Dream Passionate, Live Simple

통계강의

1

Data

1. 정의

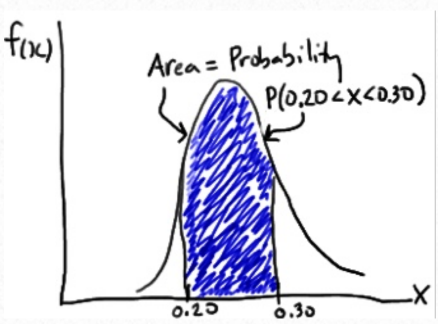
- 추론, 토론, 계산에 사용되는 실제의 정보 (측정값 혹은 통계) (웹스터)
- 정량적, 정성적 변수의 값들의 모임
- 정보를 가진 숫자의 모임
- 데이터 행은 개체 subject , 열은 변수 variable, 셀은 관측값 observation 으로 구성

변수 : 개체의 속성을 설명(구별)하는 속성, 데이터와 변수는 혼재하여 사용하는 개념임

2. 변수 정의

- 개체마다 변하는 관측값
- $X(w) = x$, w=확률실험(관측)의 결과, 관측된 결과를 input, 숫자가 output인 함수
- 통계학의 (확률) 변수는 반드시 확률분포함수 probaility distribution function을 갖고 있음

확률분포함수 : 변수가 가질 수 있는 값이 input, 그에 대응하는 확률이 output인 함수, 변수가 가진 정보를 모두 표현



3. 변수 variable 종류

1) 사회과학

- 범주형 categorical : 명목 nominal 개체를 범주화 , 순서 oedinal 명목 변수 중 크기가 있는 경우
- 측정형 measuremant :구간 interval 배수 개념이 없는 경우, 비율 ratio 배수 개념 적용이 가능한 경우

2) 수리

- 이산형 discrete : 변수가 가질 수 있는 값이 유한
- 연속형 : 변수가 가질 수 있는 값이 무한

3) 분석

- 정량적, 양적 quantitative : 관측값이 측정 가능한 숫자 값으로 표현
- 정성적, 질적 qualitative : 개체를 구별하는 인덱스 (예) 성별, 직업, 거주 지역

4) 통계모형

- 종속 dependent 변수, 반응 response 변수, 목표 target 변수 ; 모형의 결과(y, output)
- 설명 exploratory 변수, 독립 indepenent 변수, 예측 predictor 변수: 모형의 input

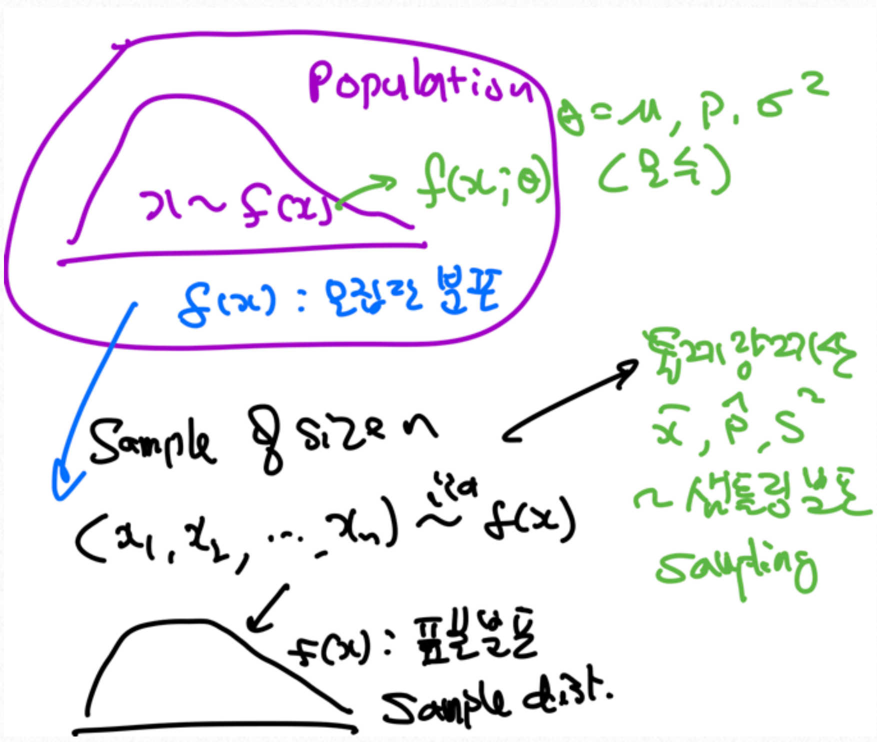
5) 시계열

- 횡단 cross-sectional : 시간의 시점
- 종단 longitudal : 데이터가 시간적 순서를 갖는 경우

4. 데이터분석

1) 일변량분석

- 단일 변수 분석 방법
- 모집단 변수에 대한 정보는 확률분포함수에 있음
- 그러나 모집단 개체 모두를 관측하지 않는 이상 모집단 (확률)분포함수를 알 수 없다 $\sim f(x)$
- 모집단 개체 모두를 관측하여 변수 관측값을 모두 얻어 히스토그램(분포함수)를 그리더라도 확률분포함수 형태를 정확하게 알 수 없음 - 그러므로 이론적 분포함수에 적합성만 검정
- 그러므로 모집단에 대한 추론은 분포함수에 대한 것이 아니라 확률분포의 요약값(평균, 비율, 분산) - 이를 모수 parameter, θ
- $\theta = \mu, p, \sigma^2, \mu_1 - \mu_2, p_1 - p_2, \dfrac{\sigma_1}{\sigma_2}$

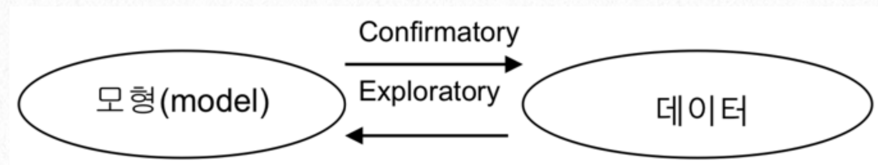


2) 이변량(다변량분석)

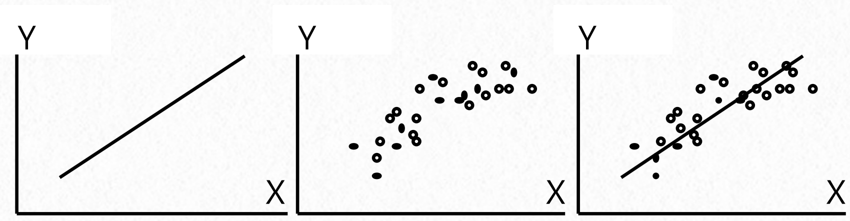
- 두 변수와의 관계 분석 : 선형관계(해석과 다루기 쉬움) 분석

X. I Y	범주형	측정형
범주형	교차분석	분산분석
측정형	범주형 로그선형	상관분석 / 회귀분석

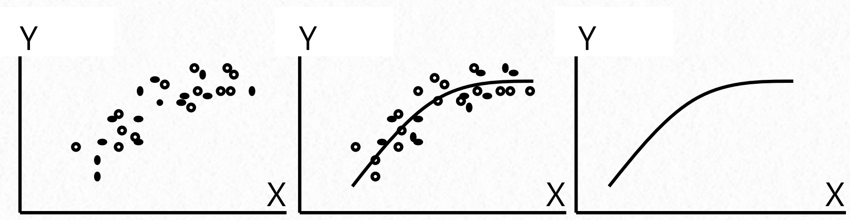
3) 확증적, 탐색적 방법



1)통계적 가설(statistical hypothesis), 모형을 설정하고 2)데이터 수집하여 3)가설 혹은 모형의 유의성(significance)을 검정



1)수집된 데이터를 숫자 요약이나 그래프로 표현하고 2)데이터를 보다 유용하게 만드는 데이터를 재표현하여3)데이터에 내재된 정보나 이론을 도출 - 이론이나 모형은 CDA 방법에 의해 검정



2

변환(정규변환)

1. 변환이 필요한 이유

1) 일변량 : 변수가 하나인 경우

- 통계학의 모수적 검정방법(통계량의 샘플링분포)에는 데이터(변수) 모집단 분포는 정규분포(좌우 대칭인 분포)를 가정한다.
- 물론 평균, 비율에 관한 검정(차이 포함)은 대표본 이론(중심극한정리)에 의해 데이터 모집단 분포에 관계없이 표본평균과 표본비율의 샘플링 분포는 정규분포에 근사하므로 데이터 변환의 필요 없음
- 그러나 분산에 대한 추론에서는 표본분산의 샘플링분포는 카이제곱분포를 이용한다. 그러려면 모집단 분포는 정규분포를 따라야 한다

모집단 분포 population dist.: 데이터가 추출된 모집단의 분포 $x \sim f(x)$, 일반적으로 알 수 없으며 필요한 경우 분포 가정을 함,모집단 분포에 대한 관심보다는 모집단 분포의 특성 값 (모수, parameter)을 추론(가설검정, 추정) 하게 됨

표본 분포 sample dist. : 확률표본 데이터 분포 $f(x)$, 모집단의 분포와 동일함 - 히스토그램을 그려 분포함수의 형태를 볼 수 있음, 이론적 분포(가정한 분포)와 같은지 검정은 분포 적합성(goodness od fits) 검정 실시

샘플링 분포 : 표본 데이터로부터 계산된 통계량의 분포를 의미한다. 통계량이 추정에 사용되면 추정량, 가설 검정에 사용되면 검정통계량이라 한다. (예) 표본평균은 모평균의 추정량이며, 표본평균의 함수인 $(\bar{x} - \mu_0)/(s/\sqrt{n})$ 은 모평균 가설검정 통계량임. 통계량의 분포를 샘플링분포라고 하고 모집단의 분포와 관계없이 표본평균의 샘플링분포는 (중심극한정리에 의해)정규분포에 근사한다.

- 표본이 대표본이 아닌 경우(모집단 분포의 치우침의 크기에 좌우하나 표본크기 20~30이면 대표본) 표본평균의 샘플링 분포는 모집단이 정규분포(적어도 좌우 대칭)이어야 표본평균이 정규분포를 따르고 t-검정방법을 사용한다.
- 소표본이고 모집단이 정규분포를 따른다는 가정이 없다면 모수적 검정방법(통계량 샘플링분포(t-분포)를 이용한 추론)을 사용할 수 없고 비모수적 방법(non-parametric, dist. free 분포자유)을 사용해야 한다.
- 모수적 방법을 사용하려면 모집단의 분포가 정규분포를 따라야 하므로 (1) 모집단의 분포가 정규분포를 따르는지 검정? <=> 표본 데이터의 정규성 검정 (2) 정규성 검정을 만족하지 않으면 변수변환(변수 분포의 정규분포 검정)

2) 모형

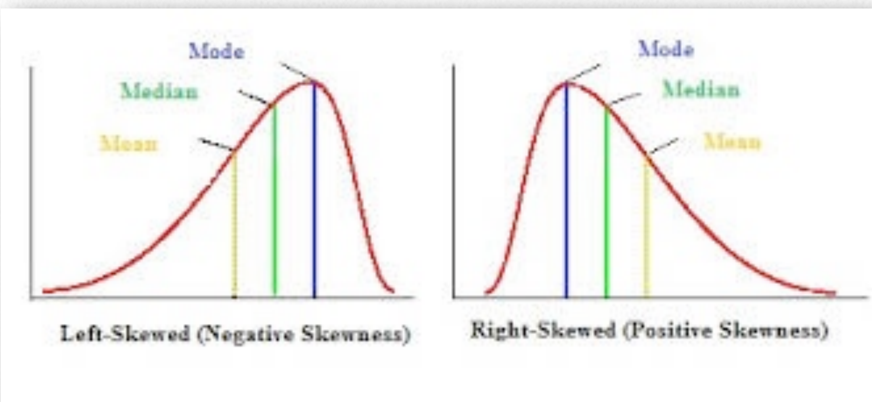
(1) $y = f(x) + e$ 통계모형에서 오차항의 분포는 정규분포를 가정한다.

(2)오차항의 분포가 정규분포를 따라야 모형의 모수에 대한 모수적 방법(통계량의 샘플링 분포, 회귀계수에 대한 t-검정, 분산분석 F-검정) 가능하다

(3)통계모형에서는 종속변수 y만 확률변수이고 설명변수(input)는 결정변수(확률분포함수를 가지지 않음)이므로 오차항이 정규분포이면 종속변수도 정규분포에 근사해야 한다

(4)그러나 통계 선형모형에서 모든 변수(X, Y)는 정규분포에 근사해야 모형의 적합성이 높아짐 - 그리하여 치우침이 있는 데이터의 경우 미리 변환하여 모형에 삽입하게 된다. (예) 소득, 가격, 수능점수

2. 정규변환 normal transformation



1) 일반적 접근

- 우로 치우침 : $\sqrt{x} \rightarrow x^{1/3} \rightarrow ln(x)$-> $log(x)$
- 좌로 치우침 : $x^2 \rightarrow x^3$

2) Tukey Ladder of Power $x' = x^{\lambda}$

- 좌로 치우침 : $\lambda = 2(x^2) < 3(x^3)$

우로 치우침 : $\lambda = 1/2(\sqrt{n}), 1/3(x^{1/3}), 0(ln(x))$

$- > -1/2(-1/\sqrt{x}), -1(-1/x), -2(-1/x^2)$

(1) 단일변수 정규변환 : λ 값을 변환하면서 데이터의 정규성 검정을 실시하여 최적의 λ 값을 찾음

(2) ANOVA (선형모형) : 잔차의 정규성 검정 -> 최적의 λ 값을 찾고 종속변수를 변환하여 다시 분석함

(3) 선형모형에서는 종속변수만의 변환은 오차 분산의 변동을 초래 -> 모든 설명변수 포함 모든 변수를 개별적 정규변환이 필요함

3) Box-Cox tranformation $x' = \dfrac{(x^{\lambda} - 1)}{\lambda}$

Tukey 변환가 동일하지만 최적의 λ 값은 MLE 방법에 의해 찾음

단일변수 정규성 변환에서는 Tukey 방법이
ANOVA에서는 Box-Cox 방법이 적절함

정규성 검정 : 데이터의 분포가 이론적 정규분포를 따르는지 검정하는 적합성 검정임

귀무가설 : 데이터 모집단 분포는 정규분포이다

대립가설 : 정규분포를 따르지는 않는다 <=> 그러나 어떤 분포인지는 모른다.

검정방법 : 위키피디아

1) Anderson Darling 방법

2) Komogorov - Simirov 검정

3) Shapiro - Wilks 검정

3. 정규변환 예제

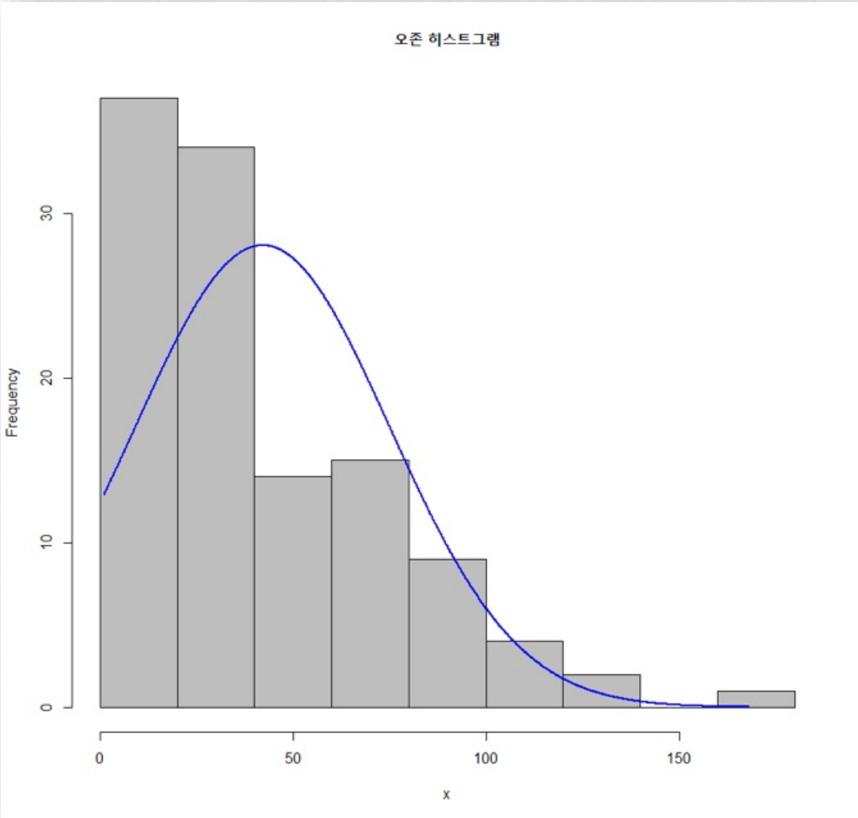
예제 데이터

> *data(airquality) #R* 패키지 데이터 불러오기

히스토그램 그리기 (확률분포함수 포함)

> *library(rcompanion)*

> *plotNormalHistogram(airquality$Ozone, main="오존 히스토그램")*



◉ 우로 치우친 분포를 가짐

정규성 검정 : AD 검정, S-W 검정

> *library(nortest)*

> *ad.test(airquality$Ozone)*

```
        Anderson-Darling normality test

data:  airquality$Ozone
A = 4.5211, p-value = 2.787e-11

> shapiro.test(airquality$Ozone)

        Shapiro-Wilk normality test

data:  airquality$Ozone
W = 0.87867, p-value = 2.79e-08
```

◉ 유의확률(P-VALUE)이 0.01보다 작으므로 귀무 가설(정규분포 따름) 기각됨

◉ 정규변환이 필요함 - 우로 치우침 해결

치우침 해결 : 정규성 검정 포함

>*ad.test(sqrt(airquality$Ozone))* #제곱근변환

> *ad.test(airquality$Ozone^(1/3))* #세제곱근변환

> *ad.test(log(airquality$Ozone))* #로그변환

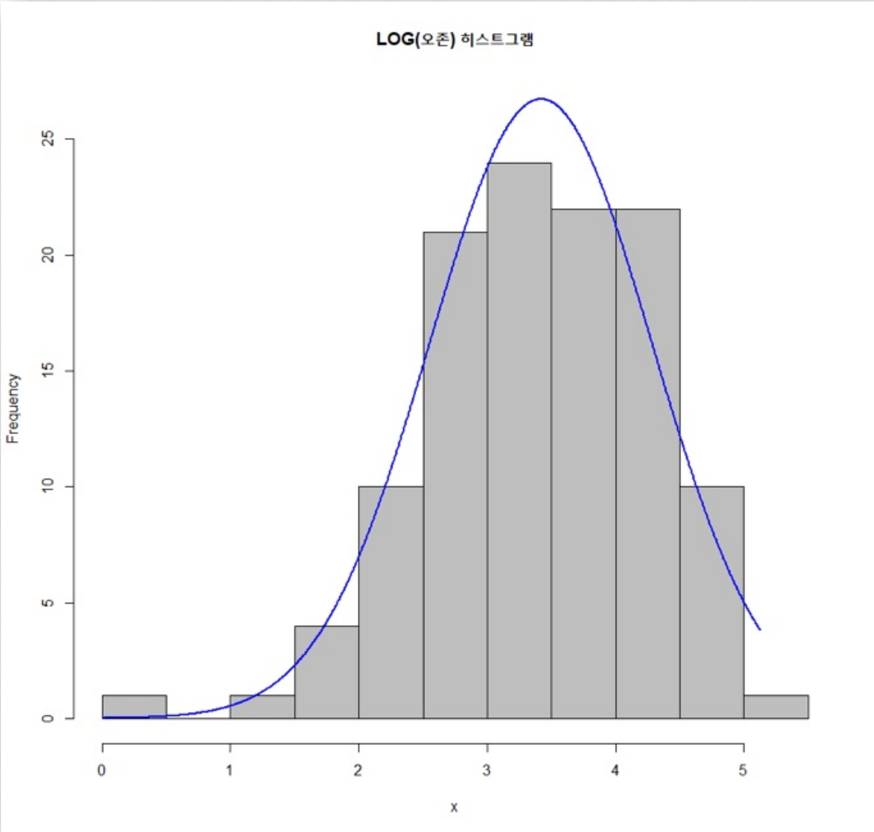
```
> ad.test(log(airquality$Ozone)) #로그변환

        Anderson-Darling normality test

data:  log(airquality$Ozone)
A = 0.46497, p-value = 0.2497
```

◉ 최종 로그변환 시 정규분포를 따름

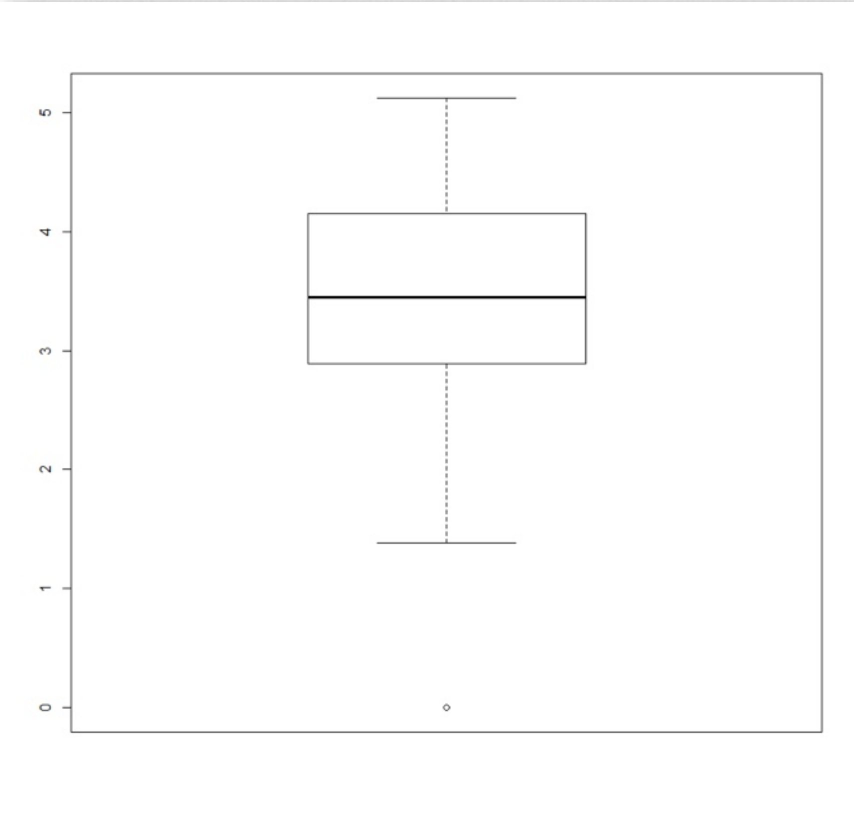
>*plotNormalHistogram(log(airquality$Ozone), main="LOG(오존) 히스토그램")*



결과가 좌로 치우쳐 보이는 것은 치우침의 문제가 아니라 이상치(극단값)이 존재하기 때문임

이상치 진단

```
> boxplot(log(airquality$Ozone))$out
[1] 0
```



- 이상치를 진단할 수 있는 나무상자 그림을 그려본 결과 0 값이 이상치로 판단되었음
- 이를 제거하면 보다 더 정규분포에 근사할 것임

그러나 로그변환의 경우 Shairo-Wilks 검정에 의한 정규분포 적합성은 통과하지 못한다.

```
> shapiro.test(log(airquality$Ozone)) #로그변환

        Shapiro-Wilk normality test

data:  log(airquality$Ozone)
W = 0.97168, p-value = 0.01471
```

이처럼 정규성 검정 방법에 따라 정규분포 적합성 결과가 달라진다? 그럼 어느 방법을 따를까? 통계학에서는 가장 많이 사용되는 방법이 Sapiro -Wilk 방법이다. 왜냐하면 정규성 검정을 가장 보수적으로 하기 때문이다. 보수적이라 함은 정규분포를 따른다고 결론 내리리기 어려움을 의미한다. A-D 방법이 보다 쉽게 정규분포를 따른다고 결론 내린다.

패키지 이용하여 Tukey Power 변환의 최적의 λ 값 찾기 transformTukey()함수

```
> library(rcompanion)
> t.ozone<-transformTukey(airquality$Ozone)

    lambda      W Shapiro.p.value
409    0.2 0.9871          0.3399

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```

최적의 λ=0.2 -> S-W 정규성 검정 결과 유의확률은 0.34로 정규성 만족

$\lambda = 0.5$ = 제곱근 변환과 동일

$\lambda = 0.33$ = 세제곱근 변환과 동일

$\lambda = 0$ = 로그 변환과 동일함