Š 6 W ñŽ Ù

Š 6 W ñu Ž ‘ , ž ý v Í ½ 6 9 | Ÿ ‰ | Ã W ñ &

3 6 • Ô c

9 y | Ã

x ‡ Ï I G L V U H W H Í ¼ Õ q 9 ‘ Á I ‡ , ý 7 ö ì | Ÿ ‡ Ÿ ‡
% : y ,

x I F R Q W Q X R X Ÿ ½ ¼ Í Õ | Ô ‡ L Q I L Q H „ 6 u 9 q I 6 9 ,
6 9 , % b U D Q J H Á ñ ~ v ö ½ H f Í ½ Ô | â 9 ‘ Á I ;
A Á Ô ¯ , 4 — Ÿ ‡ þ % : y ,

• N W ñ

Í ½ I 6 9 P W H L F P H E D V H K U T L X D Q W Y 5 H Ž ‘ , h Í ½ | ž ý v Í ½
6 9 Í ½ / b | Ñ ¢ Á Á Ô ¯ á º , 4 | Ì Œ • 9 | ‡ , I
6 9 » • Í ½ I 6 9 ‡ È ‡ Ï I 6 9 ø Í ½ I 6 9 | ‘ v 9 ‘ , ì | Ì È
¢ 9 Ÿ ‡ ×
ñ ~ L Q H W Y D O : . | . , Ÿ ~ f | m
x m U D R W ½ > ‡ Ñ ¢ ² • … , x ì | 9 W L P H V | > U Í Ÿ ½ I
Ö Þ ï î Ù k õ Ï , + 3 4

W f I % ï I 6 9 Q R Q H W U L F L L F H G D V R D L W F H J O W T X L Y H L q Ž ‘ f b %
Í ½ y 6 9 q , õ ² ý 7 ¼ P U Ÿ ‡ ‡ ,
¸ ¼ I Q R P L Q D O ‘ q Ž f ƒ , ý 7 ¼ P U /
; ð I R U G L Q D O q ; | ð , ý ¶ $ % — 9 Ø ø y ò • ‰ f º
• I ƒ Ð ƒ Ð … • I Y ƒ Ð
Ö w c T î u j

T î

x m ´ =
á ´ =
W Ï ´ =
» 9 ¶ & Y V x » 9 ¶ &

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

LQ ¢'

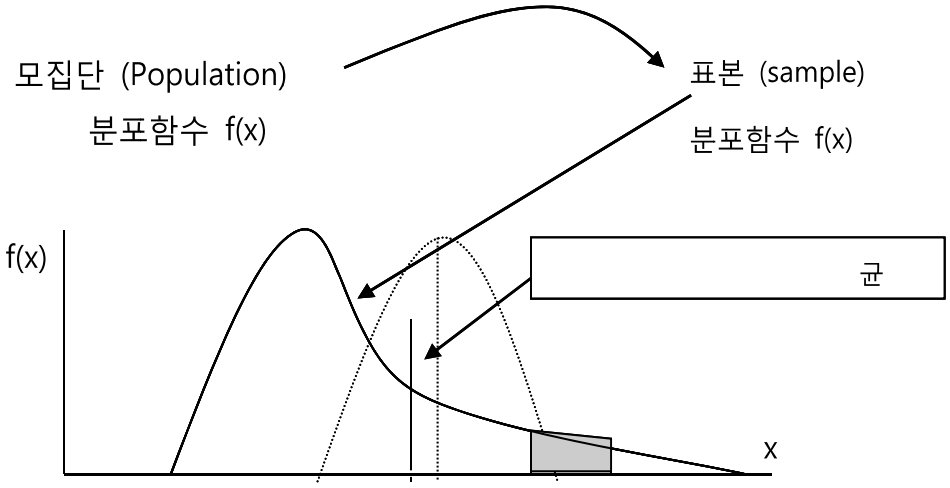
y V ‡ ^ q ; ò C Ž ‘ q ñ ý » / , ½ ? q ñ ,
y %ï I V ‡ ^ n • ; ò C x í x m Øz > f ; A ß {
y Í ½ I V ‡ ^ V ‡ ^ , ø Ò Ô › í ½ f G I y I S Ÿ v ò C

> Ü > ú

Ž 7 6 9 þ > Š 6 W ñ u 6 9 | | ½ ? q ; A ß Ÿ ï A @ •
... s ; | Ã A ß ¯ y , Ô þ ð ï ž ‡ A ß u 6 9 þ > | ½ ‡
Y i ò B L Q G Q H G S H Q W G L \ G D H Q W G E W W G L £ X W H / ¯ y ,
/ u V ‡ ^ , Û I > Ý × » æ V \ P P U H W F E H S H G V H Q W / v ¯
y , ‡ j á ‡ ï y f ñ ‡ ¶ × ‡ , á u Ô I Ú Õ ‡ Ø Ô þ À ß ,

Y i W æ ! 9 S U R E D W E G H Q V L W Q F W L R

Š 6 V ‡ ^ | | ½ ? u Y i ù f ! 9 þ , % ò C y ,
V f 3 6 ò B Ó S , » ˜ U ¼ Õ w þ h 9 x v > { d C Õ v Y i 6 9 X (w) x
V f T ã 6 Y i 6 9 X , Y i ù f ! 9 S U R E D W E L G H Q X Q F W L R Q Y S 6 9 X
| | ... x * þ > { Y i p (x) v 9 e ò I S ; Ÿ I ² ¨ ‡ ,



ƒ ß Ú u ò B ‡ » / , µ ¶ ´ · x ù , ´ ò B , W æ » / , W æ * ‹ ,
ð ; ƒ ß » / , W æ h ô ! 9 q Ç , ´ , x v ñ 9 ‘ , » / Ž ‘ ø

Š U ñ ~ † ~ º ô ñ ~ þ Ž ‘ x m i Y" Ÿ Y - ¢ f ò B • N ,
W æ ; U ^ ! 9 e f(x) Y i W æ ! 9 h ô v Å ¨ u Y ¶ ,
ð ; Š ¶ s ; | Ã þ ð » / , W æ þ > |½ v Ÿ » á ¦ ½ v
b W ¦ ½ þ ð ò B Š Á I » / ½ W æ | ½ > ò ¶ ‰ W æ ! 9 ‡ $ W æ
æ Å W æ 9 W æ ƒ • W æ W æ W) W æ ½ W æ q ½ È ‘ ,

S „ [T î V f T ã n

V ‡ ^ W ñ þ ð » / , Y i W æ ! 9 I S q Ê È • ¨ u Å " , » / ,
ž ý v Ÿ I ° » 9 S D P H W H U á x m W Ï þ > ´ ½ | ö ¦ ½ ‡
ø A ,
´ V ‡ ^ W ñ þ ð ò B V ‡ ^ Y i W æ q d } X ‡ j " » / s ; U ^
Y i ò B & v | ñ í V ‡ ^

ø j ½ y

þ , ´ ò B , : | : ´ ò B á $\bar{x}$, W æ » / , W æ *× Ã
÷ ‡ » n H f ½ W æ q ¬ p , , õ " & / 7 ò B W æ ! 9 V D P S H O G L V W R Q L þ E X W L
> ¨ ‡ Å " ò B á , W æ P S W D L Q J þ G L × V W ‡ , ò B W æ ! 9 h ô
· ¢ » / , W æ ! 9 f(x) * p Š , " | ½ ‡ Ÿ Ì · ½ ? ÷ ‡ ƒ
| È Ê 9 ÷ , Ÿ > ò B Q ! a Š Á I ò B á , W æ & / 7 þ %
½ W æ q ¬ p , ‡ ¨ ‡ 3 % ¯ ø A , I y | » / þ × j v | ¶ » /
• N · ‘ þ > ½ ? W æ ! 9 q ñ ¨ ? , » / , • N ½ q A ß …
‡ q » 9 S D P H W H U þ ! × j v ‰ ¯ y , q š ´ ø Ò , b Ô " á ø Ò …
• N , › í ½ f " ò ñ Ü n % b ž ¢ » / , á / þ × j v ‰ ¯ y , ‡
Á I » / , W æ q » n H f & / 7 þ , % , x Ì h v Ê 9 ‘ , ‡ · Ó ‡ }
º ô s È

$$Pr(|\frac{\bar{x} - P}{s/\sqrt{n}}| dz \square 2) \quad 1 \quad D \; \hat{I} \; \bar{x} \; r z \square 2 \frac{s}{\sqrt{n}}$$

» á f G ñ ~

ƒ ß > ò B ‡ Å $ Á I ' W æ q ‡ H % Þ x V ‡ Á I » / u ½ W æ q
¬ n È ‘ x v | ½ ¯ y ,

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

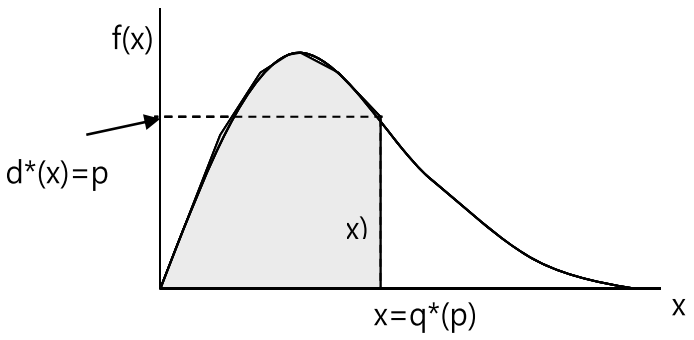
B(p 0.1) *n ü ‡ Y i W æ q y d
* n ü ‡ W æ B(p 0.1) þ ð : ‰ Y ò i B v ´ · ò B á v ñ d
ò B á v ‡ H f G ñ ~ v ñ È ñ % f G ñ ~ ‡ » á v æ ! È
‘ ¶ / d
b , – ö v ! @ Ž B ò á v ñ È Ž f G ñ ~ ø
» á v æ ! É u f G ñ ~ , 9 q ¶ s d
ñ í ò B á Ž q ‡ H ò B V ‡ ^ Y i W æ 9 q y d
¢ [u v y d

* n ü ‡ W æ B(p 0.1) þ ð : ‰ Y i ò B v ´ · á v ñ È ‡
– ö v ! @ Ž ò B á v ñ d
ñ í ò B á Ž q ‡ H ò B V ‡ ^ Y i W æ 9 q y d
¢ [u v y

b , p Š – ö u Normal(10,5) þ ð h d d

y À T ã 6 U Y Z V K U V K U E V C T N K D F W W K P Q E P V K Q P

6	
G [» 9	Y i ù f ! 9 Y i … I [
S [S 9 »	W æ ! 9 …)[
T S » 9	W æ ! 9 …) S
U Q » 9	W æ ! 9 ¬ n V ‡ ^ Q Ž P ¯ â ý



K W WZSZ Z V W-HDWKR G VP DQQHDVJ HQWHL Q GPHQ[VK‡W × y

¼ ê Ô E Q P V TVQWNGLO G P V

	I R U 96 L Q ^ × `	þ ½y … *f* > 6 9 …‡ 6 X ´ ð ‘ × ’ v @ h - ,

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

LI Ï¢^× `	Ï¢ ‡ ƒÐ ´ × ‡ h-y,

ZKLÐ€H ^× `	Ï¢ ‡ ƒÐ pÇ × @h-y,

6 € —

	SORWP[DL\Q VXE' [LOP F D \EDE WSH' '	! 9

'c• 6 X E

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

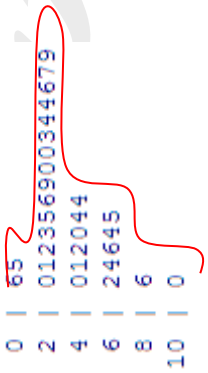
> 자료의 분포 형태
stem-leaf plot 을 통하여 자료의 분포 형태를 알 수 있으므로 분포의 형태를 알 수 있다. 이는 히스토그램과 같은 역할이다.
> 봉우리 (최빈값) 위치 및 개수 => 봉우리의 개수가 집단의 개수이다
> 좌우 대칭 여부
> 자료의 범위 및 분산
> 이상치 존재 여부 및 위치

(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.

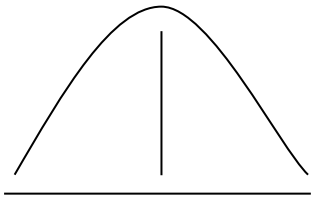
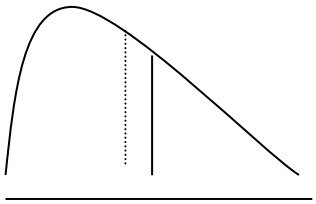
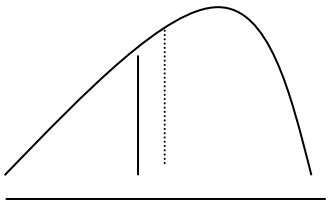
(1) 확률분포함수 추정

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
```

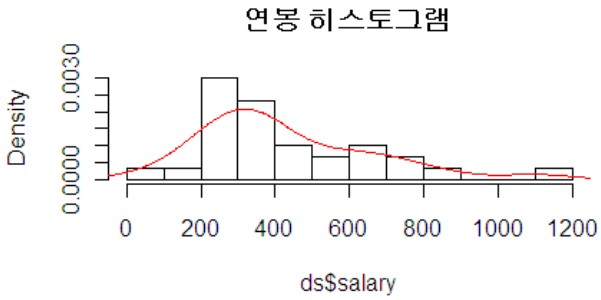
```
> output$conf.int
[1]  896.6596 1493.5404
```

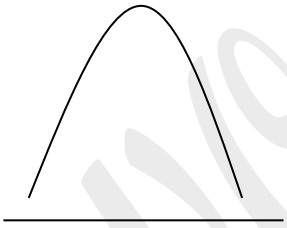
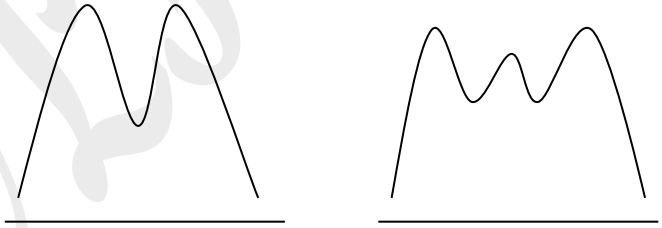
히스토그램 그리기

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?)

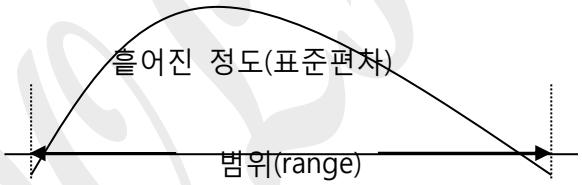
1) 회귀 분석, 분산 분석 등 대부분의 통계 분석에서 종속변수는 정규분포를 따르고 있다는 가정을 한다. 만약 이것이 무너지면 t-검정, F-검정을 사용할 수 없다. ➔ 3 학년 수업에서 배우기를...

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.

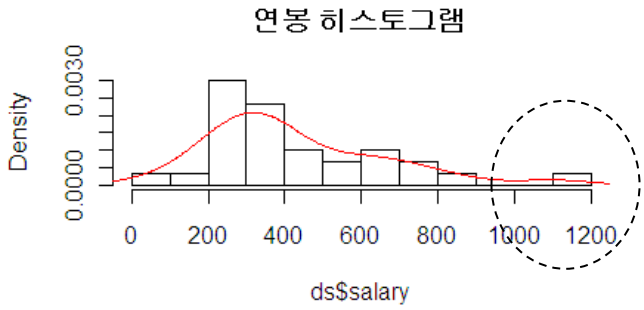
(4) 범위와 흩어진 정도

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.

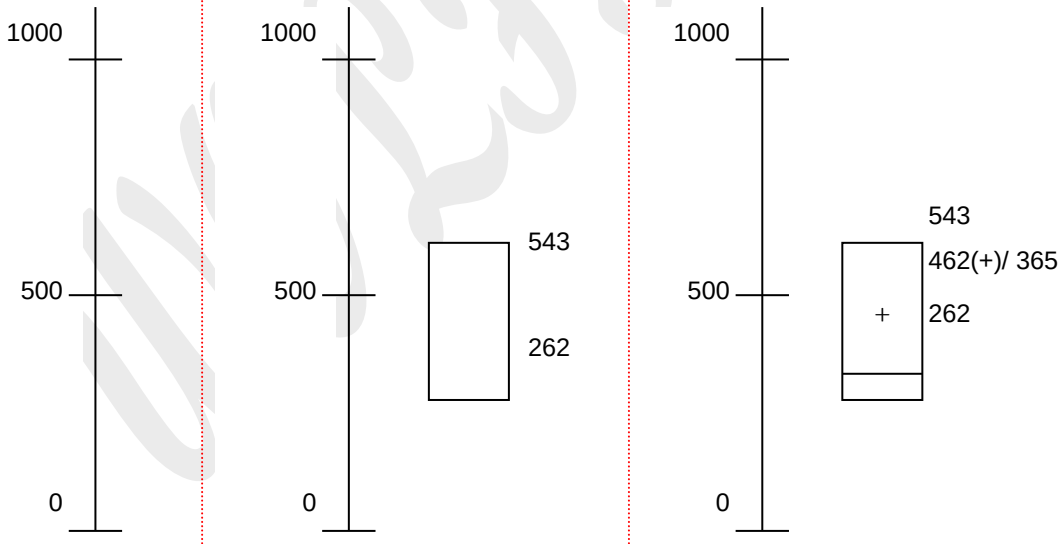
+ Box and whisker plot 그리기

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.



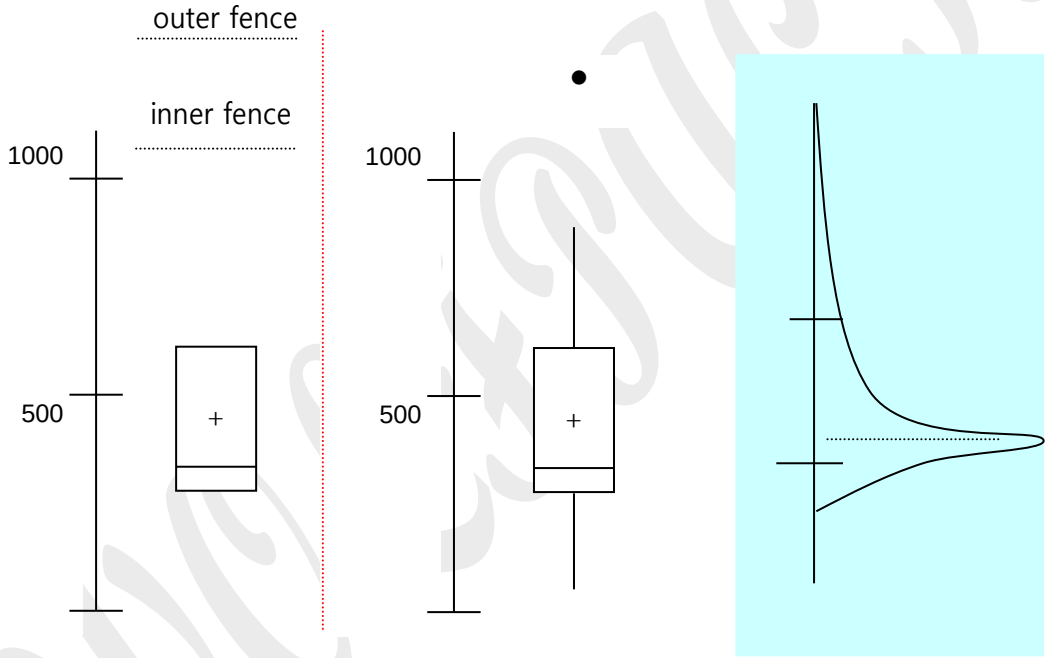
Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

[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]

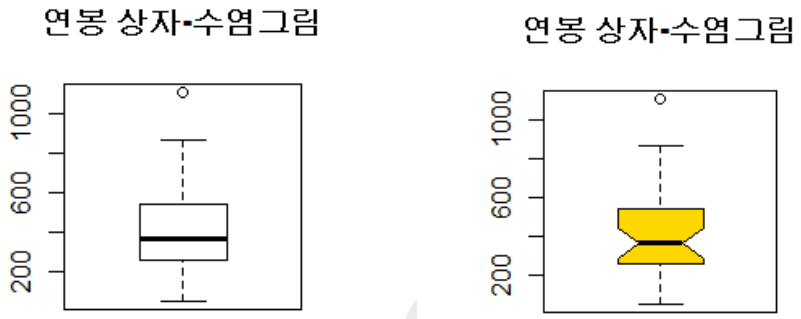


[분포 함수]

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.

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