로 구할 수 있다. 각각의 확률의 최대우도 추정량은 N_{i+}/n, N_{+j}/n으로 구해지므로, 최종적으로 E의 최대우도추정량은 아래와 같이 구해진다.

$$\hat{E}_{ij} = n\left(\frac{N_{i+}}{n}\right)\left(\frac{N_{+j}}{n}\right) = \frac{N_{i+}N_{+j}}{n}$$

결과적으로, 위의 결과들을 종합하여 우리는 검정통계량 Q를 구할 수 있고, 이는 표본 n이 커질 때 근사적으로 $(R-1)(C-1)$의 자유도를 갖는 χ^2-분포에 근사한다.

▷ **예제 10.3.2 : College Survey.**

이제 앞선 가정들을 갖고 이 분할표에 대한 독립성 검정을 실시해보자. 우선 분할표를 관찰하면, $N_{1+}=59$, $N_{2+}=48$, $N_{3+}=38$, $N_{4+}=55$ 로 구해지고, $N_{+1}=92$, $N_{+2}=64$, $N_{+3}=44$로 구해진다. 전체 모집단의 표본 개수는 $n=200$이므로, $\hat{E}_{ij}$는 아래의 Table 10.13과 같이 구할 수 있다.

Table 10.13 Expected cell counts for Example 10.3.2

	Candidate preferred			
Curriculum	A	B	Undecided	Totals
Engineering and science	27.14	18.88	12.98	59
Humanities and social sciences	22.08	15.36	10.56	48
Fine arts	17.48	12.16	8.36	38
Industrial and public administration	25.30	17.60	12.10	55
Totals	92	64	44	200

이제 앞선 논의에 의해 제시된 검정통계량의 식을 이용하면, $Q=6.68$이다. 이 때, 행 $R=4$, 열 $C=3$이므로, 검정통계량에 대응하는 꼬리영역은 자유도 $(R-1)(C-1)=6$인 χ^2-분포표로부터 찾을 수 있고, 그러한 값은 0.3보다 크다. 그러므로, 우린 귀무가설 H_0을 유의수준 $\alpha_0 \geq 0.3$인 경우에 기각할 수 있다.

▲

R Code in Example 10.3.2 of χ^2 test of independence.

Results

```
> df_candi_prefer
   candidate_A candidate_B Undecided
1           24          23        12
2           24          14        10
3           17           8        13
4           27          19         9

> chisq.test(df_candi_prefer,p=df_candi_pro)

        Pearson's Chi-squared test

data:  df_candi_prefer
X-squared = 6.6849, df = 6, p-value = 0.351
```

R Code

```
# variables
candidate_A<-c(24,24,17,27)
candidate_B<-c(23,14,8,19)
Undecided<-c(12,10,13,9)

# probabilities
Pr_A<-c(27.14,22.08,17.48,25.30)
Pr_B<-c(18.88,15.36,12.16,17.60)
Pr_Undecided<-c(12.98,10.56,8.36,12.10)

# contingency table
df_candi_prefer=data.frame(candidate_A,candidate_B,Und
ecided)
df_candi_prefer
df_candi_pro=c(Pr_A,Pr_B,Pr_Undecided)

# chi-square test of independence
chisq.test(df_candi_prefer,p=df_candi_pro)
```

▷ **예제 10.3.3 : Montana Outlook Poll.**

몬타나에 거주하는 거주민들의 자신들의 개인적인 재정 상태에 대한 의견을 조사하였다. 그 결과는 다음 Table 10.14 와 10.15로 주어지고, 이를 참고하여 독립성 검정을 실시한다. 이 예제는 R을 이용한다.

Table 10.14 Responses to two questions from Montana Outlook Poll

	Personal finantial status			
Income range	Worse	Same	Better	Total
Under $20,000	20	15	12	47
$20,000~$35,000	24	27	32	83
Over $35,000	14	22	23	59
Total	58	64	67	189

Table 10.15 Expected cell counts for Table 10.14 under the assumption of independence

	Personal finantial status			
Income range	Worse	Same	Better	Total
Under $20,000	14.42	15.92	16.66	47
$20,000~$35,000	25.47	28.11	29.42	83
Over $35,000	18.11	19.98	20.92	59
Total	58	64	67	189

R Code in Example 10.3.3 of χ^2 test of independence.

Results

```
> df_Montana
  Worse Same Better
1    20   15     12
2    24   27     32
3    14   22     23


> chisq.test(df_Montana,p=df_Montana_pro)

        Pearson's Chi-squared test

data:  df_Montana
X-squared = 5.2104, df = 4, p-value = 0.2664
```

R Code

```
# variables
Worse<-c(20,24,14)
Same<-c(15,27,22)
Better<-c(12,32,23)

# probabilities
Pr_Worse<-c(14.42,25.47,18.11)
Pr_Same<-c(15.92,28.11,19.98)
Pr_Better<-c(16.66,29.42,20.92)

# contingency table
df_Montana=data.frame(Worse,Same,Better)
df_Montana
df_Montana_pro=c(Pr_Worse,Pr_Same,Pr_Better)

# chi-square test of independence
chisq.test(df_Montana,p=df_Montana_pro)
```

10.4 동질성 검정 (Tests of Homogeneity)

§. 여러 모집단으로부터의 표본 (Samples from Several Populations)

χ^2 동질성 검정 (The χ^2 Test of Homogeneity)

분할표로 표현된 데이터 셋에서 for $i=1,\dots,R$ (rows) and $j=1,\dots,C$ (columns), p_{i+}, p_{+j} : 주변확률(marginal probabilities) 라 할 때, 다음과 같은 가설이 주어진다고 가정하자.

$$H_0 : p_{1j}=p_{2j}=\cdots=p_{Rj}$$
$$H_1 : H_0 \text{ is not true.}$$

$\hat{E}_{ij}$를 귀무가설이 참인 경우의 최대우도추정량(M.L.E.)라고 하자. 이 때, 검정통계량 Q는 다음과 같다.

$$Q=\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(N_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$

여기서 $n\to\infty$일 때 Q의 c.d.f.는 자유도가 $RC-1-s=(R-1)(C-1)$인 χ^2-분포의 c.d.f.로 수렴한다. E의 최대우도추정량은 아래와 같이 구해진다.

$$\hat{E}_{ij}=n\left(\frac{N_{i+}}{n}\right)\left(\frac{N_{+j}}{n}\right)=\frac{N_{i+}N_{+j}}{n}$$

이 결과들은 사실 독립성 검정과 거의 동일하며, 귀무가설에 대한 설정만 다름에 유의하자.

▷ **예제 10.4.3 : A Clinical Trial.**

다음 Table 2.1을 참고하여 동질성 검정을 실시하자. 우리는 귀무가설에서 No relapse일 확률이 모든 네 개의 Treatment group에서 같다고 하고 싶다. 이러한 분할표의 검정통계량 $Q=10.80$으로 구해진다. 이는 자유도가 3인 χ^2-분포의 0.987 분계수에 해당한다. 즉, p-값이 0.013이고, 귀무가설은 $\alpha_0 \geq 0.013$인 모든 유의수준 α_0에서 기각할 수 있다. 정확한 결과는 통계 프로그래밍 언어인 R을 이용하여 구하였다.

Table 2.1 Results of the clinical depression study in Example 2.1.4

	Treatment group				
Response	Imipramine	Lithium	Combination	Placebo	Total
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

R Code in Example 10.4.3 of χ^2 test of Homogeneity.

Results

```
> df_Clinical
  Imipramine Lithium Combination Placebo
1         18      13          22      24
2         22      25          16      10


> # chi-square test of independence
> chisq.test(df_Clinical)

        Pearson's Chi-squared test

data:  df_Clinical
X-squared = 10.803, df = 3, p-value = 0.01284
```

R Code

```
# variables
Imipramine<-c(18,22)
Lithium<-c(13,25)
Combination<-c(22,16)
Placebo<-c(24,10)

# contingency table
df_Clinical=data.frame(Imipramine,Lithium,Combination,Placebo)
df_Clinical

# chi-square test of independence
chisq.test(df_Clinical)
```

▲

10.5 심슨의 역설 (Simpson's Paradox)

§. 심슨의 역설에 관한 예제 (An Example of the Paradox)

▷ (**개요**) : 심슨의 역설은 각 부분에 대한 평균이 크다고 해서 전체에 대한 평균까지 크지는 않다는 의미이다. 즉, $\frac{a_1}{A_1} > \frac{a_2}{A_2}$ 이고 $\frac{a_2}{A_2} > \frac{b_2}{B_2}$ 라 해서 반드시 $\frac{a_1+a_2}{A_1+A_2} > \frac{b_1+b_2}{B_1+B_2}$ 를 만족하는 것은 아니다. 이는 곧 각각의 변수에 신경 쓰지 않고 전체 통계 결과를 유추하다 일어나는 오류를 일컫는다.

▷ **(예제1)** : 눈, 코, 입, 귀 등 각각의 이목구비가 예쁘다고 하여 예쁜 얼굴이 나오는 것은 아니다.

▷ **(예제2)** : 수학과에서 900명의 학생을, 영어영문학과에서 100명의 학생을 모집하는 미래의 경희대학교에서, 남학생 1000명과 여학생 1000명이 지원했을 때, 지원자 수와 합격자 수가 다음과 같다고 하자.

	지원자	합격자	합격률
남학생	900명	720명	80%
여학생	200명	180명	90%

(a) 수학과에서의 합격자 비율

	지원자	합격자	합격률
남학생	100명	10명	10%
여학생	800명	90명	11.25%

(b) 영어역문학과에서의 합격자 비율

각각의 합격자 비율에 대한 위의 표를 보면, (a)와 (b) 둘 다 여학생의 합격률이 남학생의 합격률보다 높음을 확인할 수 있다. 이제 전체적인 합격률을 알아보면 다음과 같다.

	지원자	합격자	합격률
남학생	1000명	730명	73%
여학생	1000명	270명	27%

(c) 전체 모집단에서의 합격자 비율

전체 모집단에서의 합격자 비율 (c)를 살펴보면, 전체적으로 남학생의 합격률이 여학생의 합격률보다 훨씬 높음을 확인할 수 있다. 만약 이러한 비율을 보고 있는 화자가 두 과 모두 여학생의 합격률이 높으므로, 경희대학교의 이번 신입생 모집은 여학생의 합격률이 높다고 말한다면, 이는 심슨의 역설에 빠진 것이다. 혹은 그 반대로, 전체적인 통계만 확인하고 남학생의 합격률이 월등히 높다고 생각하여 수학과를 지원할 후배에게 수학과도 마찬가지로 남학생의 합격률이 더 높다고 이야기 한다면 이 또한 심슨의 역설에 빠진 것이다. 수학과의 남학생 합격률은 여학생의 그것보다 낮기 때문이다. 물론 이는 예시를 위한 상황이므로, 상식을 갖고 적절히 이해하도록 한다.

10.8 부호 검정과 순위 검정 (Sign and Rank Tests)

§. 부호 검정과 순위 검정(Sign test and Rank test)

▷ **(개요)** : 부호 검정은 표본들이 서로 관련되어 있는 경우, 짝지어진 두 개의 관찰치들의 크고 작음을 (+)와 (-)로 표시하여 그 개수를 가지고 두 개의 분포의 차이가 있는가에 대한 가설을 검증하는 비모수적 방법이다. 이 때, (+)나 (-)가 나올 확률이 동일하다는 가정 하에 이항분포를 이용하여 가설을 검정한다. 만약 두 부호 중 어느 한쪽이 지나치게 많이 나오면 귀무가설을 기각한다.

▷ **(특징)** : 적용하기가 쉽고, 가정들이 그다지 제한적이 아니다. 그러나 이 방법이 원점수들에 담겨 있는 특정한 정보들은 무시하기 때문에, 약간의 민감성(sensibility)를 잃게 한다는 것이다. 정규분포가정이 위반되었을 경우 t-test 대신 사용할 수 있다.

(1) 단일 표본(One-sample) 부호검정

하나의 모집단의 중심위치에 대하여 검정하기 위해 부호를 이용하는 방법. 데이터의 중앙값보다 큰 표본에 (+)를 주어 그 개수를 Y라 하면, Y는 이항분포를 따르게 되며, 이를 통해 검정을 수행하게 된다.

- **검정통계량** : 표본 중에서 μ_0보다 큰 표본의 수 (n^+로 두자)
- p**-값** : 주어진 유의수준보다 작으면 귀무가설을 기각

Ⅰ. 소표본인 경우 : 이항분포 이용

대립가설 H_1	$\mu > \mu_0$	$\mu < \mu_0$	$\mu \neq \mu_0$
$p-value$	$\sum_{x=n^+}^{n} \binom{n}{x} 0.5^x (1-0.5)^{n-x}$	$\sum_{x=0}^{n^+} \binom{n}{x} 0.5^x (1-0.5)^{n-x}$	$2 \times \sum_{x=0}^{n^+} \binom{n}{x} 0.5^n$

Ⅱ. 대표본인 경우 : 정규분포 이용

대립가설 H_1	$\mu > \mu_0$	$\mu < \mu_0$	$\mu \neq \mu_0$
$p-value$	$1-\Phi\left(\frac{n^+-0.5n-0.5}{\sqrt{0.25n}}\right)$	$\Phi\left(\frac{n^+-0.5n+0.5}{\sqrt{0.25n}}\right)$	$2\times\left\{1-\Phi\left(\frac{n^+-0.5n-0.5}{\sqrt{0.25n}}\right)\right\}$ or $2\times\left\{1-\Phi\left(\frac{n^+-0.5n+0.5}{\sqrt{0.25n}}\right)\right\}$

(2) 단일 표본(One-sample) Wilcoxon 검정

단일표본 부호검정은 부호만을 검정통계량에 이용하지만, Wilcoxon 검정은 부호에 순위(Rank)를 결부시켜 부호검정보다 큰 검정력을 갖도록 한 방법이다. 데이터가 연속형이고 대칭인 모집단에서 추출되었다고 가정하며, 모집단이 정규분포를 따를 경우 이 검정은 t-검정보다 검정력이 약간 떨어지며, 신뢰구간도 넓어진다. 다른 모집단에서는 검정력이 훨씬 높아지고 신뢰구간은 평균적으로 더 좁아질 수 있다. (**여기서 검정력(Power)이란 귀무가설을 기각할 확률을 말한다.)

- **검정통계량** : 가설의 중위수를 초과하는 왈쉬 평균 수 + 0.5 × (가설의 중위수와 같은 왈쉬 평균 수)
- p**-값** : 주어진 유의수준보다 작으면 귀무가설을 기각

$$E(W)=\frac{n(n+1)}{4} \quad , \quad Var(W)=\frac{n(n+1)(2n+1)}{24}$$

대립가설 H_1	$\mu > \mu_0$	$\mu < \mu_0$	$\mu \neq \mu_0$
$p-value$	$1-\Phi\left(\frac{W-E(W)-0.5}{\sqrt{Var(W)}}\right)$	$\Phi\left(\frac{W-E(W)+0.5}{\sqrt{Var(W)}}\right)$	$2\times\left\{1-\Phi\left(\frac{W-E(W)-0.5}{\sqrt{Var(W)}}\right)\right\}$ or $2\times\left\{1-\Phi\left(\frac{W-E(W)+0.5}{\sqrt{Var(W)}}\right)\right\}$

(3) Mann-Whitney 검정

두 모집단 사이의 동일성을 검정하기 위한 비모수적 방법이다. Mann-Whitney 검정에서는 두 모집단이 서로 같은 형상 모수를 갖는다고 가정한다. 그리고 두 모집단 사이의 위치 모수가 서로 같은지를 검정하는 방법이다. Mann-whitney 검정에서는 데이터가 동일한 형태를 갖고 분산도 동일하며 연속형 또는 순서형 척도를 갖는 두 모집단에서 추출된 독립적인 두 확률 표본이라고 가정한다. 모집단이 정규분포를 따를 때에 이표본(Two-sample) 검정보다 검정력이 더 떨어지지만 대부분의 다른 모집단에서는 검정력이 훨씬 더 높다(신뢰구간이 평균적으로 더 좁다). 모집단의 형태 또는 표준편차가 서로 다른 경우에는 합동 분산을 사용하지 않는 이표본 t-검정이 더 적당할 수 있다.

- **검정통계량** : 첫 표본의 순위 합
- p**-값** : 주어진 유의수준보다 작으면 귀무가설을 기각

1) 순위에 동점이 없는 경우

$$E(W)=\frac{n_1(N+1)}{2},\quad Var(W)=\frac{n_1n_2(N+1)}{12}$$

2) 순위에 동점이 있는 경우

$$E(W)=\frac{n_1(N+1)}{2},\quad Var(W)=\frac{n_1n_2}{N(N-1)}\sum_{i=1}^{N}R_i-\frac{n_1n_2(N+1)^2}{4(N-1)}$$

대립가설 H_1	$\mu>\mu_0$	$\mu<\mu_0$	$\mu\neq\mu_0$
$p-value$	$1-\Phi\left(\frac{W-E(W)-0.5}{\sqrt{Var(W)}}\right)$	$\Phi\left(\frac{W-E(W)+0.5}{\sqrt{Var(W)}}\right)$	$2\times\left\{1-\Phi\left(\frac{W-E(W)-0.5}{\sqrt{Var(W)}}\right)\right\}$ or $2\times\left\{1-\Phi\left(\frac{W-E(W)+0.5}{\sqrt{Var(W)}}\right)\right\}$

(4) 모수적 방법과 비모수적 방법의 검정 비교표

구분	모수 검정	비모수 검정
When to use	정규분포 가정 만족 시	정규분포 가정 불충족 시, 혹은 모집단이 어떤 분포를 따르는지 모를 때
Statistic	평균(Mean)	중앙값(Median)
1 sample	1 sample t-test	1 sample Wilcoxon signed rank test
2 samples	2 sample t-test	Wilcoxon rank sum test, Mann-Whitney test
	paired 2 sample t-test	Wilcoxon signed rank test
more than 2 samples	one-way ANOVA	Kruskal-Wallis test

11.1 최소제곱법(The Method of Least Squares)

§. 직선접합(Fitting a Straight Line)

Theorem 11.1.1 최소제곱(Least Squares)
$(x_1,y_1),\ldots,(x_n,y_n)$을 n개의 점들의 집합이라 하자. 모든 점 $(x_1,y_1),\ldots,(x_n,y_n)$에 대하여 가장 가까운 직선, 즉 각 점과 직선 사이의 거리를 최소화하는 직선은 다음과 같은 기울기와 절편을 갖는다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \ , \ \hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} \ \text{ where } \ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, \ \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \ .$$

(Proof) 임의의 직선 $y=\beta_0+\beta_1 x$를 고려하자. 이 때 변수 x를 제외한 나머지는 고정된 상수이다. $x=x_i$로 두어 각 y_i에 대한 직선 접합을 고려할 것이다. 이제 다음과 같은 직선 y와 점 (x_i,y_i) 사이의 수직 거리에 관한 n개 제곱합을 생각하자.

$$Q=\sum_{i=1}^{n}\left[(y_i-(\beta_0+\beta_1 x_i)\right]^2$$

이러한 제곱합 Q를 β_0과 β_1에 관하여 최소화하고자 한다. 이는 Q를 β_0과 β_1에 관하여 편미분한 식을 0으로 두고 방정식을 푸는 것과 같다.

$$\frac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0 \ , \ \frac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0$$

위의 편미분된 방정식을 풀면, 다음과 같은 방정식계를 얻는다.

$$\beta_0 n+\beta_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$$
$$\beta_0\sum_{i=1}^{n}x_i+\beta_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_i y_i$$

위와 같은 방정식계를 β_0과 β_1에 관한 ***정규방정식(normal equation)***이라 한다. 이제 Q에 대한 편미분을 한 번 더 진행하여 위 정규방정식을 만족하는 β_0과 β_1을 구하면 그러한 β_0과 β_1가 Q를 최소화하는 값임을 찾을 수 있고, 따라서 정리 11.1.1의 결과를 얻는다.

▲

Definition 11.1.1 최소제곱선(Least-Squares Line)
최소제곱법을 적용하여 얻은 β_0과 β_1를 각각 $\hat{\beta}_0$, $\hat{\beta}_1$라 두고, 이러한 $\hat{\beta}_0$, $\hat{\beta}_1$에 관한 직선의 방정식 $y=\hat{\beta}_0+\hat{\beta}_1 x$를 **최소제곱선**이라 부른다.

▷ **예제 11.1.1~2 : Blood Pressure.**

10명의 환자가 두 개의 다른 혈압에 관한 약품을 동일한 양으로 투여 받았다고 가정하자. 첫 번째로 일반 의약품인 A를 투여한 후 혈압을 측정하였고, 약의 효능이 다한 후 신약인 B를 투여한 후 혈압을 측정하여 이를 기록하였다. 이 때 각 환자가 약품을 투여 받고 변화된 혈압의 변화를 반응(reaction)이라 한다. 우리는 이러한 반응을 예측하고 싶어 한다. 자세하게 말하면, 기존에 알려진 약물 A의 반응을 가지고 약물 B의 반응을 예측하고 싶다. 두 약물에 대한 반응을 기록한 표는 아래와 같이 주어진다.

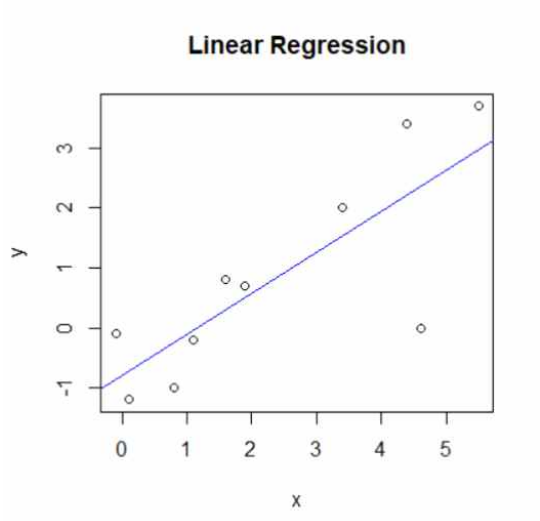
Table 11.1 Reactions to two drugs

i	x_i	y_i
1	1.9	0.7
2	0.8	-1.0
3	1.1	-0.2
4	0.1	-1.2
5	-0.1	-0.1
6	4.4	3.4
7	4.6	0.0
8	1.6	0.8
9	5.5	3.7
10	3.4	2.0

Table 11.1로부터, 전체 관측 수 $n=10$이고, 정리 11.1.1에 의해 $\hat{\beta}_0=-0.786$, $\hat{\beta}_1=0.685$으로 구할 수 있다. 따라서 이 관측값들에 대한 최소제곱선은 $y=-0.786+0.685x$로 구해진다.

R Code in Example 11.1.1~2

Results




```
> model

Coefficients:
(Intercept)            x
    -0.7861       0.6850
```

R Code

```
# Table
drug_A<-c(1.9,0.8,1.1,0.1,-0.1,4.4,4.6,1.6,5.5,3.4)
drug_B<-c(0.7,-1.0,-0.2,-1.2,-0.1,3.4,0.0,0.8,3.7,2.0)
df_blood=data.frame(drug_A,drug_B)
df_blood

# Variables
x=df_blood$drug_A
y=df_blood$drug_B

# Simple Linear Model
model=lm(y ~ x)
plot(x,y,main="Linear Regression")
abline(model,col="blue")
```

§. 최소제곱법을 이용한 다항식 접합 (Fitting a Polynomial by the Method of Least Squares)

앞서 논의하였던 직선접합에 대한 일반화로, 최소제곱법을 이용하여 다항식의 경우도 일반화할 수 있다. 즉, $y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_k x^k$로 일반화할 수 있다. 이를 구하는 방법은, 마찬가지로 점과 임의의 다항식 사이의 제곱합 Q를 최소화하는 계수 값을 찾으면 된다. 이 때 정규방정식은 $k+1$개의 Q에 대한 편미분을 푸는 것으로써, $k+1 \times k+1$ 행렬이 된다. 일련의 과정을 거치면, 최소제곱법을 이용한 다항식 접합을 $y=\hat{\beta}_0+\hat{\beta}_1 x+\hat{\beta}_2 x^2+\cdots+\hat{\beta}_k x^k$ 로 얻을 수 있다.

▷ **예제 11.1.3 : Fitting a Parabola.**

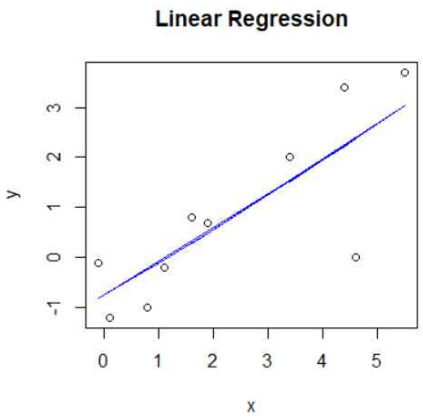
앞선 혈압과 약물 복용에 대한 예제 11.1.1을 상기하자. 이전 예제에서 우리는 산점도에 표시된 점들을 최적근사하는 선을 직선접합으로 택하였다. 이번 예제에서는 이를 곡선접합, 즉 다항식 접합으로 근사해보고자 한다. 다항식은 포물선의 형태, $y=\beta_0+\beta_1 x+\beta_2 x^2$로 근사하고자 한다. 이전 예제에서 했던 가정들을 그대로 이용하여, 최소제곱해인 각 β_i에 대한 값들을 구하려면 다음 정규방정식을 풀어야한다.

$$
\begin{aligned}
10\beta_0+23.3\beta_1+90.37\beta_2 &= 8.1\\
23.3\beta_0+90.37\beta_1+401.0\beta_2 &= 43.59\\
90.37\beta_0+401.0\beta_1+1892.7\beta_2 &= 204.55
\end{aligned}
$$

이러한 정규방정식을 만족하는 각 β_i의 값은 $\hat{\beta}_0=-0.744$, $\hat{\beta}_1=0.616$, $\hat{\beta}_2=0.013$으로 구할 수 있다. 따라서 최소제곱 포물선 $y=-0.744+0.616x+0.013x^2$ 을 얻는다.

R Code in Example 11.1.3

Results




```
> model

Coefficients:
(Intercept)          x      new_x
    -0.7451     0.6186     0.0126
```

R Code

```
# Table
drug_A<-c(1.9,0.8,1.1,0.1,-0.1,4.4,4.6,1.6,5.5,3.4)
drug_B<-c(0.7,-1.0,-0.2,-1.2,-0.1,3.4,0.0,0.8,3.7,2.0)
df_blood=data.frame(drug_A,drug_B)
df_blood

# Variables
x=df_blood$drug_A
y=df_blood$drug_B
new_x=x^2

# Simple Linear Model
model=lm(y ~ x + new_x)
plot(x,y,main="Linear Regression")
lines(x,fitted(model),col="blue")
```

§. 다변수에서의 선형함수 접합 (Fitting a Linear Function of Several Variables)

앞서 논의하였던 다항식 접합과 마찬가지로, 최소제곱법을 이용하여 선형함수의 경우로 일반화할 수 있다. 즉, $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k$로 일반화할 수 있다. 이를 구하는 방법은, 마찬가지로 점과 임의의 다항식 사이의 제곱합 Q를 최소화하는 계수 값을 찾으면 된다. 이 때 정규방정식은 $k+1$개의 Q에 대한 편미분을 푸는 것으로써, $k+1 \times k+1$ 행렬이 된다. 일련의 과정을 거치면, 최소제곱법을 이용한 선형함수의 접합을 $y=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_k x_k$ 로 얻을 수 있다.

▷ **예제 11.1.5 : Fitting a Linear Function of Two Variables.**

혈압에 관한 이전 예제 11.1.1을 참고하자. 예제 11.1.1의 내용 중, 설명변수인 심박수 x_{i2}를 추가할 것이다. 우리는 반응 y_i를 예측하기 위해서 곡선접합을 사용하여 점들의 추세선이 되는 함수를 구할 것이다. 즉, $y=\beta_0+\beta_1 x_1+\beta_2 x_2$인 선형함수의 계수들을 구하고자 한다. 다음 표를 참조하자.

Table 11.2 Reactions to two drugs and heart rate

i	x_{i1}	x_{i2}	y_i
1	1.9	66	0.7
2	0.8	62	-1.0
3	1.1	64	-0.2
4	0.1	61	-1.2
5	-0.1	63	-0.1
6	4.4	70	3.4
7	4.6	68	0.0
8	1.6	62	0.8
9	5.5	68	3.7
10	3.4	66	2.0

Table 11.2의 내용을 토대로 선형함수의 계수들을 구하기 위해 최소제곱법을 적용한다. 이 때, 정규방정식은 다음과 같이 구해진다.

$$\begin{aligned} 10\beta_0+23.3\beta_1+650\beta_2&=8.1\\ 23.3\beta_0+90.37\beta_1+1563.6\beta_2&=43.59\\ 90.37\beta_0+1563.6\beta_1+42334\beta_2&=563.1 \end{aligned}$$

이러한 정규방정식을 계수들에 관해 풀면 $\widehat{\beta_0}=-11.4527$, $\widehat{\beta_1}=0.4503$, $\hat{\beta}_2=0.1725$로 구할 수 있다. 따라서 최소제곱 선형함수는 $y=-11.4527+0.4503x_1+0.1725x_2$로 구해진다. R 프로그래밍을 이용하여 구한 결과는 아래에 나타내었다.

R Code in Example 11.1.5

Results

```
> df_blood
   drug_A heart_rate drug_B
1     1.9         66    0.7
2     0.8         62   -1.0
3     1.1         64   -0.2
4     0.1         61   -1.2
5    -0.1         63   -0.1
6     4.4         70    3.4
7     4.6         68    0.0
8     1.6         62    0.8
9     5.5         68    3.7
10    3.4         66    2.0

> model

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
   -11.4528       0.4503       0.1725
```

R Code

```
# Table
drug_A<-c(1.9,0.8,1.1,0.1,-0.1,4.4,4.6,1.6,5.5,3.4)
heart_rate<-c(66,62,64,61,63,70,68,62,68,66)
drug_B<-c(0.7,-1.0,-0.2,-1.2,-0.1,3.4,0.0,0.8,3.7,2.0)
df_blood=data.frame(drug_A,heart_rate,drug_B)
df_blood

# Variables
x1=df_blood$drug_A
x2=df_blood$heart_rate
y=df_blood$drug_B

# Simple Linear Model
model=lm(y ~ x1 + x2)
```

▲

11.2 회귀(Regression)

§. 회귀함수(Regression functions)

Definition 11.2.1 **반응변수/예측변수/회귀** (Response/Predictor/Regression)
회귀분석과 관련한 통계적 문제에서, 변수 $X_1, ..., X_k$을 **예측변수(Predictor)** 라고 부르며, 이에 대한 확률변수 Y를 **반응변수(Response)** 라고 부른다. $X_1, ..., X_k$의 관측값 $x_1,...,x_k$이 주어진 Y에 대한 조건부 기댓값, 즉 $E(Y|x_1,...,x_k)$을 Y에 대한 **회귀함수(Regression function)** 또는 $X_1, ..., X_k$에서의 Y에 대한 **회귀(Regression)** 이라 부른다.

▷이 장에서는 $E(Y|x_1,...,x_k)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$로 생각하자. 이 때, 이 선형함수의 각 계수를 **회귀계수(regression coefficients)**라고 부르며, 이들은 알려져 있지 않은 것으로 간주한다.

▷ **예제 11.2.1 : Pressure and the Boiling Point of Water.**
1857년, Forbes는 고도(altitude)를 추정하기 위한 방법을 얻기 위한 실험의 결과를 보고하였다. 그러한 고도를 측정하는 방법은 기압계 상의 압력을 측정하는 것으로 얻을 수 있으나, Forbes가 살던 시기에는 높은 고도로 기압계를 가져가기 어려웠던 상황이었다. 그러나 현대에는 많은 여행객들이 산을 오를 때 가지고 다니는 온도계를 소지하고 물의 끓는점을 측정하는 것으로 고도를 손쉽게 구할 수 있다. 다음 Table 11.5 에서 이러한 끓는점과 압력을 나타내었다.

Table 11.5 Boiling point of water in degrees Fahrenheit and atmospheric pressure in inches of mercury from Forbes' experiments.

Boiling Point	Pressure
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

우린 여기서 끓는점과 압력 사이의 선형관계를 접합하기 위해 최소제곱법을 이용할 수 있다. y_i를 Forbes가 관측한 관측값들 중 어느 값이라 하고 x_i를 끓는점이라 하자. 이 때, $i=1,...,17$이다. Table 11.5의 데이터들을 이용하면, 우린 최소제곱선을 계산할 수 있다. 그러한 최소제곱선을 계산하면 절편과 기울기는 각각 $\hat{\beta}_0 = -81.049$와 $\hat{\beta}_1 = 0.5228$로 구해진다. 따라서 최소제곱선 $y = -81.049 + 0.5228x$로 구할 수 있다. 물론, 우리는 이러한 최소제곱선이 끓는점과 압력 사이의 관계를 자세히 제시할 수 없다는 것을 알고 있다. 만약 우리가 물의 끓는점 x를 계속해서 얻을 수 있고 알려지지 않은 압력 Y에 대한 조건부 분포를 계산하길 원한다면, 그러한 Y를 계산할 수 있게 해줄 통계적 모델의 존재 여부가 궁금할 것이다.

▲

§. 단순선형회귀(Simple Linear Regression)

단순선형회귀에서는 회귀식이 단일 변수인 X에 대한 회귀 Y에 관한 식으로 간주한다. 우리는 각 설명변수가 $X=x$로, 확률변수 Y는 $Y=\beta_0+\beta_1 x+\epsilon$, where random variable $\epsilon \sim N(0,\sigma^2)$ 로 가정한다. 이 장에서 우린 각 점 (x_1, Y_1), ..., (x_n, Y_n)에 대한 회귀 문제를 다루게 될 것이다. 다음의 가정들은 단순선형회귀를 포함한 여러 개의 예측변수들(설명변수)에서의 회귀를 위해 반드시 필요한 가정들이므로 잘 숙지해야 한다.

Assumption 11.2.1 예측변수가 알려져 있다 (Predictor is known)
값 $x_1, ..., x_n$이 미리 알려져 있거나 $(Y_1, ..., Y_n)$의 결합분포를 계산하기 전에 확률변수 $X_1, ..., X_n$의 관측 값이어야 한다.

Assumption 11.2.2 정규성 (Normality)
$i=1,...,n$에 대하여, $x_1,...,x_n$이 주어진 Y_i의 조건부 분포가 정규분포여야 한다.

Assumption 11.2.3 선형성을 갖는 평균 (Linear Mean)
$i=1,...,n$에 대하여, $E(Y_i|x_1,...,x_n)$이 $\beta_0+\beta_1 x$ 의 형태인 모수 β_0과 β_1이 존재함을 말한다.

Assumption 11.2.4 등분산성 (Homoscedasticity)
$i=1,...,n$에 대하여, $Var(Y_i|x_1,...,x_n)=\sigma^2$인 모수 σ^2가 존재함을 말한다. 여기서 다른 분산을 갖는 확률변수는 **이분산성(Heteroscedasticity)**를 갖는다고 말한다.

Assumption 11.2.5 독립성 (Independence)
관측 값 $x_1,...,x_n$가 주어진 확률변수 $Y_1,...,Y_n$이 독립적임을 일컫는다.

앞서 설명한 가정 11.2.1-11.2.5는 $\vec{x}=(x_1,\ ...,x_n)$이 주어진 조건부 결합 분포 $Y_1,...,Y_n$을 특정하고, 모수(매개변수) β_0, β_1, σ^2를 특정한다. 특히 조건부 결합 p.d.f.인 $Y_1,...,Y_n$는 다음과 같이 표현할 수 있다.

$$f_n(\vec{y}\,|\,\vec{x},\beta_0,\beta_1,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\right]$$

우린 이러한 β_0,β_1,σ^2의 최대우도추정량을 찾을 것이다.

Theorem 11.2.1 단순선형회귀의 최대우도추정량 (Simple Linear Regression M.L.E.'s)
Assumption 11.2.1-11.2.5를 만족한다고 가정하자. β_0, β_1의 최대우도추정량은 최소제곱추정량이고, σ^2의 최대우도추정량은 다음과 같다.

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

§. 최소제곱추정량의 분포 (The Distribution of the Least-Squares Estimators)

Theorem 11.2.2 최소제곱추정량의 분포 (Distribution of Least-Squares Estimators)
Assumption 11.2.1-11.2.5를 만족한다고 가정하자. 그러면 최소제곱추정량 $\hat{\beta}_1$의 분포는 평균 β_1과 분산 σ^2/s_x^2를 갖는 정규분포를 따른다. 또한, $\hat{\beta}_0$의 분포는 평균 β_0과 분산 $\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{s_x^2}\right)$ 을 갖는 정규분포를 따른다. 마지막으로, 두 최소제곱추정량의 공분산(covariance)은 다음과 같다.

$$Cov(\hat{\beta}_0,\hat{\beta}_1)=\frac{\bar{x}\,\sigma^2}{s_x^2}$$

이 정리에서의 모든 분포에 관한 설명은 확률변수 X_i에 대한 조건부로 $X_i=x_i$을 가짐을 유의한다.

▷ **예제 11.2.2 : Pressure and the Boiling Point of Water.**
이전 예제 11.2.1을 참고하자. 우린 앞서 구하였던 최소제곱추정량들의 분산과 공분산을 계산하고자 한다. 우선 평균 온도는 $\bar{x}=202.95$, $s_x^2=530.78$, 전체 관측 값의 개수는 $n=17$로 구할 수 있다. 우린 σ^2를 알지 못하므로, 이를 제하고 정리 11.2.2를 이용하여 분산과 공분산을 계산해본다. 그 결과는 아래와 같다.

$$Var(\hat{\beta}_1)=\frac{\sigma^2}{530.78}=0.00188\sigma^2,\quad Var(\hat{\beta}_0)=\sigma^2\left(\frac{1}{17}+\frac{202.95^2}{530.78}\right)=77.66\sigma^2,$$

$$Cov(\hat{\beta}_0,\hat{\beta}_1)=-\frac{202.95\,\sigma^2}{530.78}=0.382\sigma^2.$$

이 결과들을 잘 살펴보면, β_0보다 β_1의 추정치가 더 자세한 값임을 예상할 수 있다.

▲

▷ **예제 11.2.3 : The Variance of a Linear Combination.**

회귀분석과 관련한 통계학적 문제에서, 최소제곱추정량의 선형결합의 분산을 자주 계산할 필요가 있다. 그러한 하나의 예로써 '예측'에 관한 문제가 있다. 우리는 $T = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + c_*$의 분산을 계산하고 싶다고 가정하자. T의 분산은 다음과 같이 $Var(\hat{\beta}_0)$, $Var(\hat{\beta}_1)$, $Cov(\hat{\beta}_0, \hat{\beta}_1)$의 값으로 대체함으로써 찾아질 수 있다.

$$Var(T) = c_0^2 Var(\hat{\beta}_0) + c_1^2 Var(\hat{\beta}_1) + 2c_0c_1 Cov(\hat{\beta}_0, \hat{\beta}_1)$$

이는 앞서 정의하였던 $Var(\hat{\beta}_0)$, $Var(\hat{\beta}_1)$, $Cov(\hat{\beta}_0, \hat{\beta}_1)$의 정의를 대입하여 정리하면 다음과 같이 쓸 수 있다.

$$Var(T) = \sigma^2\left(\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right)$$

예제 11.2.2의 특별한 경우로, 우리는 $c_0 = 0$, $c_1 = 200$으로 두었다. 따라서 $200\hat{\beta}_1$의 분산은 $200^2\sigma^2/s_x^2 = 75.36\sigma^2$로 구해진다. 이는 $\hat{\beta}_0$의 분산 값 $77.66\sigma^2$와 매우 가까운 값으로 구해짐을 확인할 수 있다.

▲

▷ **예제 11.2.5 : Pressure and the Boiling Point of Water.**

예제 11.2.4에서, 우리는 물의 끓는 점이 201.5도일 때의 기압을 예측하길 원했다. 최소제곱선은 $y = -81.049 + 0.5228x$로 구해지고, $\hat{\sigma}^2 = 0.0478$로 구하였다. 기압 Y의 예측값의 M.S.E.(Mean Squared Error)는 다음으로부터 얻어진다.

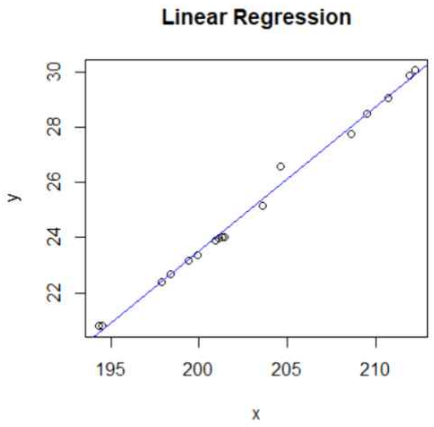
$$E[(\hat{Y} - Y)^2] = \sigma^2\left[1 + \frac{1}{17} + \frac{(201.5 - 202.95)^2}{530.78}\right] = 1.0628\sigma^2$$

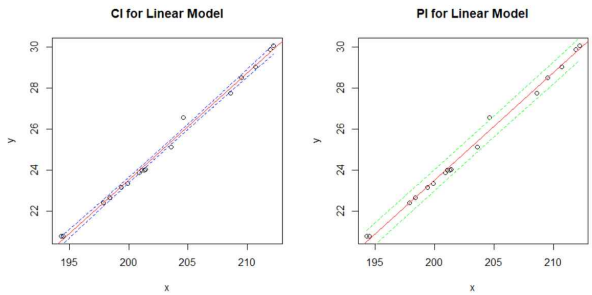
또한, 예측 $\hat{Y}$의 값은 $\hat{Y} = -81.06 + 0.5229 \times 201.5 = 24.30$으로 관측된다. M.S.E.의 값 $1.0628\sigma^2$는 다음과 같이 해석될 수 있다. : 만약 우리가 β_0과 β_1의 값을 알고 있고, Y를 예측하고자 한다면, M.S.E.는 $Var(Y) = \sigma^2$가 될 것이다. β_0과 β_1을 추정할 때 우리가 가져야할 것은 M.S.E.에서 오직 $0.0628\sigma^2$ 뿐이다.

아래에 R프로그래밍으로 이러한 예측값을 구하고 회귀선을 구한 코드를 제시하였다.

R Code in Example 11.2.5

Results






```
> # Prediction of specific value
> Linefunc(201.5)
[1] 24.29909
```

R Code

```
# Table
Boil<-c(194.5,194.3,197.9,198.4,199.4,199.9,200
.9,201.1,201.4,201.3,203.6,204.6,209.5,208.6,210
.7,211.9,212.2)
Pres<-c(20.79,20.79,22.40,22.67,23.15,23.35,23.
89,23.99,24.02,24.01,25.14,26.57,28.49,27.76,29.
04,29.88,30.06)
df_water=data.frame(Boil,Pres)
df_water
# Variables
x=df_water$Boil
y=df_water$Pres
# Simple Linear Model
model=lm(y ~ x)
plot(x,y,main="Linear Regression")
abline(model,col="blue")
#Linear function
Linefunc<-function(x){
-81.0637271+0.5228924*x}
# Prediction of specific value
Linefunc(201.5)
par(mfrow=c(1,2))
# 95% Confidence Inverval
newx=seq(min(x),max(x),by=0.05)
CI=predict(model,newdata=data.frame(x=newx),
interval="confidence",level=0.95)
plot(x,y,main="CI for Linear Model")
abline(model,col="red")
lines(newx,CI[,2],col="blue",lty=2)
lines(newx,CI[,3],col="blue",lty=2)
# 95% Prediction Inverval
newx=seq(min(x),max(x),by=0.05)
CI=predict(model,newdata=data.frame(x=newx),
interval="prediction",level=0.95)
plot(x,y,main="PI for Linear Model")
abline(model,col="red")
lines(newx,CI[,2],col="green",lty=2)
lines(newx,CI[,3],col="green",lty=2)
```

▲

11.3 단순선형회귀의 통계적 추론(Statistical Inference in Simple Linear Regression)

§. 추정량의 결합분포(Joint Distribution of the Estimators)

▷ **예제 11.3.1 : Pressure and the Boiling Point of Water.**

예제 11.2.4에서, 끓는 점이 201.5일 때의 기압을 단순선형회귀를 이용하여 예측해 보았다. 여행객들이 201.5도에서 기압이 24.5인지 그 여부를 알고 싶어 한다고 가정하자. 즉, 다음과 같은 가설을 검정하고 싶어 한다.

$$H_0 : \beta_0 + 201.5\beta_1 = 24.5$$

혹은 그 대신에, $\beta_0 + 201.5\beta_1$의 구간 추정치를 알고 싶다고 하자. 이러한 추론들은 회귀 모형의 모든 모수들 $(\beta_0, \beta_1, \sigma^2)$의 추정량의 결합분포를 찾는 것으로 가능하다.

▲

Theorem 11.3.1

확률변수 $Y_1, \ldots, Y_n$이 독립이고, 각각이 모두 분산 σ^2를 갖는 정규분포를 따른다고 가정하자. $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}$인 확률변수의 벡터라고 하자. 만약 A가 $n \times n$ 직교행렬이고 $Z = AY$라 하면, 확률변수 $Z_1, \ldots, Z_n$ 또한 독립이고 각각이 모두 분산 σ^2를 갖는 정규분포를 따른다.

Theorem 11.3.2

단순선형회귀의 문제에서, $(\hat{\beta}_0, \hat{\beta}_1)$의 결합분포는 각각이 다음의 평균과 분산을 갖는 이변수(bivariate) 정규분포를 갖는다.

$$\hat{\beta}_1 \text{ 의 분포 : 평균 } \beta_1 \text{ , 분산 } \sigma^2/s_x^2$$

$$\hat{\beta}_0 \text{ 의 분포 : 평균 } \beta_0 \text{과 분산 } \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2}\right)$$

또한, $n \geq 3$이면, $\hat{\sigma}^2$는 $(\hat{\beta}_0, \hat{\beta}_1)$와 독립적이고 $n\hat{\sigma}^2/\sigma^2$는 자유도가 $n-2$인 χ^2 분포를 갖는다.

§. 회귀계수에 대한 가설검정(Tests of Hypotheses about the Regression Coefficients)

β_0과 β_1의 선형결합에 대한 가설 검정

c_0, c_1, c_*은 c_0와 c_1이 0이 아닌 특정된 값들이라 하자. 우리는 다음과 같은 가설을 검정하고자 한다.

$$H_0 : c_0\beta_0 + c_1\beta_1 = c_*$$
$$H_1 : c_0\beta_0 + c_1\beta_1 \neq c_*$$

우리는 확률변수 $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$과 σ'를 기반으로 이러한 가설의 검정을 유도하고자 한다.

▷ **단측 가설검정(One-Sided Test)**의 경우는 앞서 해왔던 여러 단측 검정과 동일하게 설정할 수 있다. 이 때, 단측 검정의 귀무가설의 부등호와 반대인 $U_{01} \geq T_{n-2}^{-1}(1-\alpha_0)$ 또는 $U_{01} \leq T_{n-2}^{-1}(1-\alpha_0)$인 경우 귀무가설을 기각한다.

Theorem 11.3.3

각 $0 < \alpha_0 < 1$ 에 대하여, 위와 같은 가설에 대한 수준 α_0 검정은 다음과 같은 조건일 때 귀무가설 H_0을 기각한다.

$$|U_{01}| \geq T_{n-2}^{-1}(1-\alpha_0/2) \text{ where } U_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2}\left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\sigma'}\right).$$

T_{n-2}^{-1} 는 $n-2$ 자유도인 t 분포의 분계함수(quantile function) 이다.

β_0에 대한 가설검정

β_0^*은 특정된 값으로 $-\infty < \beta_0^* < \infty$인 대소 관계를 갖는다고 하고, 다음과 같은 가설을 검정하고자 한다.

$$H_0 : \beta_0 = \beta_0^*$$
$$H_1 : \beta_0 \neq \beta_0^*$$

$c_0 = 1, c_1 = 0, c_* = \beta_0^*$로 두면, $U_0 = \dfrac{\hat{\beta}_0 - \beta_0^*}{\sigma' \left[\dfrac{1}{n} + \dfrac{\overline{x}^2}{s_x^2} \right]^{1/2}}$ 이고 귀무가설이 참일 때 마찬가지로 자유도 $n-2$인 t 분포를 갖는다. 물론, p-값을 구하는 것은 이전의 t 분포와 마찬가지로 구하면 된다.

만약 회귀 계수가 아니라 회귀선 자체가 원점을 지난다 vs 원점을 지나지 않는다 에 대한 가설 검정을 실시하려면 다음과 같이 가설을 설정하면 된다.

$$H_0 : \beta_0 = \beta_0^* = 0$$
$$H_1 : \beta_0 \neq \beta_0^* = 0$$

β_1에 대한 가설검정

β_1^*은 특정된 값으로 $-\infty < \beta_1^* < \infty$인 대소 관계를 갖는다고 하고, 다음과 같은 가설을 검정하고자 한다.

$$H_0 : \beta_1 = \beta_1^*$$
$$H_1 : \beta_1 \neq \beta_1^*$$

$c_0 = 0, c_1 = 1, c_* = \beta_0^*$로 두면, $U_1 = s_x \dfrac{\hat{\beta}_1 - \beta_1^*}{\sigma'}$ 이고 귀무가설이 참일 때 마찬가지로 자유도 $n-2$인 t 분포를 갖는다.

만약 회귀 계수가 아니라 회귀선의 X와 Y가 실제로 상관관계가 없음을 검정하고 싶으면 다음과 같이 가설을 설정하면 된다.

$$H_0 : \beta_1 = \beta_1^* = 0$$
$$H_0 : \beta_1 \neq \beta_1^* = 0$$

회귀선에 대한 가설검정

회귀선 $y=\beta_0+\beta_1 x$이 특정한 점 (x^*, y^*), $x^* \neq 0$을 지나는지 여부에 대한 가설검정을 하고자 한다. 그러면 다음과 같이 가설을 설정하면 된다.

$$H_0 : \beta_0+\beta_1 x^* = y^*$$
$$H_1 : \beta_0+\beta_1 x^* \neq y^*$$

이 가설은 $c_0=1$, $c_1=x^*$, $c_*=y^*$로 두고 U_{01}이 마찬가지로 $n-2$ 자유도인 t 분포를 따른다.

▷ **예제 11.3.3 : Pressure and the Boiling Point of Water.**

이제, 앞선 예제 11.3.1에서 검정하고자 했던 다음 가설을 검정해보자. 그 가설은 다음과 같이 세울 수 있다.

$$H_0 : \beta_0+201.5\beta_1 = 24.5$$
$$H_1 : \beta_0+201.5\beta_1 \neq 24.5$$

통계량을 $c_0=1$, $c_1=201.5$로 두어 구한다. 앞서 구하였던 최소제곱추정치는 $\hat{\beta}_0=-81.049$와 $\hat{\beta}_1=0.5228$ 이다. 또한 $n=17$, $s_x^2=530.78$, $\overline{x}=202.95$, $\sigma'=0.2328$로 구하였다. 따라서 $U_{01}=-0.2204$로 구할 수 있다. 만약 H_0이 참이면 U_{01}은 $n-2=15$ 자유도를 갖는 t 분포를 따른다. 검정통계량 -0.2204에 대응하는 p-값은 0.8285로 구할 수 있다. 따라서 귀무가설은 유의수준 α_0에 대하여 $\alpha_0 \geq 0.8285$ 일 때 기각한다.

▲

§. 신뢰 구간과 예측 구간(Confidence Intervals & Prediction Intervals)

Theorem 11.3.5

c_0, c_1 이 0이 아닌 상수라고 하자. 두 확률변수 β_0, β_1 사이의 열린구간은 다음과 같다.

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \pm \sigma' \left[\frac{c_0^2}{n} + \frac{(c_0\bar{x} - c_1)^2}{s_x^2} \right]^{1/2} T_{n-2}^{-1}\left(1 - \frac{\alpha_0}{2}\right)$$

이는 $1-\alpha_0$ **계수의** $c_0\beta_0 + c_1\beta_1$**에 대한 신뢰구간(Coefficient Interval)**이라 부른다.

Theorem 11.3.6 & Definition 11.3.1

단순선형회귀의 문제에서, Y를 Y_1, ..., Y_n이 독립일 때 설명변수 x로 인해 구해진 새로운 관측값이라 하자. 그리고 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$라 하자. 그러면 Y가 다음 두 확률변수 사이에 있을 확률은 $1-\alpha_0$이다. 그러한 두 확률변수는 다음과 같다.

$$\hat{Y} \pm T_{n-2}^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma' \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_x^2} \right]^{1/2}$$

여기서 이는 $1-\alpha_0$ **계수의** Y **에 대한 예측구간(Prediction Interval)**이라 부른다.

§. 잔차 분석(The Analysis of Residuals)

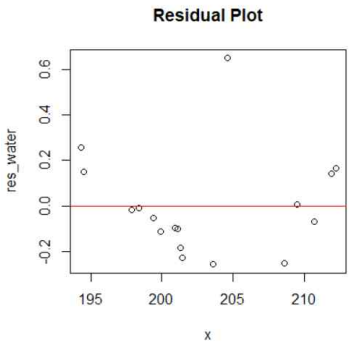
Definition 11.3.2 잔차/적합값

$i = 1, \ldots, n$에 대하여, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$의 관측값들을 **적합값(Fitted value)** 이라고 부르고, $e_i = y_i - \hat{y}_i$를 **잔차(Residuals)** 라고 부른다.

▷ **예제 11.3.6 : Pressure and the Boiling Point of Water.**

이전 같은 주제의 예제에서 소개한 Table을 참조하여, R 프로그래밍으로 구한 잔차 그래프로 대신한다.

Results

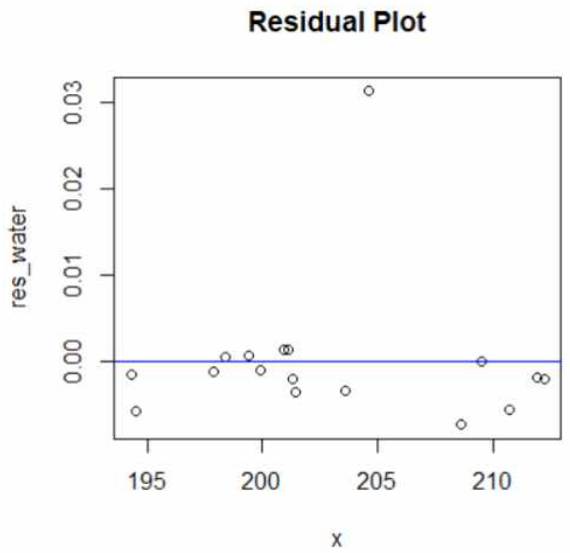


R Code

```
# Residual Plot
res_water=resid(model)
plot(x,res_water,main="Residual Plot")
abline(0,0,col="red")
```

잔차 그래프를 보면, 점들이 이상점(outlier)을 제외하고 U자 형태의 패턴으로 뿌려져 있는 것을 확인할 수 있다. 이는 구한 회귀선이 선형회귀보단 곡선이 더 적합함을 암시한다. 그러나 이는 수치상에서 나타나는 현상으로 실제로도 선형이 적합하지 않은지는 정확히 알 수 없다. 따라서 데이터를 로그를 취한 로그 스케일로 다시 분석해보는 것이 필요하다. 여기서 로그 스케일을 사용하는 이유는 나타난 실제 값들이 아닌 데이터의 본질을 파악하기 위함이다. 예를 들어 1000일 때 100000을 반환하는 함수가 있다고 하자. 그러면 이는 10일 때 1000을 반환하는 함수랑 같다. 그러나 수치상으로는 첫 번째가 훨씬 큰 값을 반환하므로 어떤 오류가 있다고 생각할 수 있다. 로그 스케일을 적용하는 것은 이러한 일들을 테크니컬하게 분별하기 위한 일종의 도구이다. 로그 스케일을 씌워서 만든 잔차 그래프는 아래와 같이 구해진다.

Results




```
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -0.97087      0.02062
```

R Code

```
# Table
Boil<-c(194.5,194.3,197.9,198.4,199.4,199.9,200
.9,201.1,201.4,201.3,203.6,204.6,209.5,208.6,210
.7,211.9,212.2)
Pres<-c(20.79,20.79,22.40,22.67,23.15,23.35,23.
89,23.99,24.02,24.01,25.14,26.57,28.49,27.76,29.
04,29.88,30.06)
Pres_lo<-log(Pres)
df_water=data.frame(Boil,Pres_lo)
df_water

# Variables
x=df_water$Boil
y=df_water$Pres_lo

# Simple Linear Model_log
model=lm(y ~ x)

# Residual Plot
res_water=resid(model)
plot(x,res_water,main="Residual Plot")
abline(0,0,col="blue")
```

▲

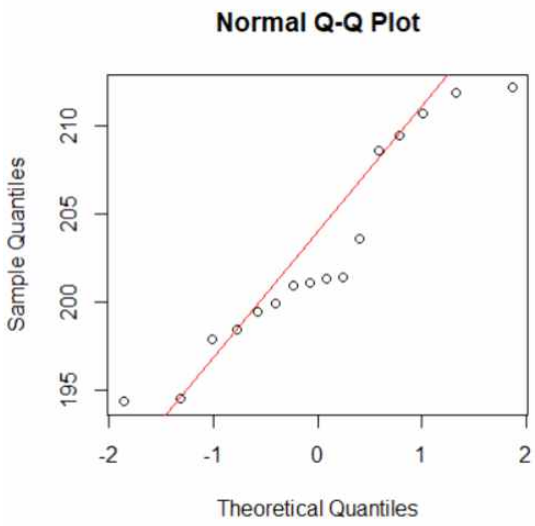
▷ **예제 11.3.7 : Pressure and the Boiling Point of Water.**

Normal Quantile Plot, 줄여서 **정규 Q-Q 플롯**에 대하여 잠시 소개하도록 하자. 자세한 배경은 교재를 참고한다. Q-Q (Quantile-Quantile) 플롯은 데이터가 특정 분포를 따르는지를 시각적으로 검토하는 방법이다. 예를 들어, 처음에 주어진 데이터가 정규분포를 따르는지 살펴보고 싶다고 가정한다. $X \sim N(\mu,\sigma^2)$이라면 표준화 하였을 때 $Z=(X-\mu)/\sigma \sim N(0,1)$임을 안다. 여기서 X를 Z에 대한 식으로 표현하면, $X=\mu+\sigma Z$ 이다. 따라서 X가 정규분포를 따를 때 (X,Z)를 좌표평면에 표시한다면, 식 $X=\mu+\sigma Z$은 직선의 형태로 평면상에 나타나게 될 것이다.

한편으로는, 이는 또한 잔차의 정규성을 확인할 때 사용하는 플롯으로, 이는 단순선형회귀의 여러 Assumption 중 하나인 오차가 $N(0,\sigma^2)$를 따른다는 것을 대략적으로 확인하는 과정이다. 이는 그래프 상에 직선과 점들이 얼마나 가까운지에 따라 데이터가 정규성을 띄는지 여부를 확인할 수 있다. 직선은 이론적으로 정규분포일 때 만들어져야 할 이상적인 직선이고($y=x$), 점은 실제 데이터가 따르는 위치이다. 이러한 점들이 직선에 가까울수록 데이터는 더욱 정규성을 띈다고 볼 수 있다.

이제 앞선 예제의 로그 스케일 데이터를 그대로 이용하여 Q-Q 플롯을 확인해보자. 통계 소프트웨어를 이용한다. 예제에서는 이상점이 되는 데이터 값(Table의 12번째 값)을 제거한 후 출력하였다.

Results



R Code

```
# Table
Boil<-c(194.5,194.3,197.9,198.4,199.4,199.9,200
.9,201.1,201.4,201.3,203.6,209.5,208.6,210.7,211
.9,212.2)
Pres<-c(20.79,20.79,22.40,22.67,23.15,23.35,23.
89,23.99,24.02,24.01,25.14,28.49,27.76,29.04,29.
88,30.06)
Pres_lo<-log(Pres)
df_water=data.frame(Boil,Pres_lo)
df_water

# Variables
x=df_water$Boil

# Q-Q Plot
qqnorm(x)
qqline(x,distribution=qnorm,col="red")
```

▲