일변량 분석 개념

일변량분석은 개체의 특성을 측정한 변수가 하나인 통계분석 방법

변수의 종류

(수리 통계)

- 이산형(discrete): 측정 결과를 셀 수 있는 경우이다. 성별, 직업, 교통량, 나이 등이 여기 해당된다.
- 연속형(continuous): 측정 결과가 무한이(infinite) 많은 변수를 연속형 변수라 한다. 즉 변수의 범위 (range) 중 어떤 구간을 설정하더라도 측정치가 발생할 수 있는 경우로 키, 몸무게, IQ, 소득 등이 여기에 해당된다.

(자료 분석)

측정형 변수(metric, measurable, quantitative): 실험 개체의 측정 가능한 특성을 측정한 변수, 측정 단위가 존재; 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다. (예) 교통사고 건수, 나이 (년)

(1) 구간 interval : 값의 크기가 등간 (온도, 증가율)

(2) 비율 ratio : 정대 0 이 존재하며 두 값의 비교가 배수(times) 가능 : 대부분 측정형

⇨ **평균&표준편차, 중앙값&IQR**

분류형 (범주형) 변수(non-metric, classified, categorical, qualitative): 개체를 분류하기 위해 측정된 변수를 의미하며 성별, 결혼여부 등이 그 예이다.

(1) 명목형(nominal) : 개체를 분류만 한다. 성별, 결혼여부, 학력

(2) 순서형(ordinal) : 순서를 가진다. 성적 (A, B, ..) 소득수준 (상, 중, 하), 리커트 척도(5 점, 매우 만족, 만족, ..., 매우 불만족)

⇨ **빈도분석, 비율**

분석방법

비율 추론

평균 추론

분산 추론

모수적 방법 vs. 비모수적 방법

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

in EDA

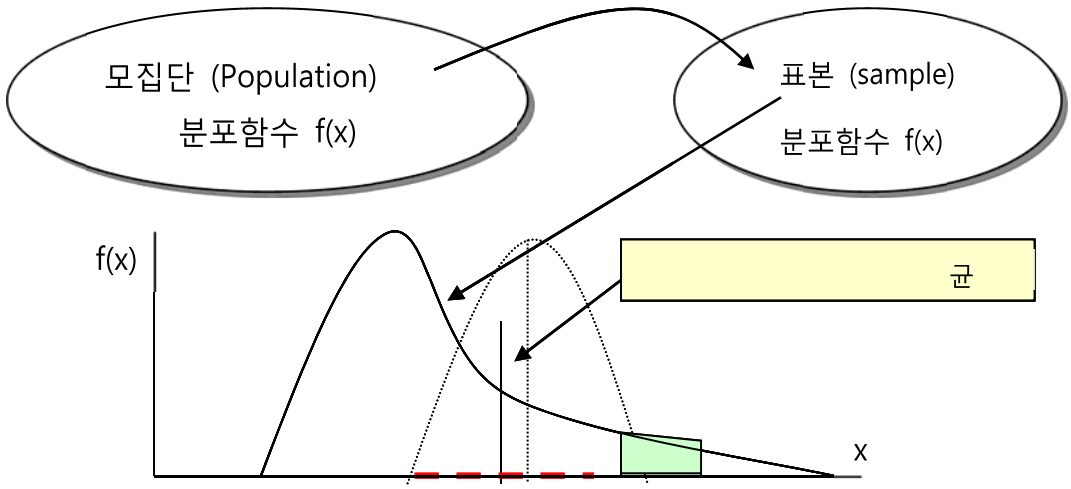
- 데이터를 그래프로 표현하여 개체를 구성한 모집단의 정보를 얻는다.
- 범주형 데이터는 바 차트로 표현되어 비율(상대빈도)로 요약됨
- 측정형 데이터는 데이터의 중앙 위치, 흩어진 정도, 봉우리 형태 등을 표현

그래프 요약 필요성

개별 변수에 대한 일변량분석은 변수가 가진 정보를 그래프로 요약하거나 주요한 숫자 값으로 (통계량) 요약하게 된다. 앞에서 언급하였듯이 그래프 요약은 변수에 대한 가정이 (확률표본, independently and identically distributed) 성립하는지 진단하게 된다.
그래프 진단은 데이터의 좌우대칭 (종모양 symmetric, bell shaped), 이상치 진단을 하게 된다. 그 이유는 평균이 주된 도구이기 때문이다. 평균은 치우침과 이상치에 취약하다.

확률분포함수 probability density function

일변량 데이터가 가진 정보는 확률밀도함수에 의해 표현된다.
확률변수 표본공간 S 의 모든 원소(결과 w)에 실수 x 을 대응시킨 규칙을 확률변수 $X(w)=x$
확률분포함수 확률변수 X 의 확률밀도함수(probability density function, pdf)는 확률변수 X 가 가지는 값 x와 그에 대응하는 확률 $p(x)$ 을 그래프, 수식, 표 형태로 나타낸 것이다.



만약 좋은 표본이 (모집단의 축소판) 추출되었다면 표본의 분포는 모집단의 분포와 같다. 그러므로 만약 모집단의 분포(실선) 함수를 안다면 다음을 구할 수 있다. 모집단개체 중

일부 구간 (빨간 점선 구간)에 속한 개체 비율 (확률)? 그러나 불행히도 표본 자료의 분포로부터 함수식 $f(x)$ (확률 분포 함수: 실선 그래프)을 아는 것은 불가능하다.
그러므로 일반적으로 통계학에서는 모집단의 분포에 대한 가정을 하거나 (예: 모평균 검정을 위한 t-검정에서 소표본일 경우 모집단 정규 분포 가정) 대표적인 분포 함수 (이항분포, 포아송 분포, 지수분포, 감마 분포, t-분포, F-분포, 정규 분포)를 규정하고 있다.

데이터 분석 : 확률분포를 역할

데이터분석에서는 모집단의 확률분포함수 형태를 알고자 하는 것은 아니다. 모집단의 특성을 나타내는 모수 parameter (예: 평균, 비율, 분산)에 대한 추정 및 가설검정이 중요하다.
그러면 데이터 분석에서 표본 데이터 확률분포를 시각화 하는 이유는? 모집단으로부터 확률표본기법을 통하여 얻어지는 데이터

중심극한정리

에 의하면 표본의 크기가 크면 표본 평균($\bar{x}$)의 분포는 모집단의 분포와 관계 없이(모르더라도) 정규 분포를 따른다 의미? CLT 는 표본 분포 함수(sample distribution)에 대한 것이 아니라 표본 평균의 분포 (sampling dist.)에 관한 것이다. 표본 분포 함수(실선)는 여전히 모집단의 분포 함수 $f(x)$와 동일하다. 그러니 가정이나 사전 정보 없이는 그래프만 가지고는 알 수 없다. 그러나 대표본(n>20~30)일 경우 표본 평균의 분포는 CLT 에 의해 정규분포를 따른다. 이것이 왜 그렇게 중요하다. 우리가 모집단에 관심을 가질 때 모집단 자료 전체에 대한 정보(분포 함수)를 구하는 것보다는 모집단의 자료 정보를 요약한 값 (이를 모수: parameter 라 함)에 관심을 갖게 된다. 예를 들면 중앙의 위치는?(평균, 중앙값) 자료의 흩어진 정도는? (표준편차, 범위) 특히 모집단의 평균(μ)에 관심을 갖게 된다. 이 경우 모집단의 분포를 모르더라도 CLT 에 의해 다음 사실을 알 수 있다. (이전 페이지 그림 점선 그래프 참고)

$$\Pr(|\frac{\bar{x}-\mu}{s/\sqrt{n}}| \le z_{\alpha/2}) = 1-\alpha \;\rightarrow\; \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} : \text{모평균 95\% 신뢰 구간}$$

만약 대표본이 아닌 경우는 t-분포를 이용해야 되는데 이런 경우 모집단은 정규 분포를 따르고 있음을 가정하게 된다.

In R RGui

(1) $B(p=0.1)$ 베르누이 확률분포를 그리시오.

(2) 베르누이 분포 $B(p=0.1)$ 에서 크기 2 인 확률표본을 추출하여 표본 평균을 구하시오

(3) 표본평균을 이용하여 95% 신뢰구간을 구하고, 구해진 신뢰구간이 모평균을 포함하고 있는지 판단하시오.

(4) 위의 (2)-(3) 작업을 100 번 반복하여 100 개 표본평균을 구하고, 100 개 95% 신뢰구간 중 모평균을 포함하지 않은 신뢰구간의 수를 적으시오.

(5) 얻어진 표본평균 100 개를 이용하여 표본 데이터 확률분포함수를 그리시오. 즉 히스토그램을 그리시오.

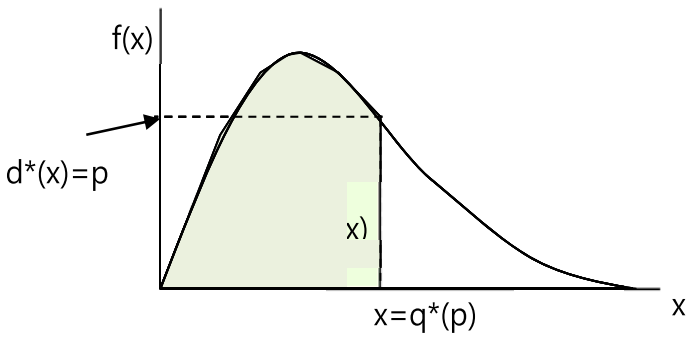
(6) 베르누이 분포 $B(p=0.1)$ 에서 크기 20 인 확률표본을 추출하여 평균을 구하고, 이런 작업을 100 번 반복하여 100 개 표본평균을 구하시오.

(7) 얻어진 표본평균 100 개를 이용하여 표본 데이터 확률분포함수를 그리시오. 즉 히스토그램을 그리기

위의 동일한 작업은 $Normal(10,5)$ 에서 실시하시오.

통계분포함수 (statistical distribution function)

함수	기능
d*(x, 모수)	확률밀도함수 확률 값, f(x)
p*(x, p, 모수)	분포함수 값, F(x)
q*(p, 모수)	역분포함수 값, $F^{-1}(p)$
r*(n, 모수)	분포함수 따르는 데이터 n 개 랜덤하게 생성



beta	beta	shape1, shape2
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df2
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's t	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

http://www.statmethods.net/management/index.html (데이터 관리)

제어문 control statement

for

```
for (var in seq) expr
```

for(변수 in 연속) {문장}	연속에 지정된 값만큼 변수 값이 변화하면서 '문장'을 반복 실행한다.

```
> for(i in 1:3){x=i*3;print(x)}
[1] 3
[1] 6
[1] 9
```

```
> x=c(50,60,80,90,95)
> for(i in length(x):1){print(i)}
[1] 5
[1] 4
[1] 3
[1] 2
[1] 1
```

if-else

```
if (cond) expr
if (cond) expr1 else expr2
```

if (조건) {문장}	조건이 만족하면 문장이 실행된다.

```
> for(i in 1:10){
+ if(i%%2==0){print(i)}}
[1] 2
[1] 4
[1] 6
[1] 8
[1] 10
```

```
> for(i in 1:10){
+ if(i%%2==0){cat('Even ',i,"\n")}
+ else{cat(i,'는 홀수','\n')}}
1 는 홀수
Even  2
3 는 홀수
Even  4
5 는 홀수
Even  6
7 는 홀수
Even  8
9 는 홀수
Even  10
```

while

```
while (cond) expr
```

while(조건) {문장}	조건이 만족하는 동안 문장 반복 실행된다.

```
> while(i<=3){print(i);i=i+1}
[1] 1
[1] 2
[1] 3
```

함수 만들기

I:₩Books₩R활용₩ch1.R - R 편집기

```
#하이 로우 게임
game=function(){
x=round(runif(1,0,1)*100)
for(i in 1:10){
cat("Guess? ")
g=scan(nmax=1,quiet=T)
if (g>x) {cat("Lower. \n")}
if (g<x) {cat("Higher. \n")}
if (g==x) {cat("Good!");break}}}
```

I:₩Books₩R활용₩ch1.R - R 편집기

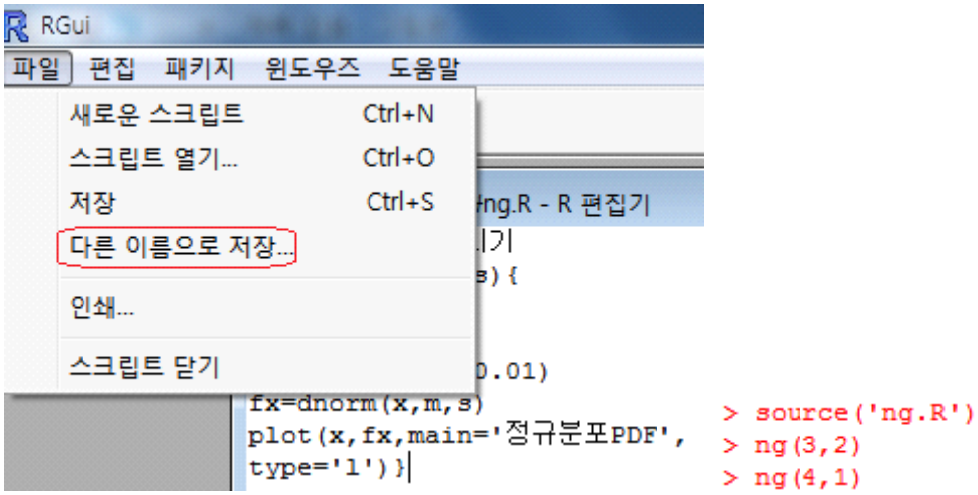
```
#정규분포함수 그리기
ng=function(m,s){
max=m+4*s
min=m-4*s
x=seq(min,max,0.01)
fx=dnorm(x,m,s)
plot(x,fx,main='정규분포PDF',
type='l')}
```

I:₩Books₩R활용₩ch1.R - R 편집기

```
#확률분포함수 그리기: 베타분포
x=seq(0,1,0.01)
fx1=dbeta(x,5,2)
fx2=dbeta(x,2,2)
x11()
split.screen(c(1,1))
screen(1)
plot(x,fx1,type='l',
 main='베타분포 PDF',col='red',
 ylab='fx',xlab='x',
 xlim=c(0,1),ylim=c(0,2.5))
screen(1)
plot(x,fx2,type='l',col='blue',
 ylab='fx',xlab='x',
 xlim=c(0,1),ylim=c(0,2.5))
```

	plot(x, y, main=‘ ’, sub=‘ ’, xlim=c(a, b), ylab=‘ ’, type= ‘ ’)	그래프 함수

자신의 함수 활용하기



줄기 잎 그림 stem and leaf

+ 진단 내용

1) 분포의 개략적인 형태를 알 수 있다.
   (1) 좌우 대칭인가? 아니면 skewed 되었는가?
   (2) 봉우리(modal)는 하나인가? 아니면 여러 개인가?
2) 이상치의 존재 여부를 쉽게 파악할 수 있다.

+ 데이터

53	145
43	621
33	262
45	208
46	362
55	424
41	339
55	736
36	291
45	58
55	498
50	643
49	390
47	332
69	750
51	368
48	659
62	234
45	396
37	300
50	343
50	536
50	543
58	217
53	298
57	1103
53	406
61	254
47	862
56	204

줄기	잎
0	5
1	4
2	00135699
3	03346699
4	029
5	34
6	245
7	35
8	6
9	
10	
11	1

(정렬)

줄기	잎
0	5
1	4
2	60931950
3	63936904
4	290
5	34
6	245
7	35
8	6
9	
10	
11	1

(정렬 않음)

+ 그리는 순서

- 자료를 크기 순으로 정리한다.
  자료의 수가 많을 때는 자료 정렬을 수작업 하기 어려움으로 이 단계는 무시해도 되지만 자료를 크기 순으로 정렬해 놓으면 plot 을 그리기 편리하다.
- 자료를 살펴 줄기와 잎을 결정한다.
  CEO 연봉 자료를 살펴보면 100 단위를 줄기로 하고 10 단위 이하를 잎으로 하여 plot 을 그리면 될 것이라는 것을 알 수 있다. 줄기 수는 히스토그램의 계급 구간 수에 해당되므로 8~12 정도가 적절하다. 적정 개수가 아닌 경우 줄기 수 조정에 대해서는 다음에 다루기로 한다.
- 한 열에 줄기(stem)를 먼저 그린다.

위에서 100 단위 이상을 줄기로 하기로 결정하였고 자료의 최소값이 58, 최대값이 1103 이므로 0 부터 11 까지 줄기를 한 열에 크기 순으로 적는다.

줄기	
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	

- 줄기(stem) 옆에 잎을 그린다.

  잎을 그리는 방법은 간단하다. 줄기 바로 뒤의 숫자를 줄기 옆에 차례로 적으면 된다. CEO 연봉 자료는 잎이 두 자리이지만 앞에 것 하나만 적으면 된다. 굳이 반올림하는 수고를 할 필요는 없다. 줄기-잎 그림의 목적은 자료의 분포 형태와 이상치를 아는 것이 주된 목적이기 때문이다.

줄기-잎 그림

```
ds=read.csv("ceo.csv",header=T)
stem(ds$CEO.Salary)
```

+ 엑셀에서 콤마가 있는 파일 형식으로 저장한 후 읽어 들인다.
+ ds$변수명; 오브젝트 ds 내의 변수명 변수를 이용 지정

```
The decimal point is 3 digit(s) to the right of the |

0 | 178889
1 | 000012334446
2 | 0
3 | 3
```

줄기-잎 그리기

http://lib.stat.cmu.edu/DASL/Stories/SingerHeights.html

합창단원의 키에 대한 (단위: inch) 데이터이다.

1) 각 파트의 줄기-잎 그림을 그리고 해석하시오.

2) 키 데이터 전체에 대한 줄기-잎을 그리고 해석하시오.

+ Stem-leaf plot 해석하기

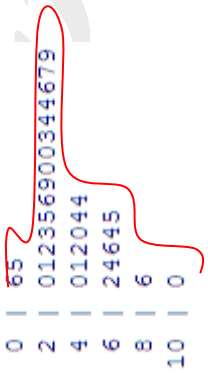
> 자료의 분포 형태
stem-leaf plot 을 통하여 자료의 분포 형태를 알 수 있으므로 분포의 형태를 알 수 있다. 이는 히스토그램과 같은 역할이다.
> 봉우리 (최빈값) 위치 및 개수 => 봉우리의 개수가 집단의 개수이다
> 좌우 대칭 여부
> 자료의 범위 및 분산
> 이상치 존재 여부 및 위치

(히스토그램과 비교)

줄기-잎 그림을 90 도 회전하면 히스토그램 (이를 bar chart 라고도 함)이 된다.히스토그램은 자료의 값의 정보가 상실되지만(실제 값은 알 수 없고 빈도만 바의 크기로 나타난다) stem-leaf plot 은 자료 값이 나타난다. 그러므로 히스토그램에 비해 더 많은 정보를 얻을 수 있다.

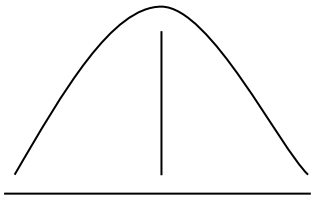
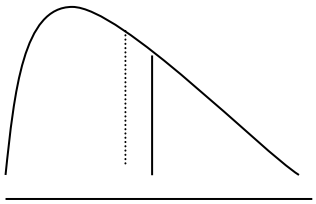
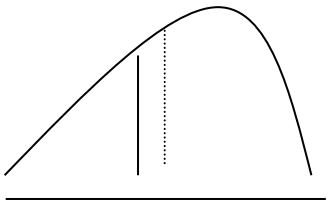
(1) 확률분포함수 추정

위의 예제처럼 stem-leaf plot 의 정점을 연결하면 확률분포함수를 얻게 된다. 아래 그림은 모집단 CEO 연봉의 확률밀도함수의 추정 형태이고 ($f(x)$) 면적은 1 이다.



(2) 대칭, 치우침 여부

symmetric (bell-shaped) 좌우 대칭, 종 모양	skewed to the right positively skewed 우로 치우침	skewed to the left negatively skewed 좌로 치우침

		
평균과 중앙값 일치	평균>중앙값	평균<중앙값
좌우대칭으로 만들려면… 자료 변환을 하면 된다.	$X^* = \sqrt{X}$ ➔ mild pos. $X^* = \log(X)$ ➔ pos. $X^* = -1/\sqrt{X}$ ➔ severe pos. $X^* = -1/X$ ➔ more severe	$X^* = X^2$ ➔ mild neg. $X^* = X^3$ ➔ extreme neg.

(정규성 검정) Anderson-Darling test for normality

```
library(nortest)
ad.test(ds$CEO.Salary)
```

```
> ad.test(ds$CEO.Salary)

        Anderson-Darling normality test

data:  ds$CEO.Salary
A = 1.1633, p-value = 0.003674
```

(연봉 모평균에 대한 95% 신뢰구간 구하기) conf.level=0.95

```
> t.test(ds$CEO.Salary)

        One Sample t-test

data:  ds$CEO.Salary
t = 8.3815, df = 19, p-value = 8.333e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  896.6596 1493.5404
sample estimates:
mean of x
   1195.1
```

```
output=t.test(ds$CEO.Salary)
names(output)
output$conf.int
```

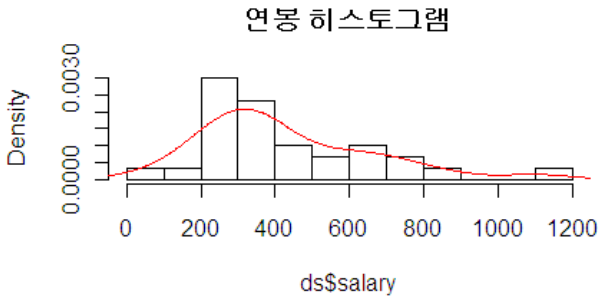
```
> output$conf.int
[1]  896.6596 1493.5404
```

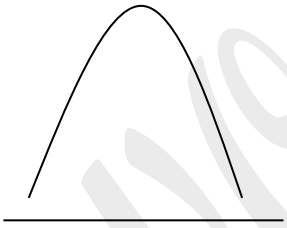
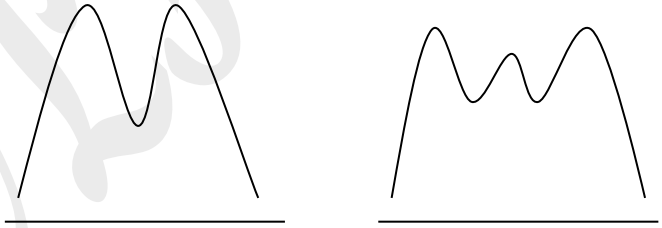
히스토그램 그리기

우로 치우침이 있으므로 제곱근 변환, 혹은 로그 변환 데이터 중 어느 변환이 더 좌우 대칭분포에 가까운지 알아보시오.

(3) 봉우리 위치 및 개수

히스토그램의 봉우리는 분포의 최빈값에 해당되는 부분으로 일반적으로 최빈값은 하나일 가능성이 가장 높다. 구간 설정에 따라 바로 옆의 구간이 동시에 최빈값이 되는 경향이 가끔 나타나기도 한다. 다음의 경우는 bi-modal 분포 함수라 하지는 않는다. 왜냐하면 구간을 조정하면 봉우리가 하나로 될 수 있기 때문이다. CEO 연봉은 단봉 형태를 갖는다.



단봉 uni-modal	다봉 bi-modal / multi-modal
	

봉우리가 2 개 이상인 의미는 모집단이 하나 이상일 가능성이 많다. 예를 들어 한남 대학생들 100 명의 몸무게를 조사하여 히스토그램을 그리면 bi-modal 형태가 될 가능성이 높다. 왜냐하면 여자와 남자 몸무게의 차이가 나기 때문에 그런 현상이 발생한다. 즉 측정 변수의 특성에 따라 모집단이 나누어진다. 용돈을 조사하여 히스토그램을 그려보면 아마 봉우리가 3-4 개일 가능성이 있다. 왜? 학년별 차이로 인하여… 이처럼 어떤 변수를 측정하느냐에 따라 같은 모집단이라도 봉우리의 개수가 다를 수 있다. 봉우리가 2 개 이상인 경우는 집단을 분리하여 추정 및 검정을 시행하는 것이 바람직하다. 그러나 집단에 대한 정보가 없다면 데이터를 분리하여 분석하는 것이 쉽지 않다.

(왜 좌우 대칭이어야 하나?)

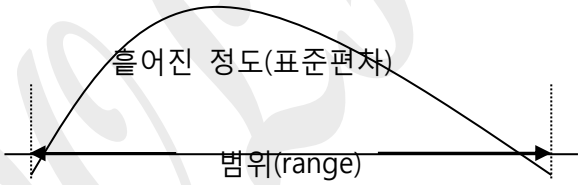
1) 회귀 분석, 분산 분석 등 대부분의 통계 분석에서 종속변수는 정규분포를 따르고 있다는 가정을 한다. 만약 이것이 무너지면 t-검정, F-검정을 사용할 수 없다. ➔ 3 학년 수업에서 배우기를...

2) 대표본 표본 크기 n 의 크기? : 자료 분석의 목적은 그래프 정리(bar chart, pie chart)나 숫자적 정리(평균, 표준편차)에서 끝나는 것이 아니라 이 정보를 가지고 모수(예:모집단의 평균)를 추정하거나 그에 대한 가설을 검정하게 된다. CEO 30 명의 연봉 자료를 이용하여 전체 CEO 의 연봉에 대해 알고 싶은 것이다.

통계 소프트웨어에서 출력되는 p-값은 two-sided(양측 검정) 가설 검정 시 값을 출력한다. 그러므로 위의 경우 대립 가설을 $H_a : \mu \neq 350$(양측 검정) 설정하면 p-값이 0.0821 로 0.05 보다 크므로 귀무가설을 기각할 수 없으나 대립 가설을 $H_a : \mu > 350$(단측 검정) 설정하면 p-값이 0.04105 이므로 0.05 보다 적어 귀무가설을 기각하고 연봉은 높아졌다고 결론 지을 수 있다. 그러므로 양측 검정 결과 귀무가설이 기각되면 같은 유의수준에서 단측 검정 결과도 귀무가설을 기각한다.

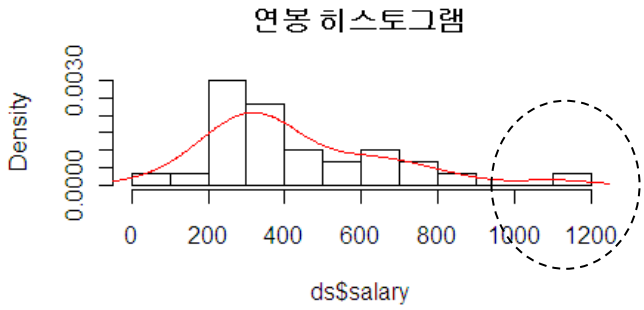
(4) 범위와 흩어진 정도

분포의 형태를 알 수 있으므로 자료의 범위(range=최대값-최소값)와 흩어진(spread) 정도를 알 수 있다.



(5) 이상치(outlier) 발견

다른 관측치에 비해 매우 크거나 적은 관측치를 이상치(outlier)라 한다. 이런 이상치는 히스토그램에서 쉽게 발견될 수 있다. 히스토그램이나 stem-leaf plot 의 경우 다른 관측치와 멀리 떨어져 있으면 이를 이상치라 한다. CEO 연봉 자료에서 이상치는 연봉이 1103(백만)인 사람이다. 물론 이 값이 이상치인지는 검정 통계량을 이용하여 (Box-plot 이나 검정 방법을 이용하여 검정해야 하지만 우선 쉽게 찾을 수 있다는 장점이 있다. CEO 연봉의 경우 다른 CEO 에 비해 연봉을 이상적으로 높게 받는 CEO (이를 이상치라 함) 가 있음을 알 수 있다.



이상치가 발견되면 그 해결책은

이상치인 관측치의 원 자료를 확인, 입력 오류인지 살펴본다. 오류가 있으면 정정한다.

이상치의 대상인 개체를 조사해 문제가 있는 개체이면 자료에서 제외한다.

예를 들면 1103(백만) 연봉을 받는 사람을 조사하였더니 외국인 전문 사장이었다. 국내 CEO 연봉으로 간주하기 어렵다면 제외

여전이 유효한 데이터이면 자료 변환을 통하여 이 문제를 해결하게 된다. 변수변환 (자료 변환)을 통하여 이상치 문제가 해결되면 이는 치우침의 한 부분이 된다.

+ 줄기 수 조정

일반적으로 자료의 분포 형태를 잘 파악하기 위해서는 줄기의 수가 8-10 개 정도 되어야 한다고 한다. 연봉 데이터 예제에서 본 것처럼 줄기 수는 변수 측정치의 범위에 의해 결정된다. 그러므로 줄기의 수를 조정하여 적절한 줄기-잎 그림을 그려야 한다.

> 줄기 수가 너무 많으면 (squeezed stems)

줄기를 일정한 수만큼 합치는 방법을 생각하면 된다. 만약 줄기가 1-20 까지 있다면 1-2, 3-4, 5-6, …, 19-20 을 각각 줄기로 하면 줄기 수가 20 개에서 10 개로 줄어든다. 이처럼 줄기 수에 따라 2 배, 3 배, 4 배씩 줄이면 된다.

> 줄기 수가 너무 적으면 (stretched stems)

줄기를 2 등분(double stem) 혹은 5 등분(five-line stem)하여 사용한다.

(예) double stem: 1_* (1.0~1.4), 1. (1.5~1.9)

(예) five-line stem: 1_* (.0, .1), 1_t (.2, .3), 1_f (.4, .5), 1_s (.6, .7), 1. (.8, .9)

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

자료
11
12
15
17
18
19
20
20
21
21
24
28

줄기	잎
1*	1
1t	2
1f	5
1s	7
1.	89
2*	0011
2t	
2f	4
2s	8
2.	

적정 줄기 수에 관한 공식

> Sturges formula $L = [1 + \log_2 n]$ (예) n=30 ➔ L=5

> Velleman formula $L = [2\sqrt{n}]$ (예) n=30 ➔ L=10

> Dixon-Kronmal formula $L \le [10 \log_{10} n]$ (예) n=30 ➔ L=14

그러나 위의 공식에 의해 줄기 수(L)를 결정하면 자료 값에 따라 줄기를 결정하기 어렵고 분포 형태를 제대로 알기 어려운 문제가 있어 이 공식들은 사용되지는 않는다. [x]의 의미는 x 보다 크지 않는 최대 정수 값을 의미한다. [2.9]=2 / [3.1]=3

R 활용

```
hist(ds$CEO.Salary, nclass=10, freq=F, main="Histogram")
lines(density(ds$CEO.Salary),col="blue")
```

- nclass 옵션은 구간의 개수를 결정한다.
- freq 옵션은 빈도 대신 상대빈도 (확률)을 y-축으로 사용하라는 옵션
- 함수 lines()는 확률밀도함수를 그리라는 옵션

히스토그램 그리기

http://lib.stat.cmu.edu/DASL/Stories/SingerHeights.html

합창단원의 키에 대한 (단위: inch) 데이터이다.

1) 각 파트별 히스토그램을 그리고 해석하시오. (확률밀도함수도 그리시오)

2) 키 전체에 대한 히스토그램을 그리고 해석하시오.

상자 수염 그림 box whisker plot

Stem and leaf(줄기-잎) plot 은 자료의 분포의 형태(좌우 대칭, 단봉) 파악과 이상치를 발견할 수 있는 도구이다. 그러나 S-L plot 만 가지고는 정확한 중앙 위치, 자료의 사분위 값의 위치 (25%, 50%, 75% 부분의 값들), 이상치라 시각적으로 판단되는 관측치가 정말 이상치인지 알 수 진단해야 한다.

Box-whisker (상자 수염) 그림은 plot 중앙값 (때로 평균까지) [상자 안의 실선], 사분위 위치, [상자 끝단], 자료의 최대값, 최소값(whisker), 이상치 존재 여부(bullet)를 그려 놓은 상자 형태의 그림이다.

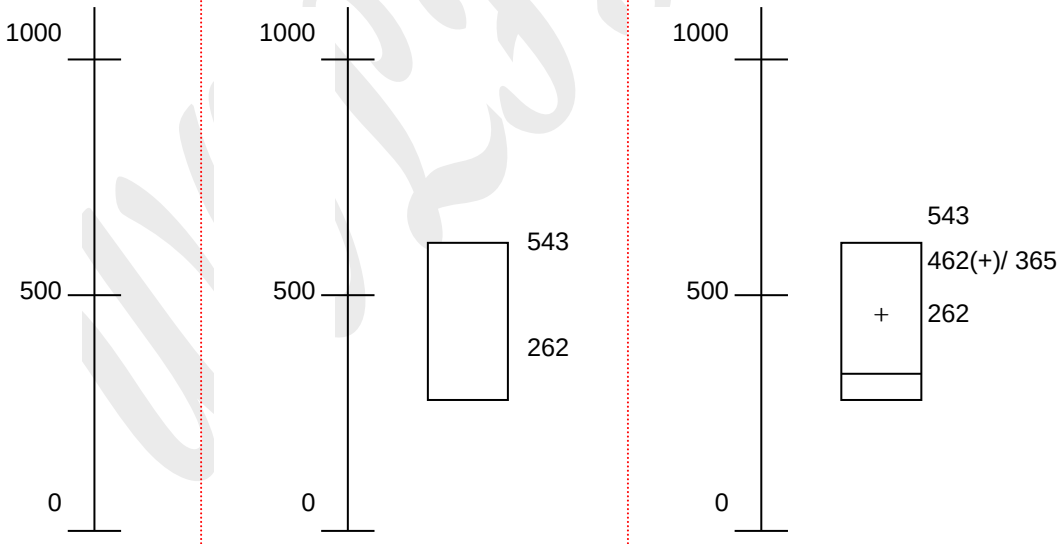
+ Box and whisker plot 그리기

상자로부터 나온 선이 수염처럼 생겨 Box and whisker plot 이라고 하는 Box plot 을 그리는 순서는 다음과 같다. [CEO 자료 이용]

[순서 1] 자료의 최소값, 최대값을 이용하여 y 축 선을 그린다.

[순서 2] Q1, Q3 를 이용하여 상자를 그린다. 상자의 넓이는 아무 의미가 없다.

[순서 3] 상자 가운데 중앙값을 그리고 평균은 기호로 (+) 표시한다.



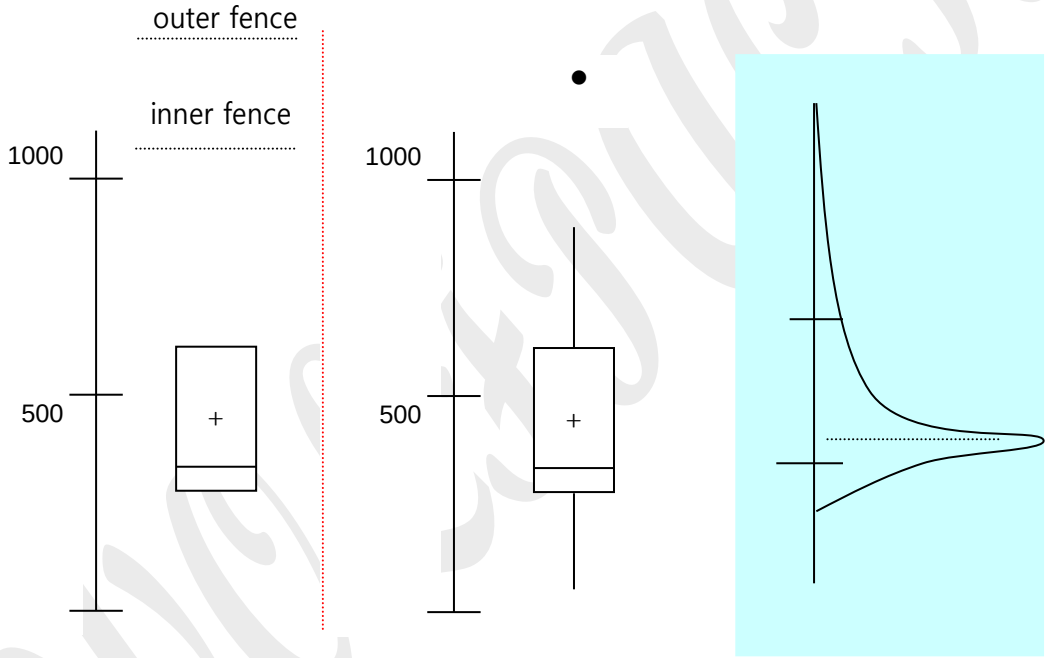
Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

[순서 4]IQR 을 이용하여 가상 선 (imaginary line) Inner fence, Outer fence 를 그린다. 가상 선은 실제 상자 그림에 표시되지 않는다. 이상치 존재 여부를 표현하기 위한 임시 선이다.
IQR=(543-262)=281

Inner fence ($Q_1 - 1.5 \times IQR$, $Q_3 + 1.5 \times IQR$) =(-159.5, 964.5)

Outer fence ($Q_1 - 3 \times IQR$, $Q_3 + 3 \times IQR$) =(-581, 1386)

[순서 5]수염과 이상치를 표시한다.관측치 중 Inner fence 를 넘지 않는 최대, 최소값까지 수염을 그린다. Fence 를 넘는 관측치를 이상치라 (outlier) 한다. outer fence 까지 넘는 관측치는 severe (극심한) 이상치, inner fence 만 넘으면 mile 이상치라 한다. [CEO 에서 1103 은 mild 이상치]

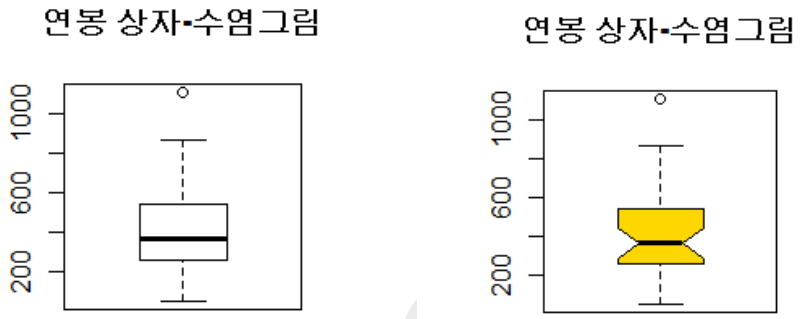


[분포 함수]

줄기-잎 그림

```
ds=read.csv('ceo.csv')
boxplot(ds$salary,main="연봉 상자-수염그림")

boxplot(ds$salary,notch=TRUE,
col=(c("gold","darkgreen")),main="연봉 상자-수염그림")
```



+ Box Plot 해석하기

분포의 형태

박스, 박스 안의 선(중앙값), 수염의 길이를 이용하여 분포의 형태를 짐작할 수 있다. 박스에 50% 자료가 있고 박스 위 부분에 25%, 박스 아래 부분에 25%가 있다. 박스 내에서도 중앙 선 위 부분에 25%, 아래 부분에 25%가 있으므로 분포의 형태를 알 수 있다. CEO 연봉은 우측으로 치우친 형태를 띠고 있음을 알 수 있다. (위쪽, 즉 확률밀도함수의 오른쪽 상자와 수염 부분이 왼쪽에 비해 살짝 길다) 그러므로 평균이 중앙값보다 크므로 역시 치우친 형태임을 알 수 있다. 확률분포함수는 빨간 선이다.

단점은 봉우리의 개수를 알지 못하는 단점이 있으므로 보완적으로 줄기-잎 그림 그린다.

중앙값, 산포 정도, 군집

자료 관측치의 중앙 위치, 그리고 관측치들의 어디에 모여 있는지 (군집), 자료 값들의 흩어진 정도를 파악할 수 있다. 중앙 값이 350 부근(실제로는 365)임을 알 수 있다. 값의 범위(range), 사분위 값을 대략적으로 알 수 있다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

이상치 존재 판단

자료 관측치 중 다른 값들에 비해 지나치게 크거나 작은 관측치, 이상치의 존재 여부를 파악할 수 있다. 연봉이 1100 이상을 받는 CEO 는 이상치임을 알 수 있다.

줄기-잎 그림을 통해서는 이상치 존재 여부를 짐작할 수 있지만, 상자 수염 그림을 통하여 판단 가능하다.

stem-leaf plot 과는 관측치 값들에 대한 정보를 얻을 수 없다는 단점이 있으나 표본 분포의 형태도 파악할 수 있고 중앙값과 이상치를 표시하여 표본 자료의 정보 파악이 용이하다. 특히 box-whisker plot 은 집단간 자료의 분포 차이 비교, 모평균 차이 검정 시 매우 유용한 plot 이다.

상자 수염 그리기

http://lib.stat.cmu.edu/DASL/Stories/SingerHeights.html

합창단원의 키에 대한 (단위: inch) 데이터이다.

1) 키 데이터 전체에 대한 줄기-잎을 그리고 해석하시오.

2) 각 파트의 상자 수염 그림을 한 그래프에 그리고 해석하시오.

상자 수염 그리기 FASTFOOD.xls

미국 fast-food 레스토랑 5 개의 Drive-through 서비스 소요시간을 측정한 것이다. 레스토랑 상자-수염 그림을 한 화면에 그리고 결과를 해석하자.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

숫자 요약 구하기 : 중앙위치

(평균 mean)

크기의 중앙, 관측치 크기 (magnitude)의 중앙으로 모든 관측치를 합한 후 데이터 크기로 나눈 값

$$\mu = \bar{x} = \frac{\sum x_i}{n}$$

(치우침 skewness 존재) 평균은 치우침에 취약하므로 다음 사항을 처치한다.

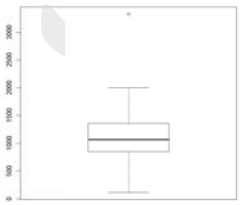
(1) 데이터 변수 변환 (예) 로그 변환, 제곱근 변환

(2) 절삭평균 trimmed mean; 양측 꼬리 일부 데이터를 제외한 평균

(3) Winsorized mean; 양측 꼬리 부분 데이터를 바로 전 데이터로 대체하여 구한 평균

```
> summary(ds.ceo$Salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  116.0   872.8  1065.0  1195.0  1354.0  3325.0
> mean(ds.ceo$Salary,trim=0.1)
[1] 1110.062
> fivenum(ds.ceo$Salary)
[1]  116.0  848.5 1065.0 1357.5 3325.0
```

```
boxplot(ds.ceo$Salary)
ds.ceo0=subset(ds.ceo,Salary<2500)
boxplot(ds.ceo0$Salary)
```



(이상치 존재) 평균은 이상치에 매우 취약하므로 제거 후 평균을 계산하다.

R 활용

다시 아래 이상치가 발생 => 제외해 보자

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

(중앙값 median)

순서의 중앙, 관측치를 크기 순으로 (순서통계량) 하였을 때 중앙에 위치한 관측치

중앙 깊이 Median Depth = $\frac{(n+1)}{2}$

$$M = x_{((n+1)/2)}$$

숫자요약 : 흩어짐 (spread)

(분산 variance)

관측치들이 평균으로부터 떨어진 거리의 제곱 합을 데이터 크기 혹은 (크기-1)로 나눈 값이다. 관측치들의 크기의 흩어짐을 측정한 값이다.

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}, \quad s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

(표준편차 standard deviation)

분산의 양의 제곱근이다. σ, $s = \sqrt{s^2}$, 단위가 평균과 동일하여 분산 대신 (평균, 표준편차)를 숫자 요약으로 제시한다.

순서 통계량 (order statistics)

크기가 n 인 표본 자료의 관측치(observation) $(x_1, x_2, \ldots, x_n)$을 크기 순으로 정렬한 후 가장 작은 관측치를 $x_{(1)}$, 그 다음 큰 관측치를 $x_{(2)}$, …, 가장 큰 관측치를 $x_{(n)}$이라 표현하고 $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$을 순서 통계량이라 한다.

1) $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

2) 최소값 (min): $x_{(1)}$, 3) 최대값 (max): $x_{(n)}$

4) 중앙값 (median): $x_{(\frac{n+1}{2})}$ (n 이 홀수) $[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]/2$ (n 이 짝수)

Quartile depth = $QD = \frac{\lfloor MD \rfloor + 1}{2}$

(범위 range)

순서의 흩어짐에 대한 측정 값, $R = x_{(\max)} - x_{(\min)}$

(midrange)

범위도 이상치나 치우침에 영향을 받으므로 제삼분위~제일사분위값을 정의한다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

```
> range(ds.ceo$Salary)
[1]  116 3325
> quantile(ds.ceo$Salary)
     0%     25%     50%     75%    100%
 116.00  872.75 1065.00 1353.75 3325.00
```

Quartile (사분위 값) 구하기

>일 사분위 (First Quartile, Low Quartile) Q1; 자료의 25%가 그 값보다 작고 자료의 75%가 그 값보다 크게 될 때 그 값을 큰 일 사분위라 정의한다.
>이 사분위 (Second Quartile, Median) Q2=Median; 자료의 50%가 그 값보다 작고 자료의 50%가 그 값보다 크게 될 때 그 값을 큰 이 사분위라 정의한다. 이를 특히 중앙값이라 한다.
>삼 사분위 (Third Quartile, Upper Quartile) Q3; 자료 75%가 그 값보다 작고 자료의 25%가 그 값보다 크게 될 때 그 값을 큰 삼 사분위라 (Lower percentile) 정의한다.
Inter-Quartile Range (IQR=Q3-Q1); (삼 사분위 값- 일 사분위 값)을 IQR(=Q3-Q1)로 정의한다.

깊이 (depth)

각 사분위 값을 구하려면 자료의 깊이 (depth) 개념을 이용하면 편리하다. (Tukey 제안) 관측치를 크기 순으로 정렬한 후 각 양쪽 끝에서 1 부터 번호를 매겨 그 번호를 자료의 깊이라 정의한다. 즉 최대값, 최소값의 깊이는 각 1 이다. Depth(중앙값=M)= $(n+1)/2$ 이다.
➔ CEO 자료에서 Depth(M)=15.5 이다. 크기 순으로 정렬했을 때 15 번째 관측치와 16 번째 관측치의 평균이다. $(x_{(15)}+x_{(16)})/2=365$ 이다.
Depth(Q1)=Depth(Q3)=([Depth(M)]+1)/2 이다. [x]=x 를 넘지 않는 최대 정수. [2.6]=2
CEO 자료에서 Depth(Q1/Q3)=([15.5]+1)/2=8 이다. $Q1=x_{(8)}=262$ 이고 $Q3=x_{(18)}=543$ 이다.

EDA 에서의 skewness 개념

$$Skew=\frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)},\quad -1\le skew\le 1$$

왜도(skewness): $\frac{E(X-\mu)^3}{\sigma^3}$ 분포의 치우침을 나타내는 값으로 0(정규분포, t-분포)이면 좌우 대칭이고 양의 값이면 우로 치우침(skewed to the right, positively skewed), 음의 값이면 좌로 치우침(skewed to the left, negatively skewed) 이다. 검정통계량은 존재하지 않는다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

kurtosis(첨도): $\frac{E(X-\mu)^4}{\sigma^4}$ 분포의 첨예(뾰족하다) 정도 나타내는 값으로 정규분포는 3, 3 보다 크면 완첨 (leptokurtic)하다고 (봉우리가 낮고 완만하여 평균 주위에 데이터가 있을 확률이 정규분포에 비해 낮고 꼬리는 가는 형태) 하고 3 보다 적으면 급첨 (platykurtic)하다 한다 (봉우리가 높고 뾰족하며, 꼬리가 두꺼운 형태). 검정 통계량 없음. 정규분포를 0 하는 경우 계산된 첨도-3 을 첨도 값으로 주는 패키지도 있음.

R 활용

함수 moment()는 패키지 moments 를 설치하여야 한다.

적률 momnent => k 차 적률 계산식 $E(X-\mu)^k$

```
library(moments)
moment(ds.ceo$Salary,order=2,central=T)
moment(ds.ceo$Salary,order=3,central=T)/sd(ds.ceo$Salary)^3
moment(ds.ceo$Salary,order=4,central=T)/sd(ds.ceo$Salary)^4
```

왜도=0.97, 우로 치우침

첨도=3.591-3=0.591, 정규분포에 비해 중심은 낮고 완만하며 꼬리가 가늘다.

변동계수(variation coefficient):

측정 단위에 따라 표준 편차의 값의 크기가 달라지므로 단위가 다른 두 집단을 비교하는 경우 두 표준 편차의 단위를 같게 할 필요가 있다. 이를 위하여 표준편차를 평균으로 나눈 값에 100 을 곱한 값을 변동 계수(CV: Coefficient of Variation)라 하고 상대 변동(분산) 개념으로 정의하고 있다. 표본 자료의 평균이 $\bar{x}$, 표준 편차가 s 인 경우 $CV = \frac{s}{\bar{x}} \times 100(\%)$ 이다.

‖EXAMPLE‖ 고등학교 3 학년인 A 학생과 B 학생의 공부 습관을 조사하여 한 달간 조사하여 A 학생은 평균 3 시간, 표준 편차는 0.5, B 학생은 6 시간 표준 편차 0.8 인 결과를 얻었다. 어느 학생이 더 꾸준히 공부하는 습관을 가지고 있을까? 이에 대한 답을 위해 변동 계수를 계산하면 된다.

A 학생 공부시간에 대한 변동 계수 $= 0.5/3 \times 100(\%) = 16.7$ (%)
B 학생 공부시간에 대한 변동 계수 $= 0.8/6 \times 100(\%) = 13.3$ (%)

위의 계산 결과 B 학생이 더 꾸준히 공부하는 습관을 가지고 있다고 결론 지을 수 있다.

```
> sd(ds.ceo$Salary)/mean(ds.ceo$Salary)
[1] 0.5335733
```

데이터 정렬

attach(ds)

newds=ds[order(CEO.Salary),]

newds2=ds[order(-CEO.Salary),]

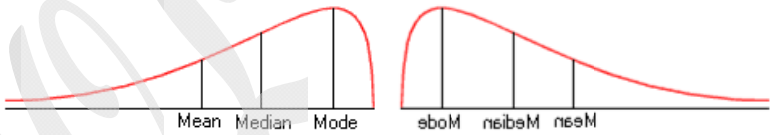
```
library(doBy)
summaryBy(bod~location,data=ds.lake, FUN=c(mean,sd))
  location bod.mean   bod.sd
1       L1      2.2 1.549193
2       L2      4.6 2.065591
3       L3     21.1 4.840799
```

R 활용

FASTFOOD.xls (레스토랑 드라이브 through 서비스 시간) 레스토랑 별 평균, 표준편차, 중위수, 사분위, 그리고 CV 값을 계산하시오.

(1) 그래프 요약(치우침, 이상치)와 비교하시오.

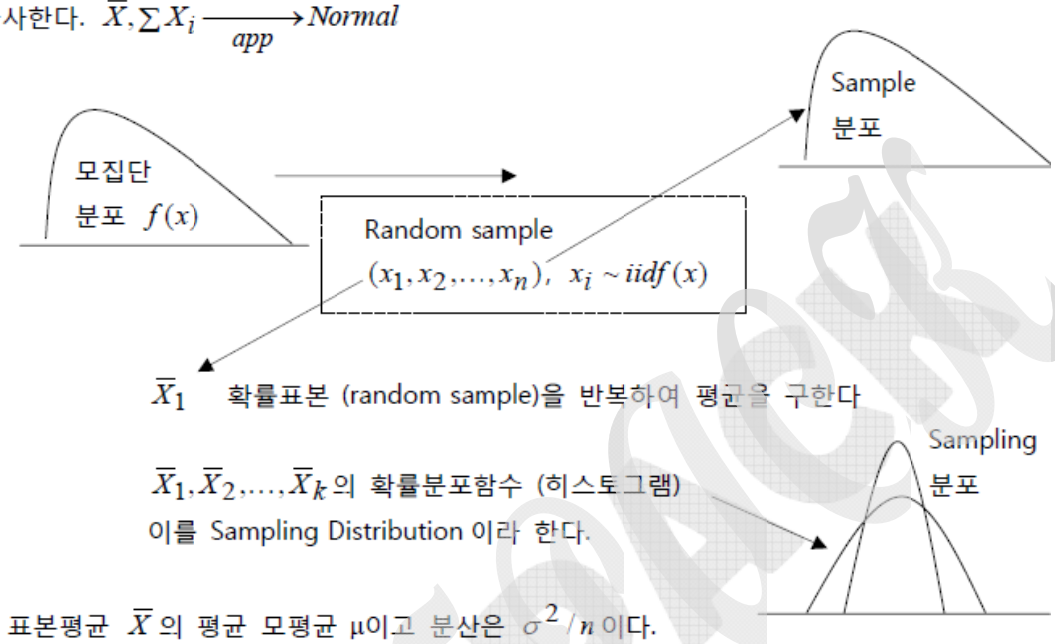
(2) 변동계수를 이용하여 레스토랑 서비스 시간에 대한 평가를 하시오.



치우침 진단 및 해결

(중심극한정리)

모집단의 분포가 어떠하든지 표본 평균의 (합의) 분포는 정규분포에 표본크기가 커질수록 근사한다. $\bar{X}, \sum X_i \xrightarrow[app]{} Normal$



> 표본평균 $\bar{X}$의 평균 모평균 μ이고 분산은 σ^2/n 이다.
> 그러므로 표본의 크기가 커질수록 표본평균의 분산은 작아진다.
> 표본의 크기가 커질수록 sampling 분포는 정규분포에 (좌우대칭) 근사한다.

iid? Independently and Identically

CLT in R

평균이 0.5 인 지수분포로부터 표본의 크기 n=5, 10, 15, 20 인 표본을 추출하여 평균을 구하고, 이런 작업을 100 번 하여 평균에 대한 히스토그램을 그리시오. split.screen(c(2,2)) 함수를 이용하여 4 개 히스토그램을 한 화면에 그리시오. 커널분포함수도 그리시오.

```
x=rexp(100,1)
h=hist(x, breaks=10, col="blue", main="EXP(lambda=1)",freq = FALSE)
lines(density(x)) # plots Kernel Density
```

Let (x_1, x_2, …, x_n) be an iid sample drawn from some distribution with an unknown density f. We are interested in estimating the shape of this function f. Its *kernel density estimator* is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \quad = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where $K(\bullet)$ is the kernel — a symmetric but not necessarily positive function that integrates to one — and $h > 0$ is a smoothing parameter called the *bandwidth*. A kernel with subscript h is called the *scaled kernel* and defined as $K_h(x) = 1/h\ K(x/h)$. Intuitively one wants to choose h as small as the data allows, however there is always a trade-off between the bias of the estimator and its variance; more on the choice of bandwidth later. A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. The Epanechnikov kernel is optimal in a minimum variance sense, though the loss of efficiency is small for the kernels listed previously, and due to its convenient mathematical properties, the normal kernel is often used $K(x) = \phi(x)$, where ϕ is the standard normal density function. **(From Wikipedia)**

시각적 진단

o 히스토그램이나 상자수염 그림의 치우침

숫자요약 진단

o 평균과 중위수 차이

o 수리 왜도 : $\frac{E(X-\mu)^3}{\sigma^3}$

o EDA 왜도 : $Skew = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)}$

o 정규성 검정 : Anderson Darling 검정, K-S 검정, Shapiro-Wilks 검정

데이터가 정규분포를 따르는지 검정한다. 히스토그램이나 상자-수염 그림으로는 데이터의 치우침 정도만 시각적으로 판단할 수 있을 뿐 정규분포를 따르는지 검정하는 것은 아니다. 데이터가 정규분포를 따르는지 알아보는 방법을 정규성 검정이라 한다.
검정방법은 다수 있으나 가장 널리 사용되는 방법은 Anderson-Darling (SPSS 사용), Shapiro 방법, 그리고 Pearson Chi-Square 방법이다.

```
> sf.test(ds) #Shapiro-Francia test for normality

> pearson.test(xbar) #Pearson Chi-Sqaure for Normality
```

치우침 해결

o Power 변환 $Y^* = \begin{cases} X^3, \text{left} \\ X^2, \text{mild left} \\ \sqrt{Y}, \text{mild right} \\ \ln(Y), \text{right} \\ 1/Y, \text{severe right} \end{cases}$

치우침 진단과 해결

http://lib.stat.cmu.edu/DASL/Datafiles/nursinghomedat.html (BED~FEXP) 변수의 치우침을 진단하고 치우침을 해결하시오. 시각적 진단 (히스토그램과 Kernel 분포함수 이용)과 정규성 검정 활용하여 치우침 진단하고, 적절한 파워 변환에 의해 데이터 치우침 해결

BED = number of beds in home
MCDAYS = annual medical in-patient days (hundreds)
TDAYS = annual total patient days (hundreds)
PCREV = annual total patient care revenue ($hundreds)
NSAL = annual nursing salaries ($hundreds)
FEXP = annual facilities expenditures ($hundreds)
RURAL = rural (1) and non-rural (0) homes

봉우리 문제 진단 및 해결

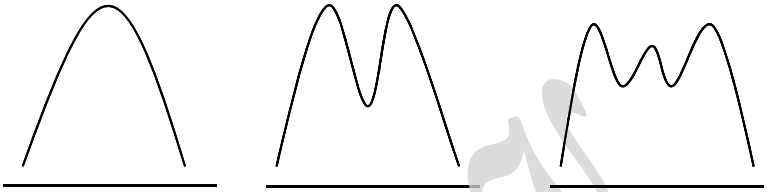
(봉우리 개수 의미)

o 봉우리 개수는 서로 다른 개체 집단을 의미

o 성별을 고려하지 않은 키/몸무게 데이터, 학년을 고려하지 않은 월 용돈 지출 데이터

진단

히스토그램 활용



해결

집단을 분리하여 분석

이상치 진단 및 해결

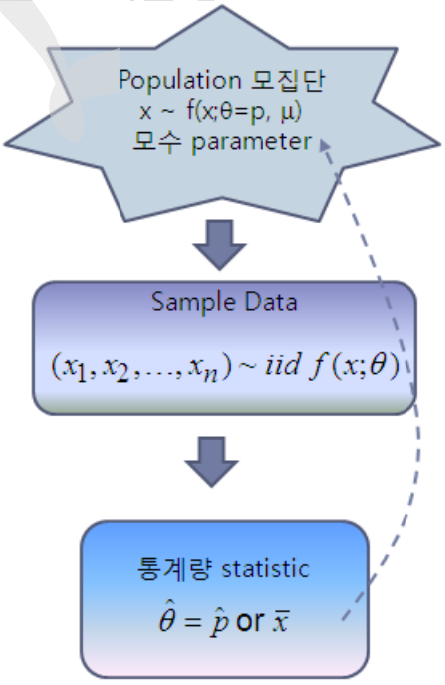
진단

상자-수염 그림

해결

이상치 제거

일변량 분석



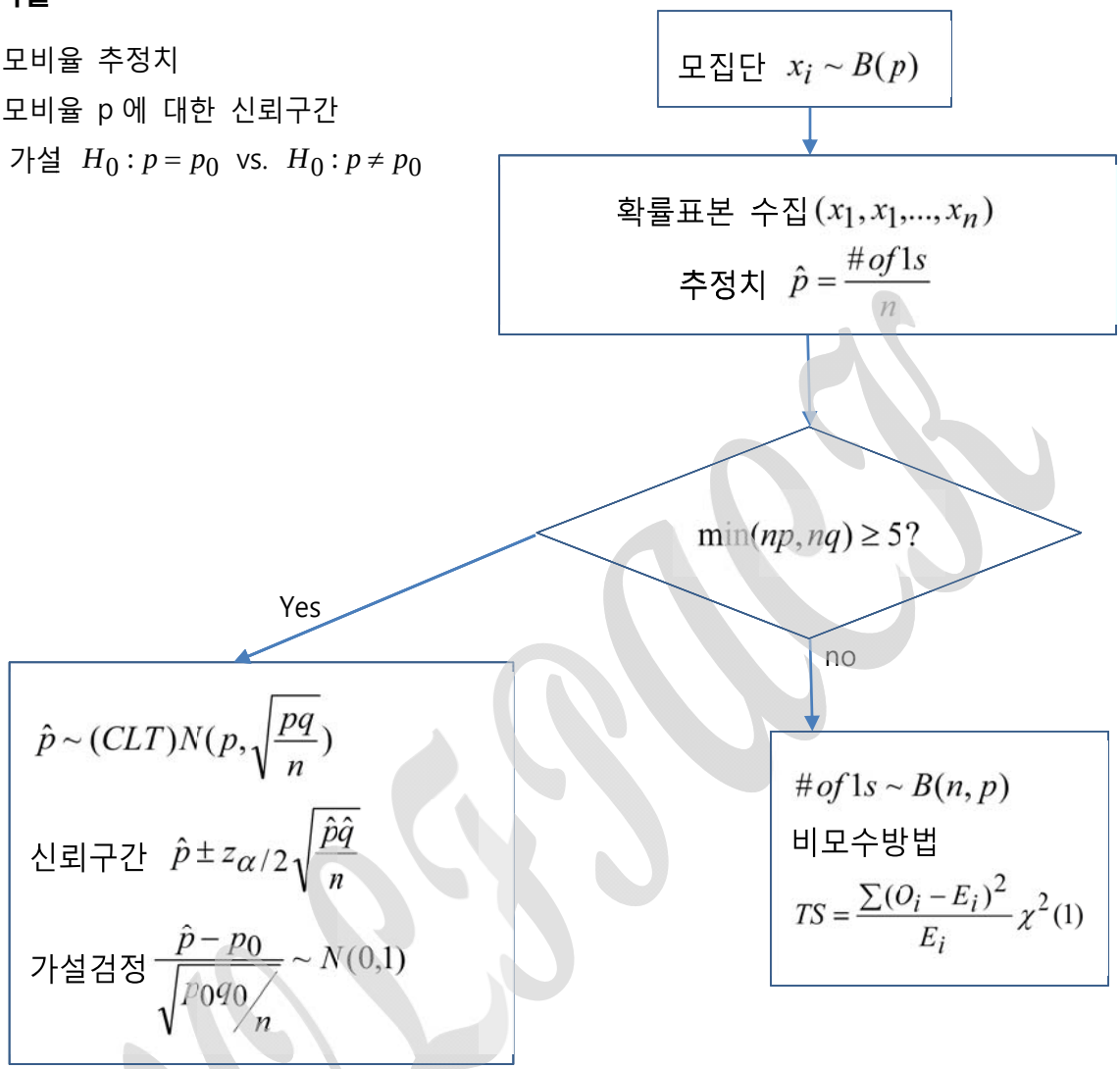
Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

모비율

o 모비율 추정치

o 모비율 p 에 대한 신뢰구간

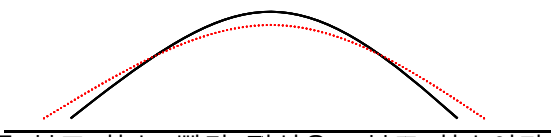
o 가설 $H_0 : p = p_0$ vs. $H_0 : p \neq p_0$



정규 분포와 t-분포의 관계

n 이 커지면 $t(df = n) \rightarrow Normal(0,1)$ 이다. 소표본(small sample)일 경우 중심 극한 정리를 사용할 수 없으므로 표본 평균 $\bar{x}$ 의 분포는 정규 분포라 할 수 없다. 대신 **모집단이 정규분포를 따르면** 다음 분포는 t-분포를 따르므로 소표본인 경우 모집단 평균 가설 검정은 t-분포를 이용한다.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(df = n-1) \rightarrow \text{평균=0, 분산=} \frac{n}{n-2}$$



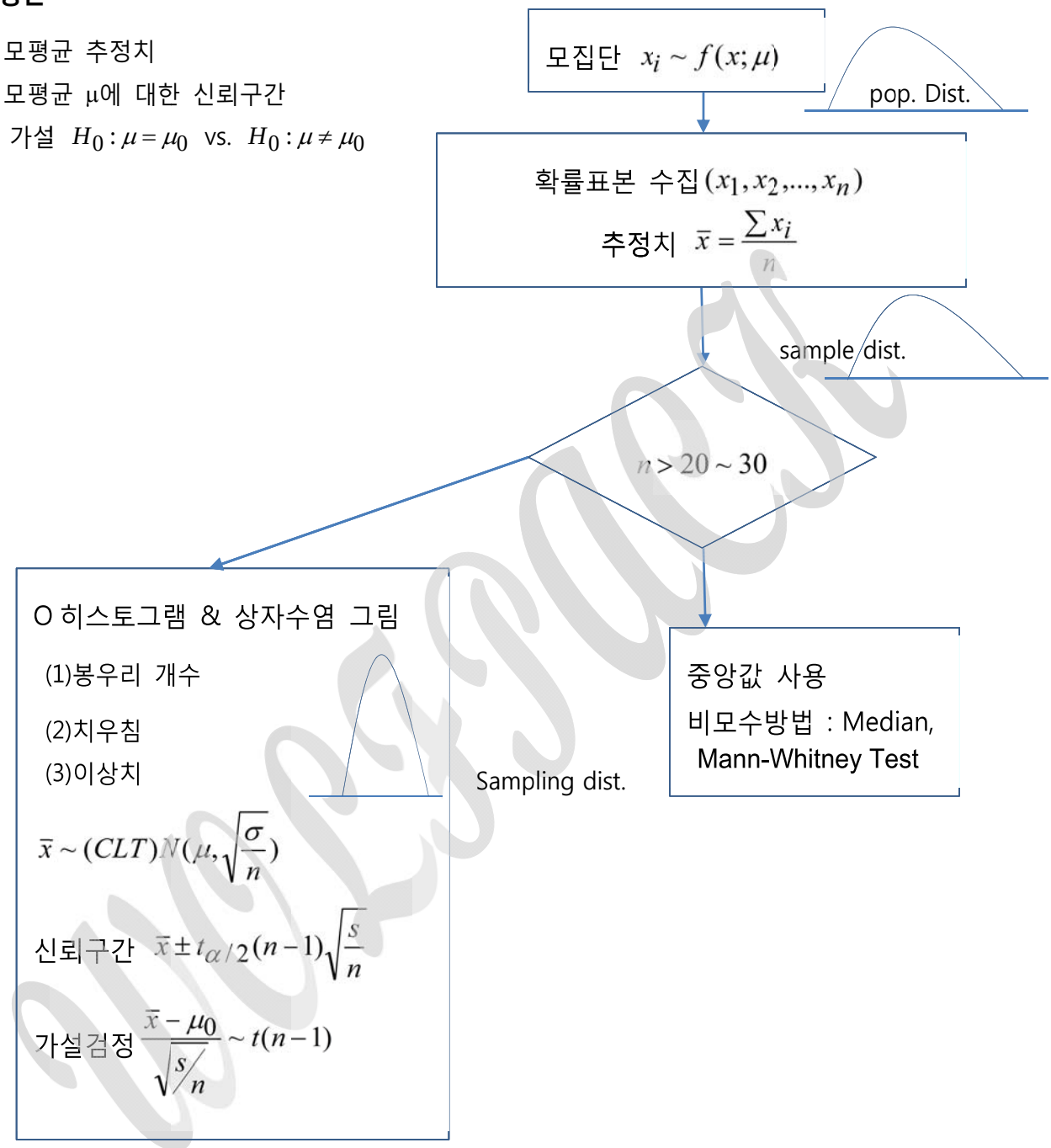
까만 선은 표준 정규 분포 함수, 빨간 점선은 t-분포 함수이다. Why? t-분포의 분산이 크다.

모평균

o 모평균 추정치

o 모평균 μ에 대한 신뢰구간

o 가설 $H_0 : \mu = \mu_0$ vs. $H_0 : \mu \neq \mu_0$



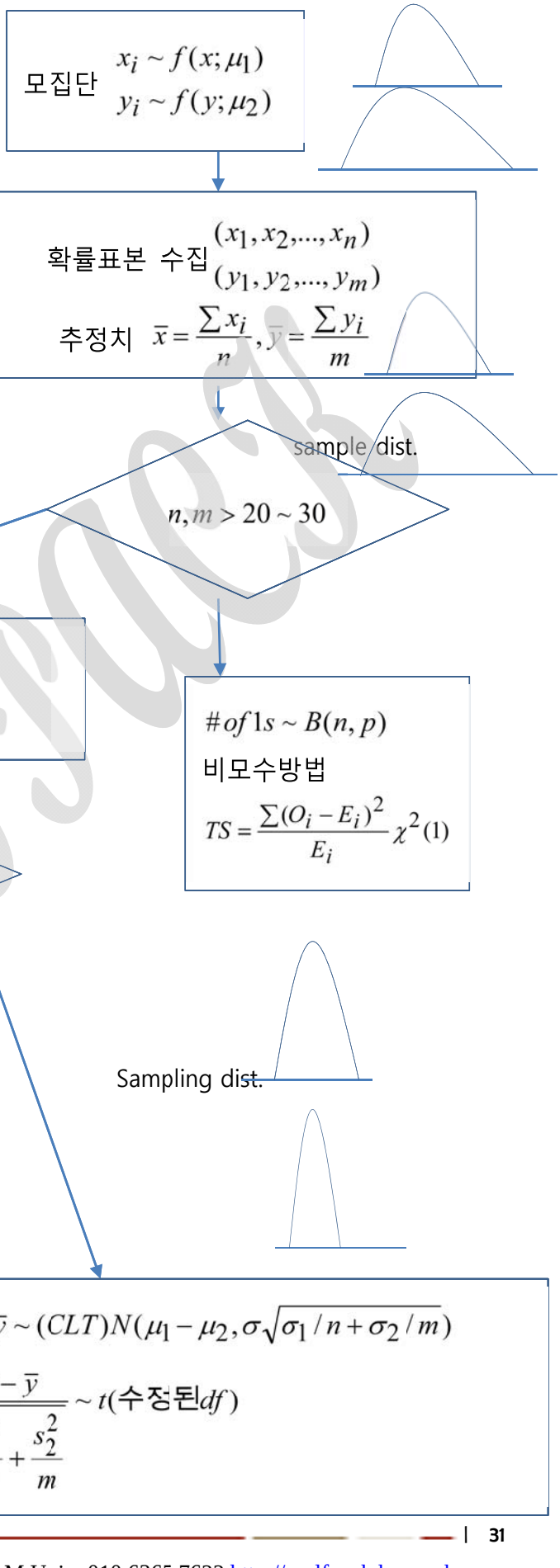
Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

모평균 차이 검정

o 모평균 추정치

o 모평균 $\mu_1 - \mu_2$ 에 대한 신뢰구간

o 가설 $H_0 : \mu_1 = \mu_2$ vs. $H_0 : \mu_1 \neq \mu_2$



METER.txt

이 데이터는 47 개월 동안 A 시 전체(7,000 개) 주차요금 징수기로부터 수금한 주차비이다. CON 은 주차비 징수 대행기관이 7,000 개 전체를, CITY 는 시청직원이 징수한 시청 주변 일부 47 개에서 징수한 요금이다.

위의 데이터를 이용하여 (대행기관, 시청직원 직접 징수) 주차요금 1 년 평균 징수요금 95% 신뢰구간을 구하시오.

Wealth.txt

지역별 기업인의 재산, 나이, 지역을 조사한 데이터이다.

ASIA 와 유럽 기업가의 재산 평균 차이에 대한 95% 신뢰구간을 구하고 차이가 있는지 검정하시오.