고급통계적 방법론

0강: Introduction

Lecturer: 장원철 *2023* 가을학기

John Tukey (1915-2000)은 위상수학자출신의 대표적인 20세기의 통계학자중 한 명이다. Fast Fourier Transform을 제안했으며 비트(bit)와 소프트웨어라는 용어를 최초로 사용한 사람으로도 알려져 있다. John Tukey는 통계분석을 크게 탐색적 자료분석 (Exploratory Data Analysis)와 확증적 자료분석 (Confirmatory Data Analysis)로 나눌 수 있으며 대부분의 통계 방법론이 후자에 치우쳐 있다고 지적했다.

그렇다면 탐색적 자료분석과 확증적 자료분석의 차이는 무엇일까? 먼저 탐색적 자료분석은 데이터의 특징과 구조에 대한 탐구에 중점을 두고 있으며 이후 확증적 자료분석을 위한 가설과 모형을 도출하는 것을 목표로 하고 있다, 확증적 자료분석의 경우 이렇게 도출된 가설과 모형의 타당성을 검증하고 이를 위해 모형적합도, 가설검정, 신뢰구간과 같은 방법을 사용한다 . 이해를 돕기 위해 다음 2가지 예를 살펴보자[1].

- 감기에 걸리는 사람과 걸리지 않는 사람들간에 어떤 차이가 있는가를 수십가지 측면에서 살펴보았다, 그 결과 비타민 C를 복용하는 사람들이 감기에 잘 걸리지 않음을 알게 되었다. 그렇다면 비타민 C 복용이 감기를 예방하는 효과가 있다고 말할 수 있는가? 이러한 가설을 제기하는 건은 탐색적 자료분석의 영역이라고 할 수 있다, 반면에 이 가설에 대한 답변을 확증적 자료분석을 통해서 할 수 있다. 구체적으로 비교실험을 설계해서 새롭게 자료를 수집하여 가설검정의 단계를 거치게 된다
- 대형마켓에서 고객들의 구매내역 자료를 분석한 결과, 일부 고객들은 다른 고객에 비해 유기농 식재료 구매 비중이 크게 나타났다. 그들은 어떤 성향의 고객들인가? 여기에 대해 몇 개의 추측을 하는 것은 탐색적 자료분석의 몫이다. 하지만 추측이 맞는 자 여부를 확인하기 위해 전체 고객중 일부를 선택하여 몇가지 인구 사회적 속성과 연소득, 그리고 소비와 삶에 대한 태도를 조사하여 구매내역과 연결해 확인하는 것은 확증적 자료분석의 단계이다.

1763년 베이즈 법칙 출간을 통계학의 기원으로 삼는다면 대략 260년간의 통계학 역사의 대부분은 확증적 자료분석에 관한 연구에 치중되어 있었다. 하지만 최근 반세기 동안 컴퓨터의 등장으로 확증적 자료분석과 더불에 탐색적 자료분석에 관한 연구도 괄목할 만한 성장을 이루었다. 예를 들면 MCMC과 재표본 추출(resampling) 같은 방법들은 컴퓨터의 도움이 없다면 실제 데이터 분석에 사용되기 어려웠을 것이다.

이 과목에서는 이러한 컴퓨팅의 발전을 통해서 통계학분야가 지난 반세기동안 어떻게 발전했는지 알아본다. 특히 빅데이터의 시대를 맞이하여 탐색적 자료분석과 확증적 자료분석을 보다 일반화한 개념인 알고리즘과 통계적 추론의 발전과정에 대해서 공부한다. 알고리즘은 기본적으로 “어떻게 분석을 하는냐”를 초첨을 맞추고

[1] 허명회, 데이터 분석의 철학과 과학성, R User Conference 2015

있으며 이러한 경향은 특히 데이터 사이언스로 대표되는 분야에서 특히 강조되고 있다, 반대로 통계적 추론은 "왜 이렇게 분석을 해야 하는냐"에 대한 답변을 제공하기 위한 수학적 논리를 제공하는 분야라고 할 수 있다.

이 과목에서는 통계학의 주요 방법론을 연대별로 다음과 같이 크게 3단계로 나누어 다룰 예정이다.

Part I 컴퓨팅 도약 이전 시점(1950년대)을 중심으로 통계학 분야의 대표적인 3개의 학파, 베이지안(Bayesian), 빈도주의(Frequentist), 우도주의 (Fisherian)에 대해서 알아본다.

Part II 컴퓨터의 초창기 도입시기인 1950년대부터 1990년대까지 개발된 대표적인 통계 방법론. Resampling methods, 생존분석과 EM 알고리즘, 일반화선형모형, Empircal Bayes, MCMC에 대해 소개한다.

Part III 21세기 빅데이터의 시대에 등장한 대표적인 통계방법론, 다중비교와 벌점화 회귀분석, Random Forests, Neural Network, SVM, Post selection inference를 대해 공부한다.

Joh Tukey는 통계학의 가장 큰 장점이 관심이 있는 모든 분야와 협력할 수 있다는 의미로 "The best thing about statistics is that you get to play in everyone's backyard."을 남겼다. 21세기에는 통계학자는 더이상 뒷마당이 아니라 앞마당에서 데이터 시대의 주역으로 등장해야 할 것이다. 이 과목을 통해서 수강생들이 그러한 역할을 할 수 있도록 성장하기를 바란다.

고급통계적 방법론

1장: Algorithm and Inference

Lecturer: 장원철 *2023* 가을학기

통계학은 경험(데이터)로 배우는 것을 연구하는 학문이다. 이러한 데이터의 사례로 혜성의 괘도, 프로야구 개별선수의 타격기록등을 들 수 있다. 통계학에는 크게 두가지 대표 이론, 베이지안 (Bayesianism)과 빈도주의 (frequentism)가 있다. 물론 여기에 하나를 더하자면 우도주의 (Fisherianism)를 들 수 있다. 이 과목에서는 다양한 측면에서 통계학의 대표적인 2가지 (혹은 3가지) 이론들의 공통점과 차이점에 관해서 대해서 알아볼 예정이다.

본격적으로 주요이론들에 대해 알아보기 전에 통계적 관점에서 "알고리즘"과 "추론"의 차이점을 명확히 하자. 가장 많이 사용되는 통계량인 평균을 이용하여 두 가지 개념의 차이를 설명해보자. 우리가 다음과 같은 자료 $x_1, \ldots, x_n$를 관측했다고 가정하자. 예를 들면 전국 기초자치단체별 자동차 사고 발생율 (10만명당 사고건수)을 관측했다고 생각해보자. 이 경우 $n = 226$이며 평균은 다음과 같은 공식을 이용하여 계산할 수 있다.

$$\bar{x} = \sum_{i=1}^{n} x_i/n \tag{1.1}$$

그렇다면 이 값은 얼마나 정확한 것일까?[1] 이 질문에 대한 답변은 표본평균의 표준오차 (standard error)를 제시하는 것으로 대답할 수 있다.

$$\widehat{\text{se}} = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2/(n(n-1))\right]^{1/2} \tag{1.2}$$

통계학이 다른 학문과 가장 큰 차이점이라면 이렇게 추정치의 불확실성을 표현하기 위해 표준오차, 또는 신뢰구간을 제시한다는 점이다. 식 (1.1)은 표본평균을 계산하는 알고리즘이지만 표준오차(1.1)는 알고리즘에서 계산되는 평균의 불확실성에 대한 추론을 제공한다고 할 수 있다.

일반적으로 알고리즘은 통계학자가 실제로 계산하는 과정을 의미한다면 추론은 이러한 계산과정을 하는 근거를 제시한다고 할 수 있다. 많은 경우 알고리즘이 먼저 개발되고 이를 뒷받침하는 추론은 이후에 개발이 되곤한다. 예를 들면 doubly truncated data의 경우 분포함수의 nonparametric MLE에 관한 self-consistent estimator를 구하는 EM 알고리즘은 1976년 Turnbull에 의해 제안되었지만 이 추정치에 관한 이론적 성질은 최근에야 밝혀졌다. 물론 이 경우에서 이론적 성질을 이용하지 않고 bootstrap(알고리즘)을 이용하여 신뢰구간을 구할 수 있기때문에 알고리즘을 이용한 추론도 가능하다고 할 수 있다.

최근 데이터의 크기가 점점 증가하고 컴퓨팅의 속도도 점차 빨라짐에 따라 알고리즘이 차지하는 비중에 더욱더 늘어가는 추세이다. 하지만 이 과목에서는 토끼 (알고리즘)와 거북이(추론)의 경주에서 거북이의 측면을 강조하

[1] 이 질문은 약간 명확하지 않을 수 있다. 구체적으로 이 평균이 추정하고자 하는 것 (즉 모수, parameter)이 명확하지 않다. 만약 전국 자동차 사고 발생율을 추정하고 싶다면 weighted average를 사용하는 것이 정확하다.

고자 한다. 컴퓨팅의 발전은 앞의 double truncation의 예와 같이 추론도 같이 성장해 나갈 수 있도록 큰 역할을 하고 있다. 다음 절에서 2가지 예제를 통해서 컴퓨팅이 어떻게 통계분석을 변화시켰는지 알아보자.

1.1 A Regression Example

그림 1.1은 스탠포드 의대의 신장기능에 관한 자료의 산점도를 보여준다. 157명의 건강한 자원자를 대상으로 나이(age)를 x축으로, 신장기능점수(tot)를 y축으로 총 점수와 나이의 관계를 보여주고 있다. 그림에서 볼 수 있듯이 일반적으로 나이가 들수록 신장기능이 저하되는 것을 알수 있으며 기능의 저하되는 속도는 신장이식시 중요하게 고려해야 할 사항이다. 또 한가지 주목할 사실은 자원자의 상당수가 40대 이하이며 나이가 증가할수록 자원자의 숫자도 적어지는 경향을 보인다. 과거에는 60세이상일 경우 기증이 금지되었지만 기증자의 감소로 지금은 더 이상 나이에 제한을 두지 않는다. 그림 1.1에서 보여주는 직선는 선형회귀식이다.

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{1.3}$$

위의 선형회귀식은 다음과 같은 제곱합을 최소로 하는 회귀계수 (β_0, β_1)를 구하여 얻어진 식이다.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

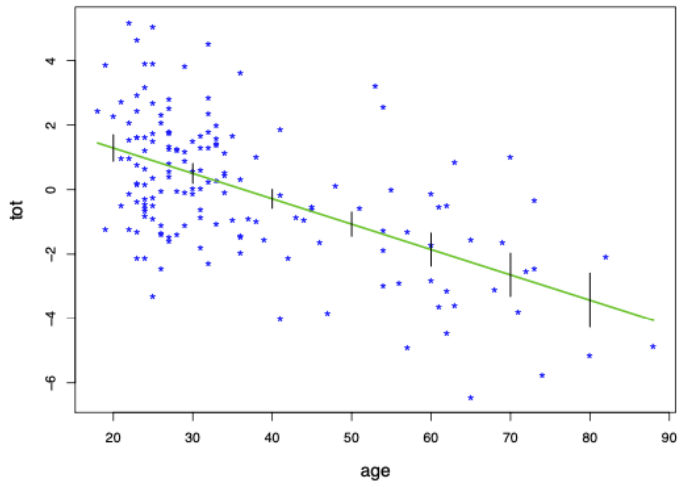

그림 1.1 157명의 자원자의 나이와 신장기능. 선형회귀직선과 일정나이에서의 신뢰구간 (±2 표준오차)을 보여준다.

최소제곱 (least square) 알고리즘은 1800년대 초반 르장드르 (Legendre)와 가우스(Gauss)에 의해 제안되었는데[2] 이 방법에 따르면 $\widehat{\beta}_0 = 2.86, \widehat{\beta}_1 = -0.079$로 계산된다. 식 (1.1)을 이용하면 특정 나이별로 신장기능점수을

[2]실제 발표는 르장드르가 1805년에 했지만 가우스가 본인이 혜성괘도 예측을 위해 1795년부터 사용했다고 주장했다. 최소제곱법과 정규분포와의 연관성과 표준오차의 계산등으로 일반적으로 가우스가 최소제곱법을 개발했다고 인정받고 있다.

예측할 수 있는데 20세일 경우 신장기능점수는 1.29 이지만 80세일 경우 예측치는 -3.43으로 감소함을 알 수 있다.

그렇다면 이러한 예측치는 얼마나 정확한가? 이 질문에 대한 답변을 하기 위해서는 예측치의 표준오차를 계산해야 하며 다행히 이미 1800년대에 가우스가 표준오차의 계산법을 제시하였다. 그림 1.1에서 특정 나이때별로 회귀직선위에 겹쳐보이는 수직선들은 그 예측지점에서는 ± 표준오차를 의미하며 95% 신뢰구간을 표시한다.

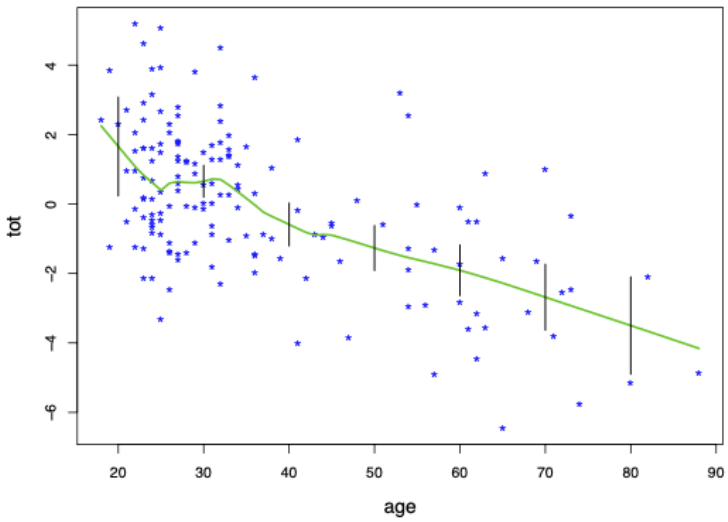

그림 1.2 Local polynomial lowess(x, y, 1/3)을 스탠포드 신장자료에 적합한 결과와 bootstrap을 이용하여 계산한 95% 신뢰구간

그림 1.2는 최신 컴퓨터기반 알고리즘 lowess를 이용하여 계산한 비선형예측곡선을 보여주고 있다. 실제로 R에서 lowess(x, y, 1/3)을 사용하여 위의 곡선을 계산했으며 여기서 1/3의 의미는 예측하고자 하는 각 나이에서 가까운 기준으로 1/3에 해당하는 데이터를 이용하여 예측치를 구한다는 뜻이다. lowess에 관해서는 잠시 후에 자세히 설명하기로 하자. 그림 1.2에서 발견할 수 있는 재미있는 사실은 20대 중반까지는 그림 1.1과 비슷하게 직선경향을 보이지만 20대중반부터 30대중반까지 기능예측치가 거의 변화가 없고 다시 이후에 감소 경향을 보인다는 점이다.

각 예측치의 표준오차를 구하기 위해서 뒤에서 배울 bootstrap이라는 방법을 사용하였다. Bootstrap의 핵심 아이디어는 resampling을 통한 표본분포 (sampling distribution)을 유추하는데 있다. 일반적으로 특정 통계량의 표본분포를 구하기 위해서 이론적인 방법 (예를 들자면 중심극한정리)를 이용하거나 방금 언급한 bootstrap을 사용할 수 있다. 사실 표본분포를 알아내기 위한 가장 쉬운 방법은 여러개의 표본이 있다면 각 표본에서 알고자 하는 특정 통계량의 값을 계산하여 그 값을 히스토그램을 이용하여 제시한다면 표본분포의 형태를 알 수 있다. 하지만 현실에서는 표본은 하나뿐이기 때문에 이 방법을 사용하는 것은 불가능하다. Bootstrap은 재표본 (resampling)이라는 아이디어를 사용하여 이 문제를 해결하였다. 즉 원래 표본에서 복원추출(sample with replacement)를 통하여 여러개의 재표본을 생성하고 각 재표본에서 관심을 가지고 있는 통계량을 계산한다는 것이 핵심 아이디어이다.

그림 1.3은 25개의 재표본에서 구한 lowess추정치를 보여주고 있다. 따라서 특정 연령대에서 25개의 예측치를 제시할 수 있으며 이를 바탕으로 예측치의 표준오차를 구할 수 있다. 실제 표준오차를 구하기 위해서 250개의

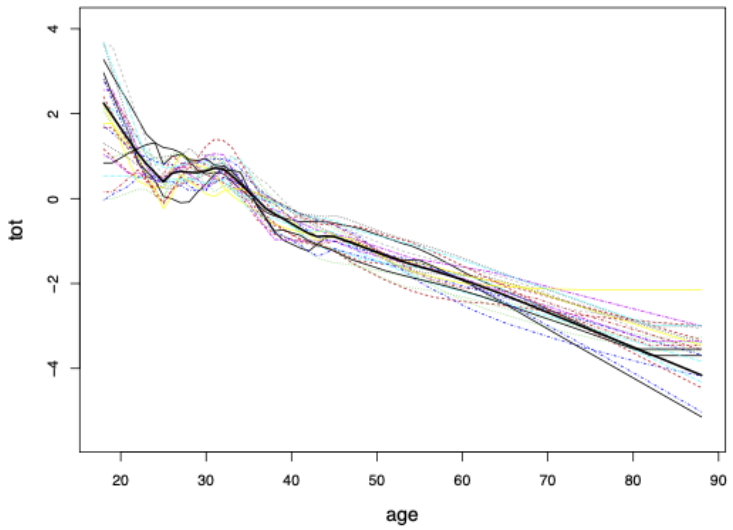

그림 1.3 Bootstrap을 바탕으로 만들어진 25개의 lowess(x, y, 1/3)

재표본을 사용하였다. 표 1.1에서 볼 수 있듯이 최소제곱법을 이용한 결과와 흡사하게 60대이상으로 갈수록 lowess추정치의 표준오차도 증가하는 경향을 보이는데 이 것은 연령이 증가하면서 상대적으로 데이터의 갯수가 줄어들기 때문이다.

표 1.1 스탠포드 신장자료의 회귀분석; (1) 나이별 선형회귀분석 예측치;(2) 선형회귀분석에서 각 나이별 예측치의 표준오차; (3) 나이별 lowess 예측치 (4) bootstrap을 이용한 lowess 예측치의 표준오차

age	20	30	40	50	60	70	80
1. linear regression	1.29	.50	-.28	-1.07	-1.86	-2.64	-3.43
2. std error	.21	.15	.15	.19	.26	.34	.42
3. lowess	1.66	.65	-.59	-1.27	-1.91	-2.68	-3.50
4. bootstrap std error	.71	.23	.31	.32	.37	.47	. .70

Lowess

Lowess (LOcally WEighted Scatterplot Smoothing)은 대표적인 비모수 함수추정방법의 하나로 이동평균 (moving average)을 일반화시킨 개념으로 생각할 수 있다. 먼저 $(x_1, y_1), \ldots, (x_n, y_n)$을 관측하였다고 가정하자, 이동평균을 계산하기 위해서는 이동평균을 계산하고자 하는 x값을 기준으로 가까운 몇개 또는 전체 자료중 특정 비율만큼의 가까운 관측치를 고려해서 이 관측치의 y값의 평균을 계산하면 된다. 실제 lowess에서는 이동평균을 보다 일반화한 다항회귀모형 (polynomial regression)을 사용하고 가중최소제곱 (weighted least square) 방법을 사용하여 최종 모형을 적합한다. 따라서 다항회귀의 차수 (degree)와 가중치, 사용하는 데이터의 비율에 따라 다양한 형태의 최종모형이 생성될 수 있다. 일반적으로 많이 사용되는 가중치는 $w(d) = (1 - |d|^3)^3$인 tri-cube weight function이며 차수는 2차이하는 권장한다. 따라서 실질적인 조절모수는 데이터의 비율 α라고 할 수 있으며 조절모수의 값에 따라서 나중에 배울 bias-variance trade-off가 발생한다.

1.2 Hypothesis Testing

통계적 추론은 크게 추정과 검정으로 나누어 지며 두번째 예제인 백혈병 환자자료는 가설검정에 관한 내용이다. 이 자료는 72명의 백혈병환자로 구성되어 있으며 이 중 45명 ALL (Acute Lymphoblastic Leukemia, 급성링프구성백혈병) 환자이며 나머지 27명은 AML (Acute Myeloid Leukemia, 급성골수성백혈병)환자이다. 일반적으로 급성골수성백혈병환자의 예후가 더 좋지 않다고 알려져 있다. 각 환자들에 대해서 7,128개 유전자 패널의 유전자 활성도를 측정하였다. 그림 1.4에서는 각 환자그룹의 136번 유전자의 유전자 활성도를 히스토그램을 이용하여 비교하였다.

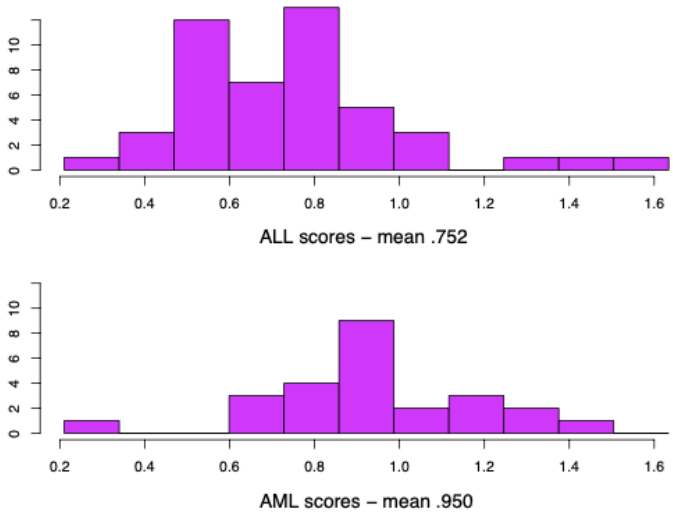

그림 1.4 백혈병 자료중 136번 유전자의 활성도: 위의 그림은 ALL ($n = 47$) 그룹, 아래그림은 AML ($n = 25$) 그룹을 보여준다. 이표본 t 검정통계량의 값은 3.01이며 p-value는 0.0036이다.

히스토그램에서 알수 있듯이 AML 그룹이 보다 활발한 활성도를 보여주고 있으며 각 그룹간의 활성도의 평균은 다음과 같다.

$$\overline{\text{AML}} = 0.752, \quad \overline{\text{ALL}} = 0.950$$

두 그룹의 평균은 어느정도 차이가 나는 것으로 보이지만 데이터의 산포정도를 고려해야 실질적으로 차이가 나는지 여부를 알 수 있다. 이 경우 가장 많이 사용되는 통계방법론은 이표본 t 검정이며 검정통계량은 다음과 같다.

$$t = \frac{\overline{\text{AML}} - \overline{\text{ALL}}}{\widehat{\text{sd}}}$$

여기서 $\widehat{\text{sd}}$는 $\overline{\text{AML}} - \overline{\text{ALL}}$의 표준편차(표준오차)의 추정치이다. 만약 두 집단이 모두 정규분포를 따른다면 귀무가설 (즉 두 집단의 평균이 같은 경우)하에서 위의 이표본 검정통계량은 자유도가 70인 t 분포를 따른다. 이표본 검정통계량의 값은 3.01로 주어지면 이 경우 양측검정을 사용한다면 p값은 0.036으로 주어진다. 일반적으로 유의수준이 0.05로 주어진다는 걸 고려한다면 이 경우 우리는 귀무가설을 기각할 수 있다.

하지만 이러한 결과가 사실 여러개의 가설검정을 동시에 진행해서 나온 결과중 하나라면 3.01이라는 검정통계량의 값은 그렇게 흥미있는 결과라고 하기는 힘들다. 그림 1.5는 모든 7,128개의 유전자 패널 각각에서 계산된

이표본 검정통계량의 히스토그램을 보여준다. 이 그림에서 3.01이라는 값은 생각보다 극단적인 값이 아니며 상위 5.6%에 위치하고 있다. 하지만 다음과 같은 2가지 요인때문에 여기서 0.056이 p값을 의미하지는 않는다.

1. 상대적으로 많은 가설검정의 갯수: 검정하고자 하는 가설이 많을 경우 모든경우에 귀무가설이 참이라 하더라도 극단적인 값이 나올 수 있다.

2. 이론적인 null distribution과 실제 데이터에서 보이는 null distribution의 괴리: 그림 1.5에서 실제 이론적인 null distribution인 자유도가 70인 t분포가 제시되어 있다. 이 경우 이론적인 분포보다 실제 데이터 분포의 꼬리가 훨씬 두터움을 알 수 있다.

위와 같은 요인은 다중검정 (multiple testing)의 문제점으로 알려져 있으며 이러한 문제를 해결하기 위해 15장에서 위발견율(False Discovery Rate: FDR)의 개념을 소개할 예정이다. FDR을 적용할 경우 백혈병 자료에서 흥미있게 보이는 검정통계량의 값은 6.16이상이 되어야 한다. 즉 136번째 유전자의 경우 실질적으로 그렇게 흥미있는 결과를 제시한다고 할 수 없다.

여기서 주목할 점은 이러한 새로운 형식의 데이터가 쏟아져 나올때 마다 새로운 분석 알고리즘이 통계학계 외부에서 종종 제시되곤 한다. Neural net이나 support vector machine, boosting이 이러한 예라고 할 수 있다. 하지만 통계학자들은 이러한 알고리즘은 어떻게 작동하는지 여부를 기존의 frequentist 혹은 Bayesian의 관점에서 설명하고 이론적 토대를 제공한다. Boosting이 대표적인 사례로 뽑을 수 있다.

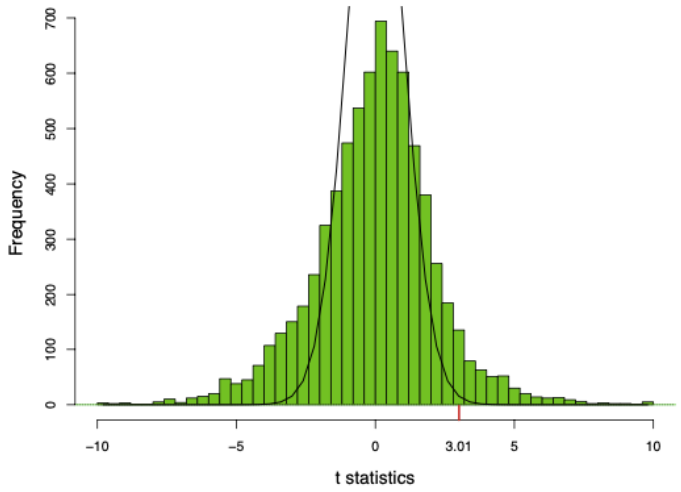

그림 1.5 7,128개의 유전자 활성도를 이용하여 계산한 이표본 검정통계량의 히스토그램. 가운데 분포곡선은 자유도가 70인 t분포를 나타낸다.