Statistical-Inference 스터디 소개 & 2장 빈도주의 추론

Sang Yeol Lee

DataBreak
Statistical-Inference

July 03, 2019

스터디는?

- 데이터뽀개기 커뮤니티(https://www.facebook.com/groups/databreak/) 스터디입니다
- 매주 수요일 진행 (7 ~ 8월)
- 장소는 판교에서 진행 (다만 이후 변경되면 재공지 할 예정)

github 주소

- github 레퍼지토리는 clone 하셔서 발표 자료는 github 통해서 올려주세요!

스터디원 명부 : github 주소 적어주세요, 구글드라이브

교재?

- http://www.yes24.com/Product/Goods/71829251 (번역본)
- https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf (영문판)
- 영문판으로 보셔도 되고, 번역본으로 보셔도 됩니다. 편하신 대로

스터디 진행 방식

- 발표 주제별로 ~최소 30분 준비 (+ 시간이 남는다면 추가로 본인이 하는 일과 관련있게 준비해도 되고, 시간을 못 채우시면 다른 데이터 분석 주제를 넣으셔도 됩니다.)
- 파트1 스터디 하면서 최소 1번 발제를 진행 (**2번 권장**)
- 스터디 빠지면 벌금을 걸을까?

스케줄 계획표 (Part1 커리큘럼)

요일	주제	발표자	발표자료	비고
2019.07.03(수)	1. 알고리즘과 추론	이희재		
	2.빈도주의 추론	이상열		
2019.07.10(수)	3.베이즈 추론			
	4.피셔 추론과 최대우도 추정			
2019.07.17(수)	5.모수적 모델과 지수 패밀리			
	6.경험적 베이즈			
2019.07.24(수)	7.제임스-스타인 추정과 리지 회귀			
	8.일반화된 선형 모델과 회귀 트리			
2019.07.31(수)	9.생존 분석과 EM 알고리즘			
	10.잭나이프와 부트스트랩			
2019.08.07(수)	11.부트스트랩 신뢰구간			
	12.교차 검증과 Cp 예측 오차 추정			
2019.08.14(수)	13.객관적 베이즈 추론과 마르코프 체인 몬테 카를로(1)			
	13.객관적 베이즈 추론과 마르코프 체인 몬테 카를로(2) + 14장 마무리			
2019.08.21(수)	15.대규모 가설 검정과 거짓 발견율			
	16.희소 모델링과 라소			
2019.08.28(수)	17.랜덤 포레스트와 부스팅(1)			
	17.랜덤 포레스트와 부스팅(2)			

자기소개 (TMI)

- 자신을 편하게 소개해주세요!
- 이름 / 하는 일 / 스터디 하는 이유 / TMI (관심있는 취미, 활동, 스포츠, 아무거나 말하고 싶은 것)

스터디 1번째 시간

- 1장 알고리즘과 추론 (이희재)
- 2장 빈도주의 추론 (이상열)

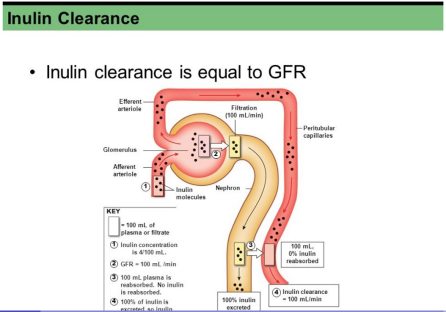
발표

- 2장 빈도주의 추론

2-1. GFR

- 신장환자 예제 (211명의 신장 기능검사 결과)
- 스탠포드 대학의 브라이언 마이어스 (Bryan Myers) 박사 신장학 연구소의 211 명의 신장 환자의 사구체 여과율 측정
- gfr(사구체여과율 : glomerular filtration rate)
- 신장기능의 저하를 일으키는 질환에서 수치가 낮고, 경중을 판단하는 지표로 사용
- 수치가 90 이상일 경우 신장 기능의 90%을 유지됨을 의미

gfr, kidney function



2-2. GFR (R code)

```r
gfr_url <- "https://web.stanford.edu/~hastie/CASI_files/DATA/gfr.txt"
gfr <- read_csv(gfr_url, col_names = F)
print(paste0("mean? ", mean(gfr$X1)))
```

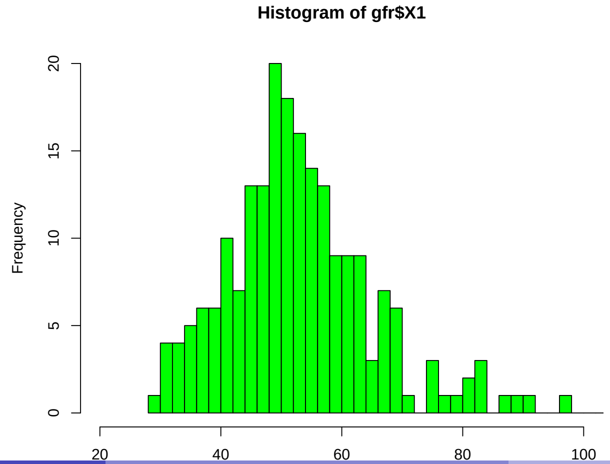
```
## [1] "mean? 54.2654028436019"
```

```r
print(paste0("sd? ", sd(gfr$X1)))
```

```
## [1] "sd? 13.7208804105898"
```

2-3. GFR (R code)

```
hist(gfr$X1, col = "green",breaks = 31, xlim = c(20, 100))
```



2-4. 모수와 통계량

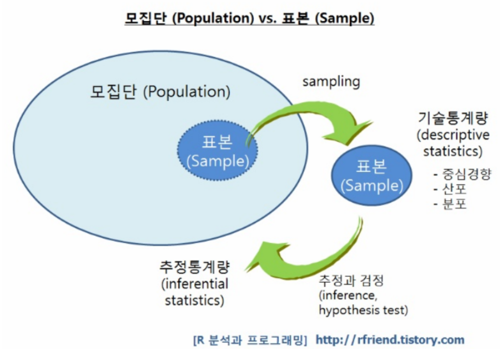


Figure 2: popluation

관심정보	모수(parameter)	통계량(statistic)
분포의 중심위치 (central location)	모평균 (μ)	표본평균 ($\bar{x}$)
분포의 산포 또는 흩어짐 (dispersion)	모분산 (σ^2) 모표준편차 (σ)	표본분산 (s^2) 표본표준편차 (s)
비율 (proportion)	모비율 (p)	표본비율 ($\bar{p}$)

Figure 3: parameter

2-4. 미지의 분포 F (1)

참고 블로그

- 관측 추정치 theat hat(단일 숫자)의 정확도란 theta의 추정치로서의 확률적 정확도
- 빈도주의는 대게 **미래의 무한한 연속적 시도**에 대해 정의

모수 θ에 대한 점추정량(point estimator)을 $\hat{\Theta}$ 라 하면,

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

점추정량은 표본들의 함수입니다. 만약 모수가 평균이라면, 점추정량은 다음과 같이 됩니다.

만약 모수가 평균이라면, 즉 $\theta = E(X)$ 이면

추정량 $\hat{\Theta} = \bar{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ 이다.

따라서 모평균의 추정량은 표본평균이다.

마찬가지 방법으로 표분분산은 모분산의 점추정량이 됩니다.

대표적인 점추정량은 표본평균과 표본분산이다.

- 가상의 데이터 집합 X(1), X(2), X(3), ··· (빈도주의 원칙은 theat hat은 대문자 세터 값들의 앙상블이 가지는 성질에 대한 정확도, 동일한 상황을 다시 실행하면 어떤 현상이 나타날까??)

1. 편향

우선 '편향'에 대해 알아보죠. 표본들로부터 얻어낸 추정량은 모수에 가까울수록 좋습니다. 생각해보면 당연하죠? 즉 **추정량의 기댓값이 모수와 같아지는 것이 가장 바람직한 경우**겠죠. 우리가 기대하는 추정량과 모수의 차이를 **'편향(bias)'**이라고 합니다.

임의의 추정량의 편향을 $B(\hat{\Theta})$ 라 하면,

$$B(\hat{\Theta}) = E(\hat{\Theta}) - \theta$$

추정량의 기댓값과 모수가 같아지는 경우가 가장 바람직한 경우인데, 이 경우는 편향이 0인 상황이겠죠. 따라서 편향이 0일 때의 추정량을 '불편추정량(unbiased estimator)'이라고 합니다.

편향이 0인 경우, 즉 $B(\hat{\Theta}) = 0$일 때의 추정량을 ´불편추정량´이라고 한다.

표본평균은 모평균의 불편추정량이고, **표본분산은 모분산의 불편추정량**입니다. 직접 한 번 유도해보시기 바랍니다.

주의할 점은 위에서 적률방법으로 유도한 분산의 점추정량은 **편향추정량(biased estimator)**입니다. 편향추정량은 **표본을 통해 얻은 추정량과 모수가 일치하지 않음을 의미**합니다.

2-4. 미지의 분포 F (2)

참고 블로그

2. 평균제곱오차

이번에는 '평균제곱오차'에 대해 알아보겠습니다. 일반적으로 오차는 근사값과 참값의 차이를 나타냅니다. 여기서 근사값을 점추정량, 참값을 모수라고 하면 오차는 다음과 같습니다.

점추정량을 $\hat{\theta}$, 모수를 θ 라 하면, 오차는 다음과 같습니다.

$$\text{Error} = \hat{\theta} - \theta$$

평균제곱오차(mean squared error)는 오차를 제곱한 값의 기댓값으로 정의합니다. 오차는 작을수록 참값에 가깝다는 의미죠. 따라서 **평균제곱오차도 값이 작을수록 좋습니다**.

점추정량의 평균제곱오차(MSE)는 다음과 같습니다.

$$MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Figure 6: P1

평균제곱오차는 다른 방법으로 표현할 수 있습니다. 바로 분산의 개념을 이용하는 것이죠.

$$Var\left(\hat{\theta} - \theta\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right] - \left\{E\left[\hat{\theta} - \theta\right]\right\}^2$$

$$\Rightarrow E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var\left(\hat{\theta} - \theta\right) + \left\{E\left[\hat{\theta} - \theta\right]\right\}^2$$

$MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$ 이고, $B\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$ 이므로

$$\Rightarrow MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left\{B\left(\hat{\theta}\right)\right\}^2$$

Figure 7: P2

3-1. 실제에서의 빈도주의

- 빈도주의에 대한 실용적 주의

  관심 대상 절차의 확률적 속성을 도출한 다음, 그대로 절차의 출력에 적용해 관측 데이터를 구성하는 것. 헛점은 참 분포 F를 알 수 없어도 참 분포 F로부터 추정량의 성질을 계산해야 함

- 우회하기 위해 여러가지 테크닉 사용

1. 플러그인 원리
2. 테일러 급수 근사
3. 모수적 패밀리와 최대 우도 이론
4. 시뮬레이션과 부트스트랩
5. 피봇 통계량

3-2. 플러그인 원리 (1)

표본표준편차와 표준오차의 구분

1. 표본 평균 : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
2. 표본 표준편차 : $s_x = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
3. 표준 오차 : $\text{s.e.}(\hat{x}) = \dfrac{s_x}{\sqrt{n}}$

3-2. 플러그인 원리 (2)

- 표준오차는 표본평균의 표준편차다.
- 표준오차는 모집단에 대해서는 딱히 어떤 정보도 주지 않는다.
- 표준오차는 아무 말이 없어도 표본에 대해서만 논한다. 즉, '모표준오차'와 같은 개념은 생각하지 않는다.
- 표준오차는 주로 신뢰구간이나 예측구간을 구할 때 필요한 것이다. 즉, 맥락 상 구간 이야기가 있을 때만 신경쓰면 된다.
- 관측치가 있다면 표본 표준편차를 표준 오차에 대입하면 표준오차(hat)이 되고, 표본 평균의 표준 오차에 대한 일반적 추정이 됨
- 표본 평균에 대한 빈도주의 측면의 정확도 추정은 관측 데이터 그 자체로부터 추정

3-3. 테일러 급수 근사 (1)

- 테일러 급수란? 좋은 동영상 테일러급수 알아보기

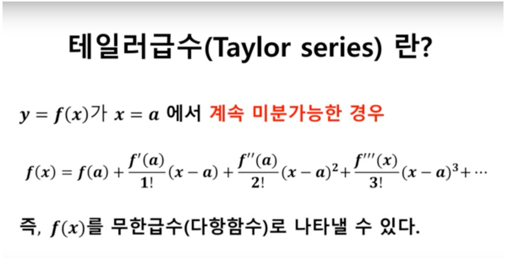


Figure 8

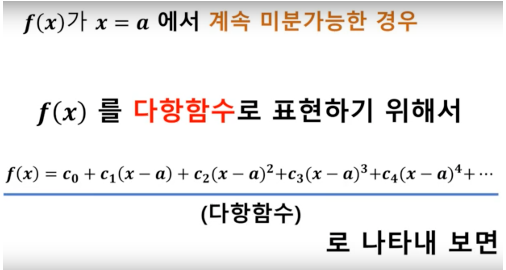


Figure 9

3-3. 테일러 급수 근사 (2)

- 테일러 급수란? 좋은 동영상 테일러급수 알아보기

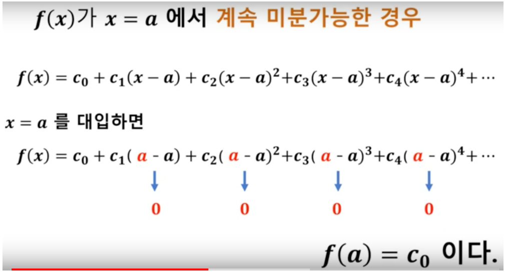


Figure 10

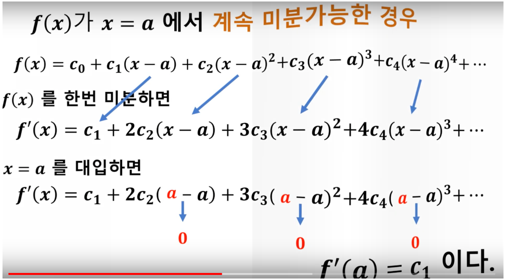


Figure 11

- 테일러 급수란? 좋은 동영상 테일러급수 알아보기

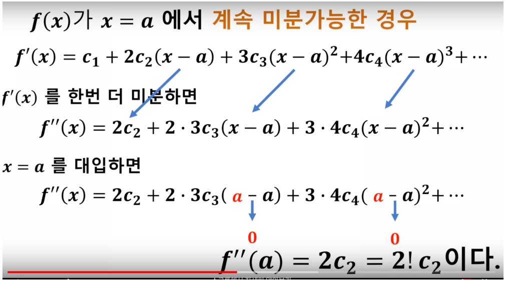


Figure 12

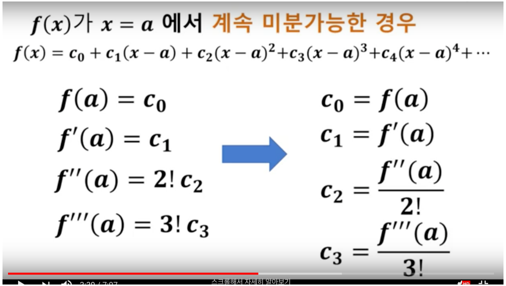


Figure 13

3-3. 테일러 급수 근사 (4)

- 테일러 급수란? 좋은 동영상 테일러급수 알아보기

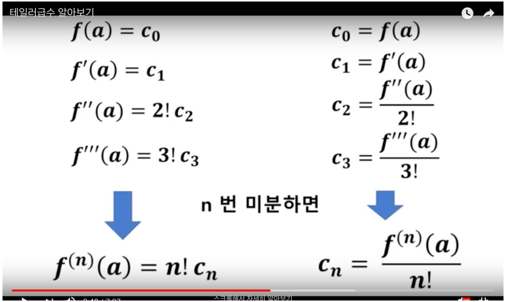


Figure 14

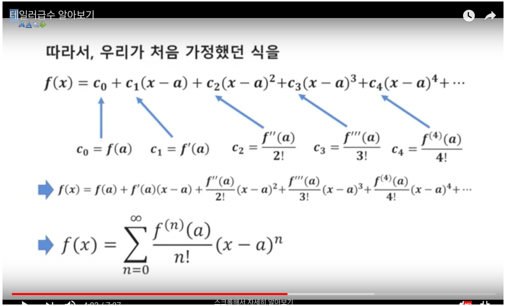


Figure 15

3-3. 테일러 급수 근사 (5)

- 테일러 급수란? 좋은 동영상 테일러급수 알아보기

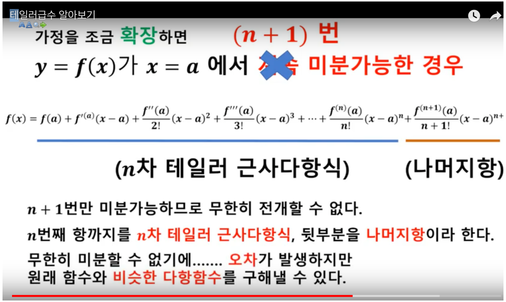


Figure 16

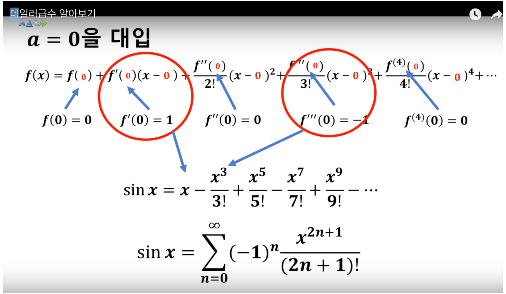


Figure 17

3-3. 테일러 급수 근사 (6)

- 테일러 급수 근사란?
- 델타 기법 : 1차 테일러 급수를 사용해 통계량 theat hat의 분산을 근사하는 방법

확률변수 함수의 분포를 알아보자 - Delta method에 대하여 (1)

예제를 통한 이해

예제 1

확률변수 $X_i, i = 1, \ldots, n$이 독립이고 같은 분포를 따른다고 가정하면, 우리는 중심극한정리를 사용하여 다음과 같은 점근 분포를 얻을 수 있다.

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

위에서 $\mu \neq 0$와 σ^2는 확률변수 X의 평균과 분산을 나타내는 모수이다. 이 경우 앞에서 설명한 delta method의 설명에서 나온 r_n는 $\sqrt{n}$에 대응하고, θ의 경우는 μ에, 수렴분포는 평균이 0이고, 분산이 σ^2인 정규분포가 된다.

이렇게 주어진 점근 분포를 바탕으로 우리는 다음의 점근 분포를 delta method를 이용하여 손쉽게 구할 수 있다.

$$\sqrt{n}\left(\frac{1}{\overline{X}} - \frac{1}{\mu}\right) \xrightarrow{d} ?$$

함수 g를 $g(x) = 1/x$라고 정의하면, 우리가 점근 분포를 알고자 하는 확률변수 $1/\overline{X}$는 $g(\overline{X})$라고 볼 수 있으므로, 앞에서의 delta method 정리에 의하여 위의 점근 분포는 다음의 분포로 수렴하게 된다.

$$\sqrt{n}\left(\frac{1}{\overline{X}} - \frac{1}{\mu}\right) \xrightarrow{d} -\frac{1}{\mu^2}Y$$

$$\sqrt{n}\left(\frac{1}{\overline{X}} - \frac{1}{\mu}\right) \xrightarrow{d} -\frac{1}{\mu^2}Y \sim \mathcal{N}\left(0, \frac{1}{\mu^4}\sigma^2\right)$$

3-4. 다른 방법들

- [3] 모수적 패밀리와 최대우도 이론 : MLE의 표준오차에 대한 이론적 표현은 4,5장에 설명되어 있음 (피셔 이론, 테일러 급수 근사, 플러그인 원리 사용)
- [4] 시뮬레이션과 부트스트랩 (10장에서 자세히 다룰 예정) : F의 추정 F hat을 시뮬레이션 하여 경험적 표준편차 등을 구하는 것
- [5] 피봇 통계량 : 확률분포 F에 종속되지 않는 것. (스튜던트 2표본 t검정)

3-5. 피봇 통계량

- 스튜던트 T 분포 통해서 확률분포 F 대신 귀무가설을 검정
- 스튜던트 2표본 T검정 수행하여 문제를 푸는데… 2표본 문제에서는 숫자의 집합 2개를 관찰
- x1 = (x11,x12,…,x1n1), x2 = (x21,x22,…,x2n2)
- 다만 스튜던트 2표본 T검정을 하기 위한 4가지 조건

1. 자료는 모두 동일 간격을 가진 연속형 수치
2. 두 집단은 서로 독립적
3. 자료의 수치는 정규성을 가져야 함
4. 두 집단 각각에서 추정된 분산은 동일해야 함

- 둘이 같은 정규분포이기 때문에, 평균이 같다라는 귀무가설을 검정
- 평균끼리의 차이는 Normal (0, sigma⊠2(1/n1+1/n2))의 분포를 따르게 됨
- 스튜던트 분포를 사용해서 시그마의 값과 상관없이 동일한 분포(n1+n2-2차의 자유도를 가지는 스튜던드 T분포) 통해서 피봇 통계량을 계산
- 다만 피봇성은 대부분 통계 상황에서 존재하지 않음…

3-6. Winsorized mean (1)

- 윈저화 평균은 이상값(outlier)이 있을 경우에 사용하는 통계량
- 절사 평균(trimmed mean)과 비슷함
- 1, 2, 3, 4, 5, 6, ⋯., 100 데이터가 있을 때
- 절사 평균은 가장 작은 5개와 가장 큰 5개를 제외한 값 {6, 7, 8, ⋯, 94, 95}를 사용
- 윈저화 평균은 가장 작은 5개 대신 그것보다 바로 위의 값인 6으로 대체하고, 가장 큰 5개 대신 그것보다 바래 아래의 값인 95로 대체
- {6, 6, 6, 6, 6, 6, 7, 8, ⋯., 94, 95, 95, 95, 95, 95, 95}

3-6. Winsorized mean (2)

```
mean(c(1:10, 300))
```

```
## [1] 32.27273
```

```
mean(c(1:10, 300), trim=.05)
```

```
## [1] 32.27273
```

```
mean(c(1:10, 300), trim=.1)
```

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

```
winsorMean <- function(x, probs=c(0.05, 0.95)) {
 xq<-quantile(x,probs=probs)
 x[x < xq[1]] <-xq[1]
 x[x > xq[2]] <-xq[2]
 return(mean(x))
}

winsorMean(c(1:10, 300))
```

```
## [1] 19.13636
```

3-7. 빈도주의 최적성

- 빈도적 기법이 대중화된 것은? 상대적으로 단순한 수학적 모델링 가정
- 오직 확률 모델(F)와 알고리즘 t(x)만 가정하면 됨
- gfr 분포의 위치를 추정하기 위해 **표본 평균**을 쓸것인가? **윈저화 평균**을 쓸것인가?
- 주어진 모델 F에 대해 최적의 t(x)를 고르는 방법이 발견됨

1. 피셔의 최대 우도 추정 이론 (4장)
2. 피셔 정보 범위 (4장)

- 같은 맥락으로 네이만-피어슨의 보조정리(lemma)는 최적 가설 감정 알고리즘 제시

3-8. 네이반-피어슨의 보조정리(lemma)

- 귀무가설 H0에 대한 가능도 함수의 값과 대립가설 H1의 가능도 함수 값을 가지고 비(ratio)를 구해보면 가능도비(likelihood ratio)라는 것이 구해지는데, 이 값이 적절한 상수 k보다 작은 경우 해당 기각역 C는 최량기각역(best critical region of size σ)
- 최량기각역은 가설검정에서 표본크기가 정해져 있다면 1/2종 오류를 동시에 줄일 수 있는 방법이 없음, 정해진 유의수준에서 검정력을 최대로 하는 기각역 설정 필요 (최적의 기각역 : 최량 기각역)
- 1종 오류? 귀무가설이 참인에도 불구하고, 귀무가설을 기각할 오류 (실제로 효과가 없는데 효과가 있다고 나타내는 것)(alpha)
- 2종 오류? 대립가설이 참인데도 불구하고 대립가설을 기각할 오류 (실제로 효과가 있는데 효과가 없다고 나타내는 것) (beta)
- 1 - beta가 검정력, a가 유의수준

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

3-9. 주석 및 상세설명

- “빈도주의”라는 이름은 리차드 폰 미제스의 **확률의 빈도주의 이론**에서 유래됨
- 네이만은 통계적 유사성으로부터 유추해 제안한 것으로 보임 (1977년 논문 “빈도주의 확률과 빈도주의 통계량”)
- 통계량 t(x)의 장기적인 행동에 근거를 두기 때문에 행동주의라는 이름이 어울리다는 말도 있음