1

추정과 검정

1. 개념 concept

2016년 H대 입학처는 우수 신입생 선발을 위하여 다음을 조사하였다. (1) 1학년 GPA (2) 고등학교 내신 등급 (3) 전체 수능점수 (4) 수학 수능점수 (5) 특목고 여부 (6) 출신고 지역

모집단과 표본

- 모집단 population : 관심, 연구 대상인 전체 개체 (사람, 사물) - 2015년 신입생 전체, 전수조사 census
- 표본 sample : 표본설계를 통하여 모집단 중 조사 대상으로 추출된 개체 - 2015년 신입생 중 확률 추출된 (n=200 신입생, 표본크기)

데이터 data

연구자가 조사목적을 위하여 수집한 숫자모임, 열은 확률변수와 행은 조사대상 개체의 관측값

(확률)변수 random var. 관측값 observation

모집단 개체에 관심을 갖는 특성, 확률의 의미는 값을 예측할 수 없다는 것을 의미하며, 변수는 관측값이 개체마다 변하므로, 개체의 변수의 조사 값을 관측값 observation - 첫 조사 대상 신입생 GPA=3.67, 고교 내신=3.2등급, 수능점수=425점, ...

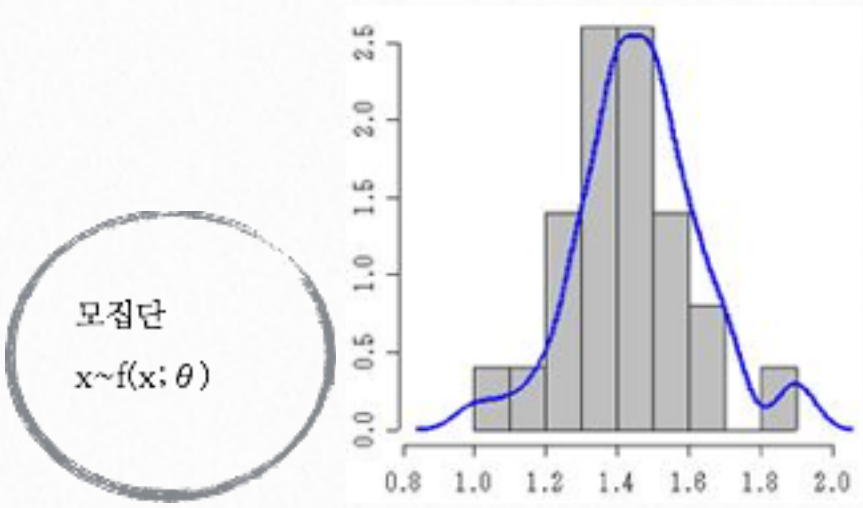
(1) 이산형 disdrete vs. 연속형 continuous

(2) 양적 quantitative vs. 질적 qualitative

(수학적 정의) 확률실험 random experiment (실험의 결과 값을 예측할 수 없어 확률이라 함)의 결과(원소 w , 입력)를 실수 값(x, 출력)으로 변환하는 함수 $X(w) = x$

확률분포함수 probability density fn.

- 확률변수 값(x , 입력), 그에 대응하는 확률($P(X = x)$, 출력)임 함수를 표, 그래프, 수식
- 확률변수의 값에 대한 정보 - 최소, 최대, 중앙 위치, 산포 정도, 치우침의 정도를 알 수 있음
- 관심 값, 범위에 대한 확률을 계산할 수 있음
- 일반적으로 전수조사를 하지 않는 경우 모집단의 확률분포함수를 알 수 없음
- 데이터의 히스토그램 (이산형 - PDF, 연속형은 막대 중앙점을 연결한 polygon 그래프가 PDF, 수식으로 정확하게 표현될 수 없음, 하여, 모집단 분포는 가정, 혹은 모른다고 설정함)



모수 parameter 통계량 statistics

(1) 모수 : θ

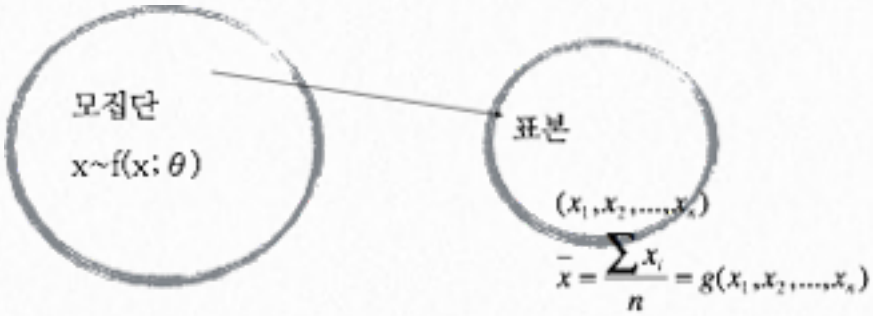
- 모집단 개체 관심 특성 값
- 모집단 확률분포함수의 요약 값
- 통계 모형의 계수

(예) GPA 평균($\theta = \mu$), 특목고 출신 비율($\theta = p$),
$GPA = a + b * Math\_score + e$

(2) 통계량 : $\hat{\theta}$

- 모수를 알기 위하여(추정) 표본으로부터 계산된 값
- 표본 관측값의 함수 : $\hat{\theta} = f(x_1, x_2, \ldots, x_n)$
- 추정에 사용되는 통계량은 추정값, 가설검정에 사용되는 통계량은 검정통계량이라 함

(예) 표본 200명의 GPA 표본평균, 200명 표본데이터로부터 계산된 OLS 추정값 $(\hat{a}, \hat{b})$



확률표본 random sample

모집단의 각 개체가 표본으로 뽑힐 가능성이 동등한 상황에서 표본이 추출할 때 이를 확률표본이라 함,

확률표본의 데이터(관측값) 분포는 모집단의 분포와 동일하다. $x_i \sim f(x; \theta)$

(수학적으로) <=> independently and identically distributed - 확률표본 결합밀도함수

$f(x_1, x_2, \ldots, x_n; \theta) =$ (독립)
$f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta) =$(동일) $[f(x; \theta)]^n$

표본 분포 sample distribution

$f(x_i; \theta)$는 모집단의 분포함수 $f(x; \theta)$와 동일하다. 모집단의 분포 함수는 알수 없다.

모집단에 대하여 알 수 없는 것은?

1) 모수 θ : 추론 대상

2) 모집단 분포 $f(x; \theta)$: 관심 대상이 아니다. 모집단의 분포가 관심대상이 경우 - 분포의 적합성 검정이며 가장 널리 사용되는 것이 정규성 검정

샘플링분포 sampling disttribution

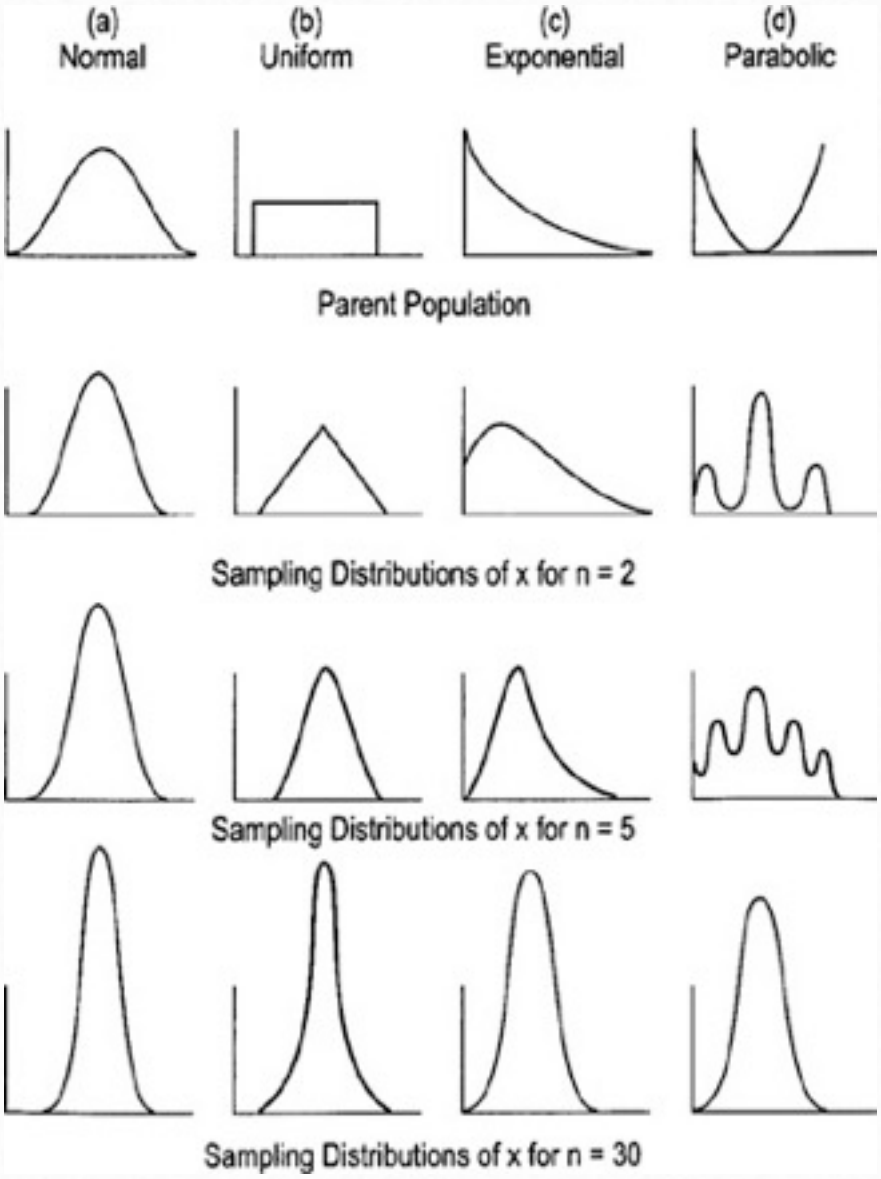
- 통계량의 확률분포함수의미한다.

추론 inference

- 통계량과 통계량의 샘플링 분포를 이용하여 모수에 대한 정보를 추론하는 과정
- 추론은 추정과 통계적 가설 검정
- 점추정은 하나의 값으로 모수를 추정하는 것이고 구간추정은 구간 값으로 추정
- 통계적 가설은 모수에 대한 가설의 진위를 데이터를 이용하여 검증
- 구간추정과 가설검정에는 통계량의 샘플링분포를 알아야 함

|중심극한정리 central limit theorem| 모집단의 분포에 상관없이 표본크기가 충분히 크면(n>20~30) 표본 데이터의 합, 표본평균의 샘플링분포는 정규분포에 근사(approximate)한다. (수학적 표현) Even though $f(x; \theta) \sim ?(\mu, \sigma^2)$

$\sum x_i$ or $\bar{x} \sim (appr.) N(\mu, \frac{\sigma^2}{n})$이다.



2. 점 추정 point estimation

모집단 : 주사위 눈금의 기대값 : $\theta = \mu = 3.5$ ->

$f(x) = \frac{1}{6}, x = 1, 2, \ldots, 6$

모수의 추정 값을 하나의 (단일) 값으로 추정

1) 추정량 estimator

모수를 추정하는 계산식

2) 추정값 estimate

표본 데이터를 이용하여 추정량 계산식에 의해 구한 모수 추정 값

3) Best estimate 베스트 추정값

MSE Minmum Sqaured Error (최소제곱오차)

$E(\hat{\theta} - \theta)^2$를 최소화 하는 추정값

4) 최소분산불편추정량 MVUE Minimum Variance among Unbiased Estimator

- MSE를 최소화 하는 베스트 추정량 계산은 불가능
- 하여, $E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}))^2 - E(E(\hat{\theta}) - \theta)^2$
  $= V(\hat{\theta}) + B^2(\hat{\theta})$
- 불편추정량 : $E(\hat{\theta}) = \theta$인 추정량, 즉 편이(bias)가 0인 추정량 = $B(\hat{\theta}) = E(\hat{\theta}) - \theta$
- 불편추정량 중 추정분산($V(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$)을 최소화 하는 추정량을 MVUE
- 불편추정량인 경우 MVUE=MSE 최소 베스트

5) MVUE 구하기

(1) MLE 추정량 구하기

- 우도함수(표본데이터 결합밀도함수이며 모수의 함수) $\min_{\hat{\theta}} L(\underline{x}; \theta) = \min_{\hat{\theta}} f(x_1, x_2, \ldots, x_n; \theta)$ 최소화 하는 추정량을 최대우도 추정량이라 함
- MLE는 CSS함수이다.

(2) MLE의 함수이며 불편 추정량인 경우 이를 MVUE라 함(Rao-Balxkwell Thm)

(3) 통계추론에서 사용하는 모든 추정량은 MVUE이며, MLE추정방법에 의해 구해진다.

모평균 MVUE= 표본평균 $\bar{x}$, 점 추정치를 구할 때는 추정량의 샘플링분포를 알 필요는 없음, 그러나 모집단의 분포는 알아야 함

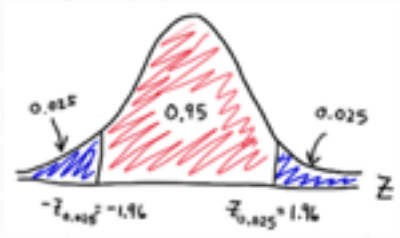
3. 구간 추정 interval estimator

점 추정량과 추정량의 샘플링분포를 이용하여 모수의 값을 구간으로 추정함

95% 신뢰구간 : $P(L(\hat{\theta}) < \theta < U(\hat{\theta})) = 0.95$, 구간 하한 추정량 : $L(\hat{\theta})$, 구간 상한 추정량 : $U(\hat{\theta})$

표본평균 $\bar{x}$의 샘플링분포는 중심극한정리에 의해 정규분포를 따르고 표본평균의 평균은 μ이고 분산은 $\frac{\sigma^2}{n}$이다. $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ => 100(1- α)% 신뢰구간

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



(95% 신뢰구간)

4.가설 검정 Hypothesis testing

1) 통계적 가설 statistical hypothesis

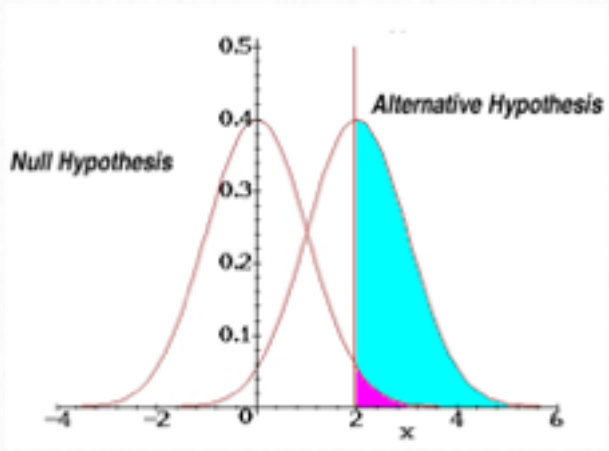
- 통계학이 적용되는 연구문제는 통계적 가설로 정의됨 - 데이터를 기반하여 통계적 가설의 진위를 검증하게 됨
- 통계적 가설은 모수의 값으로 표현됨
- 통계적 가설은 귀무가설과 대립가설로 나눔
- 귀무가설 null hypothesis : 모수 하나의 값으로 설정 $H_0: \theta = \theta_0$, 차이가 없다 $H_0: \theta_1 - \theta_2 = 0$, 영향을 미치지 않는다, 효과가 없다 $H_0: \hat{b} = 0$, nothing
- 대립가설 alternative : 귀무가설에 설정된 이외 모수 값 모두, 연구가설이라고 함, 효과가 있음,
- 대립가설은 단측가설($H_0: \theta > \theta_0$)과 양측가설($H_0: \theta \neq \theta_0$)로 나뉨

2) 검증 오류 test errors

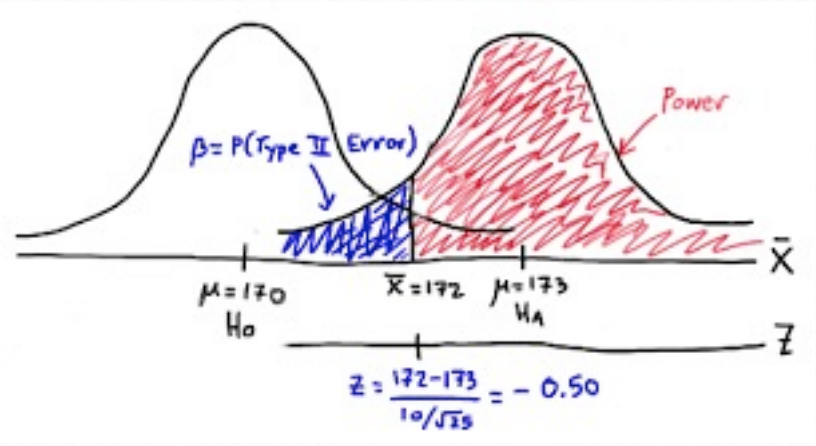
- 1종 오류 type I error = P(귀무가설 기각 | 귀무가설 참)
- 2종 오류 type II error = P(귀무가설 채택 | 대립가설 참)

실제 모집단 / 판단	귀무가설 진실	대립가설 진실
귀무가설 기각	1종 오류(α) = 유의수준	옳은 판단 : 검정력
귀무가설 채택	옳은 판단	2종 오류(β)

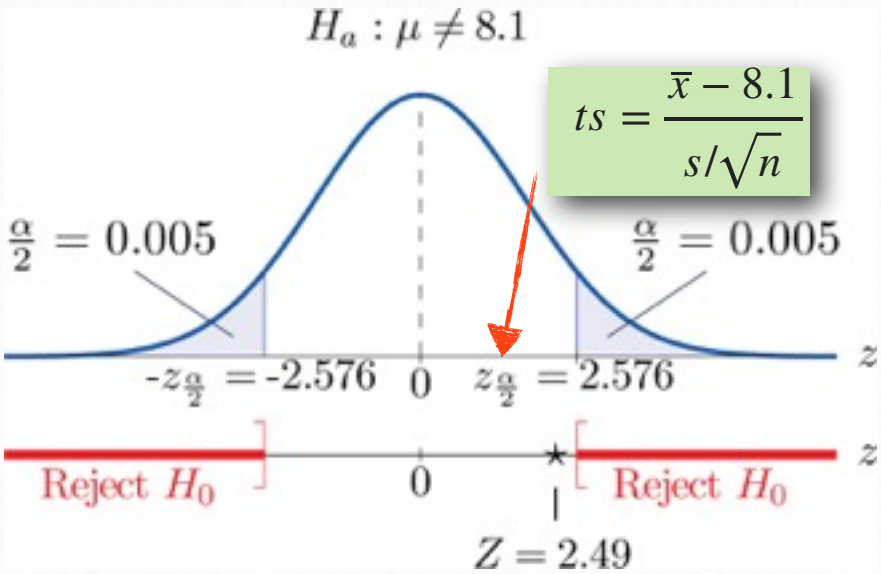
- 검정력 test power = P(귀무가설 기각 | 대립가설 참) = 1-2종오류



- 가설검정 결과 2가지 종류의 오류가 발생 - 동시에 두 오류를 줄이는 가설검정법 없음
- 두 오류 중 하나를 고정하자 - 우리 관심이 대립가설에 있으므로 1종 오류를 고정하고 (대립가설을 채택하고 싶어 이 정도 오류는 허용하자) 검정력을 최대화 하는 검정방법을 찾음'
- 고정된 1종 오류를 유의수준 significant level - 허용된 오류



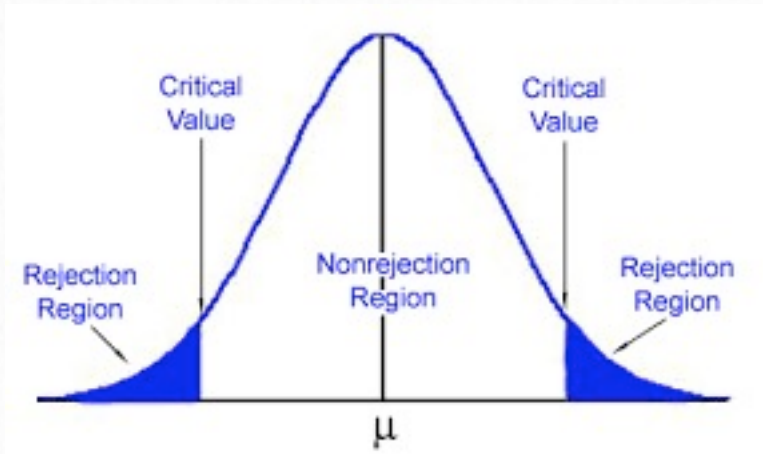
3)검정통계량 (test statistics)



- 통계적 가설의 진위 여부를 판단하기 위하여 표본 데이터로부터 계산된 통계량
- 통계량은 귀무가설에 설정된 모수의 MVUE 추정치($\bar{x}$)가 기반되고 피봇통계량($\frac{\bar{x} - \mu}{s/\sqrt{n}}$)의 샘플링 분포가 이용됨
- 검정통계량 값은 귀무가설에 설정된 모수 값을 활용하여 계산한다.
- 귀무가설 하에서 검정통계량 값이 계산되므로 확률분포의 끝 부분 값도 가질 수 있음 - 그러나 연구가설인 대립가설을 원하므로 어느 정도 오류는 감수하자.
- 이것이 유의수준임, 분포의 끝 부분(이곳을 기각역이라 함)에 계산된 검정통계량 값이 놓이면 귀무가설이 옳을 수 있는데도 불구하고 귀무가설을 기각

4)유의수준 (significant level)

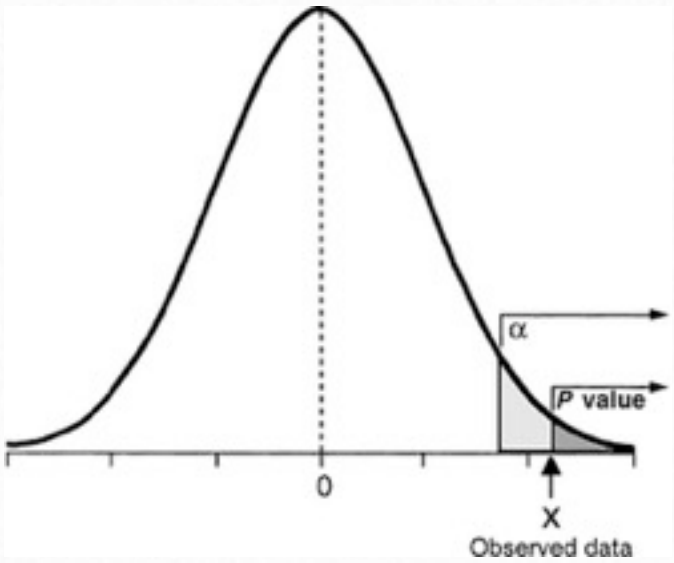
- 귀무가설이 옳음에도 불구하고 귀무가설을 기각할 확률 - 설정된 1종 오류
- 일반적으로 5%, 1%, 10% 주로 사용
- 유의수준과 신뢰수준은 역관계 - 95% 신뢰수준 <=> 5% 유의수준
- 귀무가설 설정된 모수(θ)을 활용하여 샘플링 분포를 구함



5)기각역(critical region)

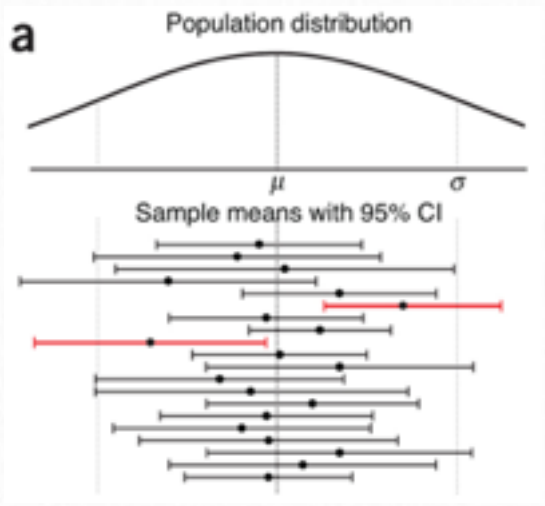
- 귀무가설 하에서의 샘플링분포의 끝 영역으로 검정통계량 값이 이 구간에 속하면 귀무가설을 기각
- 유의수준 크기에 의해 범위가 결정됨
- 기각역이 시작되는 값을 기각값 value
- 대립가설이 양측가설이면 유의수준을 1/2씩 양쪽으로 배정하고 단측가설은 유의수준 전체를 한 쪽에 배정하여 기각역을 구함

6) 유의확률 p-value probability value



- 귀무가설을 기각할 최소의 유의수준 (계산된 1종 오류)
- 계산된 유의수준, 데이터 기반 유의수준 (파란색 빗금, 양측검정이면 2배)
- 유의확률≤유의수준 ⇔ 귀무가설 기각, 유의확률> 유의수준 ⇔ 귀무가설 채택

7) 95% 신뢰구간



5. 통계적 추론 절차

1)연구문제 및 통계적 문제(가설: 모수) 정의

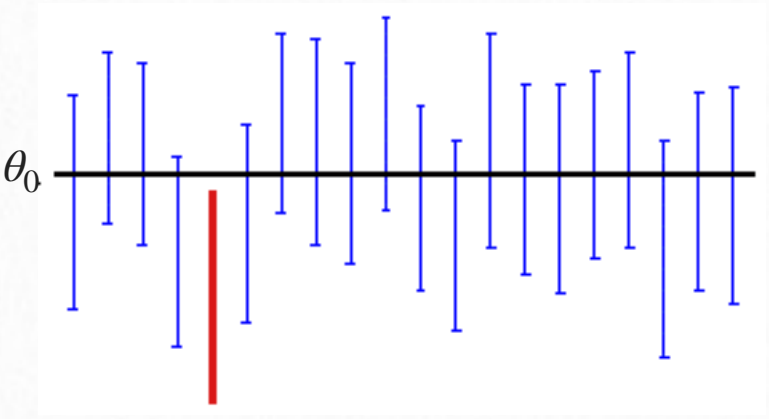
- H대 2015 신입생들의 내신등급은 3.5등급 이상일 것이다. => 귀무가설 : $\mu = 3.5$, 대립가설 : $\mu > 3.5$, 모수 μ는 H대 학생 내신평균
- 수능점수가 높으면 대학 학업능력이 높을 것이다. 학업능력은 GPA로 => 귀무가설 : $b = 0$, 대립가설 : $b > 0$

2)데이터 검증

- 모평균 추론 : 이상치, 치우침
- 선형모형 : 정규성, 등분산성, 독립성 등

3)가설검정 및 신뢰구간 계산

- 검정통계량, 유의확률 계산
- 신뢰구간 : 귀무가설 설정 모수 값 포함하면 귀무가설 채택, 포함하지 않으면 귀무가설 기각



4) 결론 및 활용

- 적절한 표 작성, 결과에 대한 활용 측면의 해석