03

데이터분석방법론(1)

Probability and Distributions

통계·데이터과학과 장영재 교수

한국방송통신대학교 대학원

학습목차

01

Basic concepts

1. Random Sampling

```
> # sampling without replacement
> sample(1:40,5)
[1] 11  2 39 15 12
> # sampling with replacement
> sample(c("H","T"), 10, replace=T)
 [1] "T" "H" "H" "H" "H" "H" "H" "T" "H" "T«
> # sampling with prob (*)
> sample(c("succ", "fail"), 10, replace=T, prob=c(0.9, 0.1))
 [1] "succ" "succ" "succ" "fail" "succ" "fail" "succ" "succ" "fail" "succ«
```

※ (*) This may not be the best way to generate such a sample, though.
See the later discussion of the binomial distribution.

2. Prob. Calculations and Combinatorics

◆ 순열 및 조합

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} \qquad \binom{n}{r} \text{로도 표현} $$

◆ The probability to choose 5 numbers out of 40

```
> prod(5:1)/prod(40:36)
[1] 1.519738e-06
> 1/choose(40,5)
[1] 1.519738e-06
```

3. Discrete distributions

binomial distributions

$$X \sim b(n, p)$$

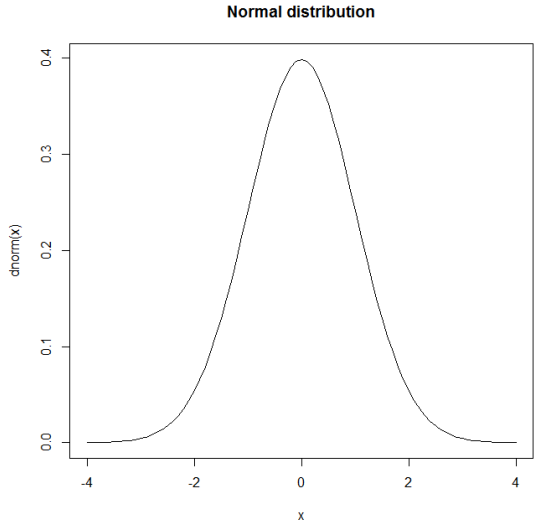
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

4. Continuous distributions

◆ Normal distributions

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



5. The built-in distributions in R

- Density or point probability
- Cumulated probability, distribution function
- Quantiles
- Pseudo-random numbers

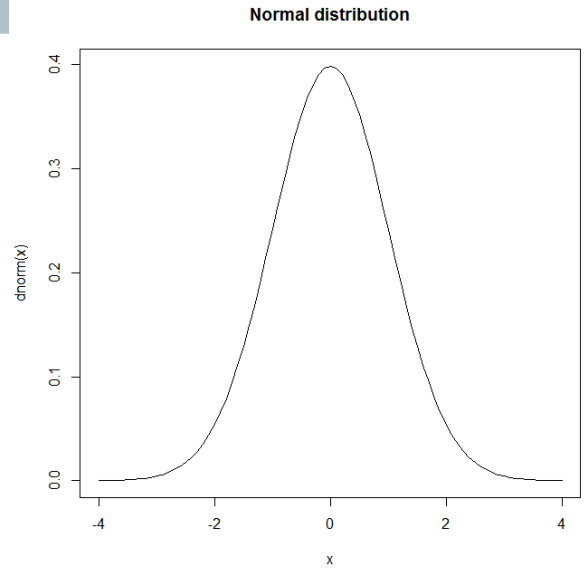
※ For the normal distribution, these are named dnorm, pnorm, qnorm, and rnorm (density, probability, quantile, and random, respectively).

5. The built-in distributions in R

Densities

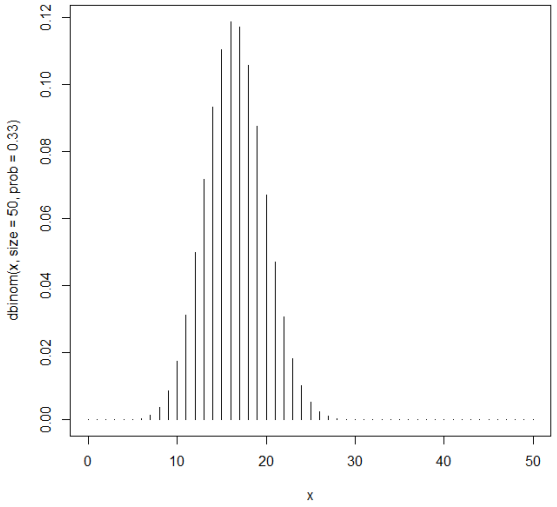
- **Normal distribution**

```
> x <- seq(-4,4,0.1)
> plot(x,dnorm(x),type="l")
> title("Normal distribution")
> # 다른 방법
> curve(dnorm(x), from=-4, to=4)
```



- **Binomial distribution**

```
> x <- 0:50
> plot(x,dbinom(x,size=50,prob=.33),type="h")
```



5. The built-in distributions in R

Cumulative distribution functions

- **Normal distribution**

$$X \sim N(132, 13^2)$$

Calculate $\Pr(X \geq 160)$

```
> 1-pnorm(160,mean=132,sd=13)
[1] 0.01562612
```

- **Binomial distribution**

Twenty patients are given two treatments each and then asked whether treatment A or B worked better. It turned out that 16 patients liked A better. The question is then whether this can be taken as sufficient evidence that A actually is the better treatment or whether the outcome might as well have happened by chance even if the treatments were equally good.

$$X \sim b(20, 0.5)$$

$$H_o : p = 0.5$$

$$H_1 : p > 0.5$$

$$p - value = P(X \geq 16 \mid H_0) = 1 - P(X \leq 15)$$

```
> 1-pbinom(15,size=20,prob=.5)
[1] 0.005908966
```

5. The built-in distributions in R

Quantiles

- The quantile function is the inverse of the cumulative distribution function. The p-quantile is the value with the property that there is probability p of getting a value less than or equal to it. The median is by definition the 50% quantile.

- 95% CI for μ

$$\bar{x} + \sigma/\sqrt{n} \times N_{0.025} \leq \mu \leq \bar{x} + \sigma/\sqrt{n} \times N_{0.975}$$

where $N_{0.025}$ is the 2.5% quantile in the normal distribution.

- 예)

```
> xbar <- 83
> sigma <- 12
> n <- 5
> sem <- sigma/sqrt(n)
> sem
[1] 5.366563
> xbar + sem * qnorm(0.025)
[1] 72.48173
> xbar + sem * qnorm(0.975)
[1] 93.51827
```

5. The built-in distributions in R

Random numbers

```
> rnorm(10)
 [1] -0.20507818 -0.09966328  0.65998810  1.72411951  1.19912241
1.30441056
 [7] -0.86466874  0.52546963  0.26023707 -0.32000708
> rnorm(10)
 [1] -0.8474839  0.2838526  1.5376718  0.1761081  1.3979563
0.6645849
 [7] -0.9653596  1.1003153  0.3086113 -0.1949667
> rnorm(10,mean=7,sd=5)
 [1]  2.776367  3.273605 15.495370  6.433699  8.942600 14.482943
6.907142
 [8] 15.290887 10.097125  5.686843
> rbinom(10,size=20,prob=.5)
 [1] 11 11  6  9 12  9  8  6  8  7
```

02

Descriptive Statistics and Graphics

1. Summary statistics for a single group

- Calculate the mean, standard deviation, variance, and median

```
> x <- rnorm(50)
> mean(x)
[1] -0.1845058
> sd(x)
[1] 0.859216
> var(x)
[1] 0.7382521
> median(x)
[1] -0.08858254
> quantile(x)
       0%        25%        50%        75%       100%
-2.10287452 -0.70814802 -0.08858254  0.33952586  2.09561948
> pvec <- seq(0,1,0.1)
> quantile(x, pvec)   # decile
       0%        10%        20%        30%        40%        50%
-2.10287452 -1.18149282 -0.88538878 -0.67435121 -0.35960028 -0.08858254
      60%        70%        80%        90%       100%
 0.03413394  0.19233202  0.46423920  0.82471082  2.09561948
```

1. Summary statistics for a single group

- In case there are missing values in data (1)

```
> library(ISwR)
> data(juul)
> head(juul,3)
  age menarche sex igf1 tanner testvol
1  NA       NA  NA   90     NA      NA
2  NA       NA  NA   88     NA      NA
3  NA       NA  NA  164     NA      NA
> attach(juul)
> mean(igf1)
[1] NA
> mean(igf1, na.rm=T)
[1] 340.168
> sum(!is.na(igf1))
[1] 1018
> summary(igf1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   25.0   202.2   313.5   340.2   462.8   915.0     321
```

1. Summary statistics for a single group

- In case there are missing values in data (2)

```
> summary(juul)
      age              menarche          sex             igf1            tanner          testvol      
 Min.   : 0.170   Min.   :1.000   Min.   :1.000   Min.   : 25.0   Min.   :1.00   Min.   : 1.000  
 1st Qu.: 9.053   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:202.2   1st Qu.:1.00   1st Qu.: 1.000  
 Median :12.560   Median :1.000   Median :2.000   Median :313.5   Median :2.00   Median : 3.000  
 Mean   :15.095   Mean   :1.476   Mean   :1.534   Mean   :340.2   Mean   :2.64   Mean   : 7.896  
 3rd Qu.:16.855   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:462.8   3rd Qu.:5.00   3rd Qu.:15.000  
 Max.   :83.000   Max.   :2.000   Max.   :2.000   Max.   :915.0   Max.   :5.00   Max.   :30.000  
 NA's   :5        NA's   :635     NA's   :5       NA's   :321     NA's   :240    NA's   :859     
> detach(juul)
> juul$sex <- factor(juul$sex,labels=c("M","F"))
> juul$menarche <- factor(juul$menarche,labels=c("No","Yes"))
> juul$tanner <- factor(juul$tanner,  labels=c("I","II","III","IV","V"))
> attach(juul)
```

1. Summary statistics for a single group

- In case there are missing values in data (3)

```
> summary(juul)
      age            menarche      sex            igf1          tanner      testvol
 Min.   : 0.170   No  :369   M   :621   Min.   : 25.0   I   :515   Min.   : 1.000
 1st Qu.: 9.053   Yes :335   F   :713   1st Qu.:202.2   II  :103   1st Qu.: 1.000
 Median :12.560   NA's:635   NA's:  5   Median :313.5   III : 72   Median : 3.000
 Mean   :15.095                         Mean   :340.2   IV  : 81   Mean   : 7.896
 3rd Qu.:16.855                         3rd Qu.:462.8   V   :328   3rd Qu.:15.000
 Max.   :83.000                         Max.   :915.0   NA's:240   Max.   :30.000
 NA's   :5                              NA's   :321                NA's   :859
> # use transform
> juu2 <- transform(juul,
+         sex=factor(sex,labels=c("M","F")),
+         menarche=factor(menarche,labels=c("No","Yes")),
+         tanner=factor(tanner,labels=c("I","II","III","IV","V")))
```

2. Graphics Display of Distributions

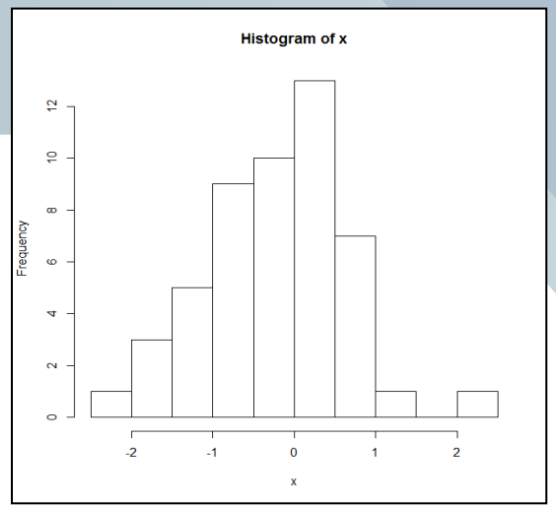
Histograms

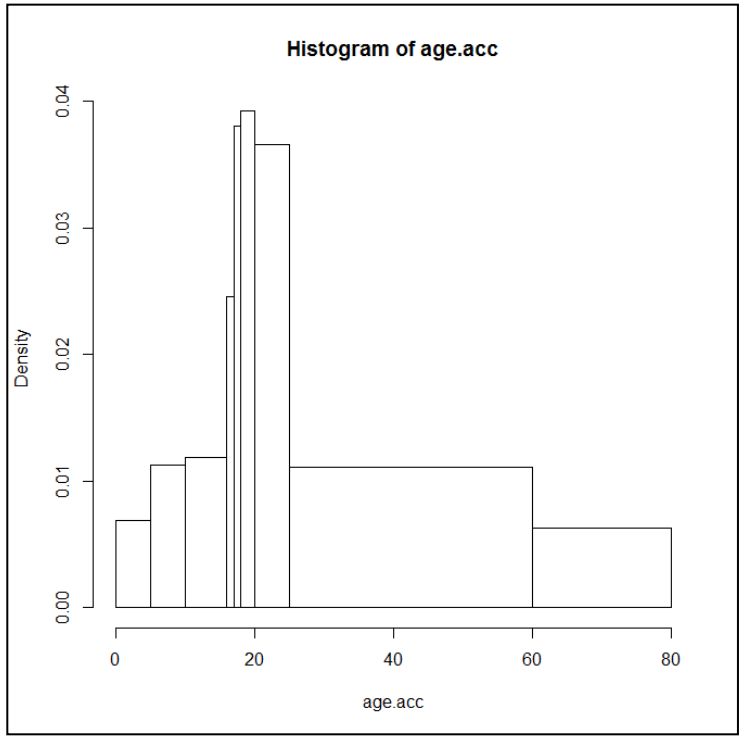
```
> hist(x)
```

- Histogram using breaks=n

ex) Altman(1991) :accident rates by age group are given as a count in age groups 0-4, 5-9, 10-15, 16, 17, 18-19, 20-24, 25-59, and 60-79 years of age

```
> mid.age <-
c(2.5,7.5,13,16.5,17.5,19,22.5,44.5,70.5)
> acc.count <-
c(28,46,58,20,31,64,149,316,103)
> age.acc <- rep(mid.age,acc.count)
> brk <- c(0,5,10,16,17,18,20,25,60,80)
> hist(age.acc,breaks=brk)
```





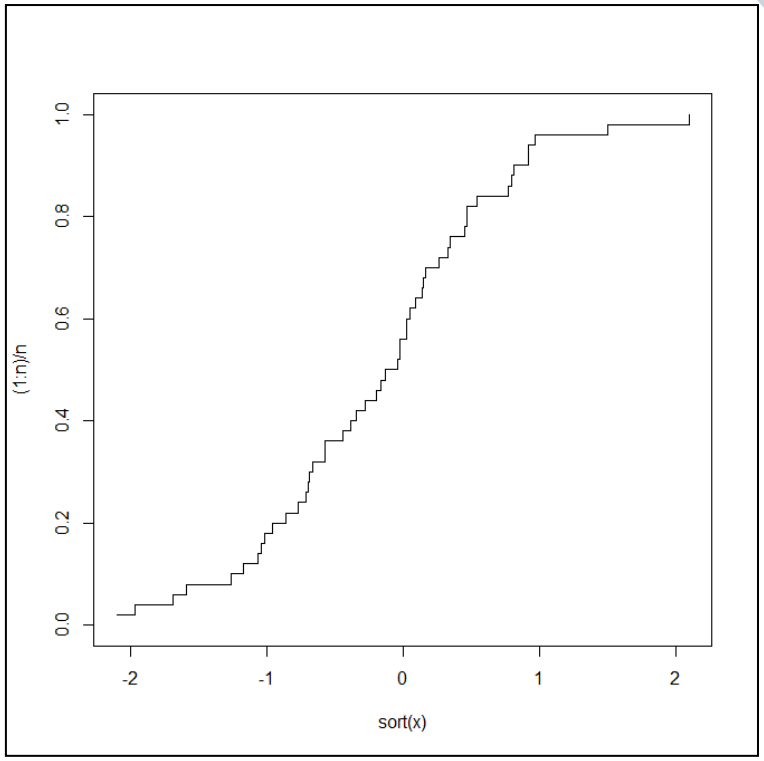
2. Graphics Display of Distributions

Empirical cumulative distribution

Definition :

$$\hat{F}_n(x) = \frac{number\ of\ elements\ \text{in}\ the\ sample \le x}{n} = \frac{\sum_{i=1}^{n} I(X_i \le x)}{n}$$

```
> n <- length(x)
> plot(sort(x),(1:n)/n,type="s",ylim=c(0,1))
```

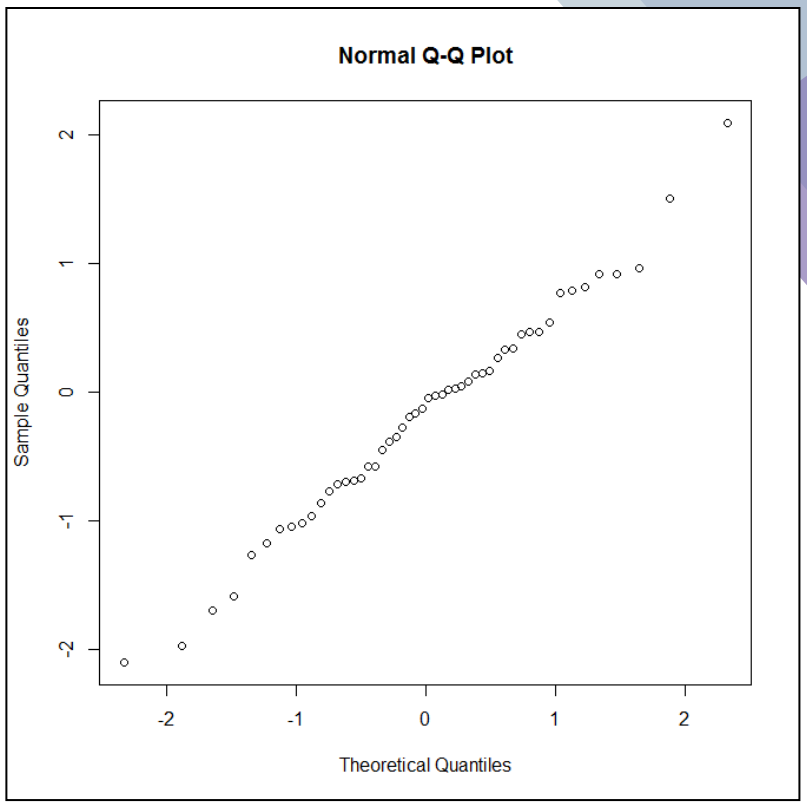


2. Graphics Display of Distributions

Q-Q plots (quantile versus quantile plots)

- One purpose of calculating the empirical cumulative distribution function (c.d.f.) is to see whether data can be assumed normally distributed. For a better assessment, you might plot the kth smallest observation against the expected value of the kth smallest observation out of n in a standard normal distribution. The point is that in this way you would expect to obtain a straight line if data come from a normal distribution with any mean and standard deviation.

```
> qqnorm(x)
```

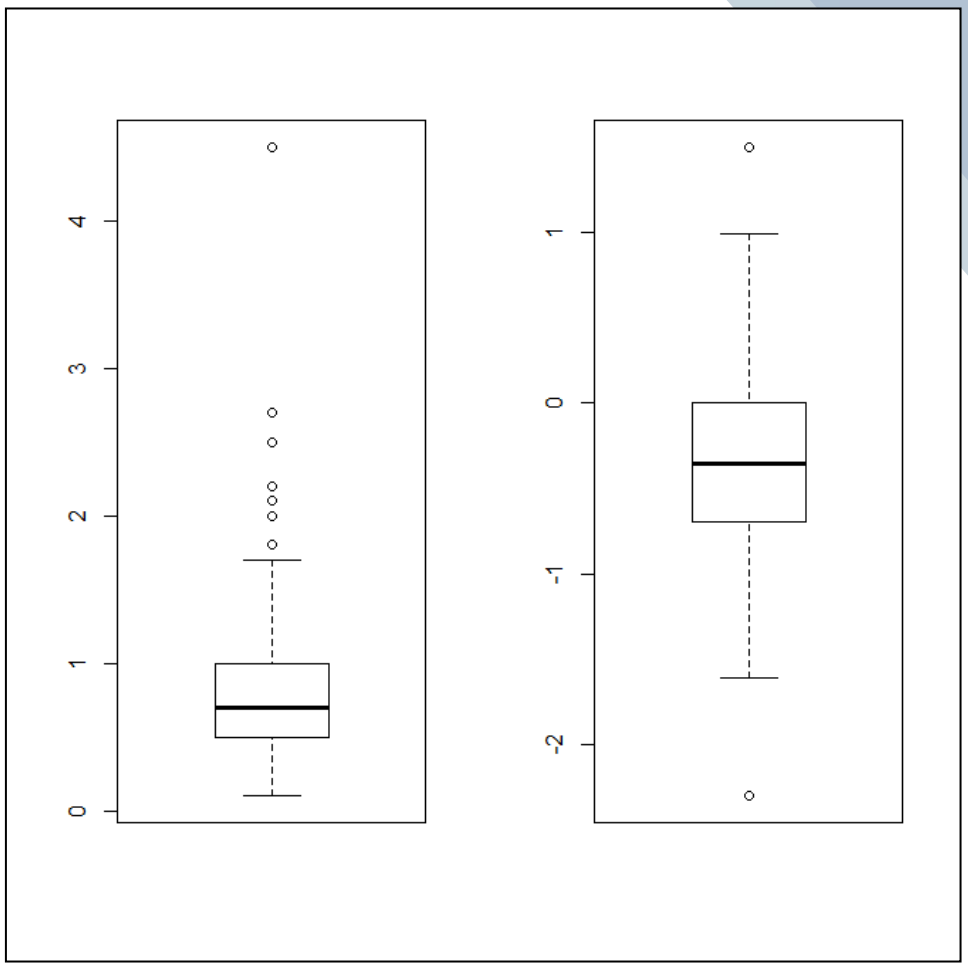


2. Graphics Display of Distributions

Boxplots

- A “boxplot”, or more descriptively a “box-and-whiskers plot”, is a graphical summary of a distribution.

```
> library(ISwR)
> par(mfrow=c(1,2))
> boxplot(IgM)
> boxplot(log(IgM))
> par(mfrow=c(1,1))
```



3. Summary statistics by groups

- When dealing with grouped data, you will often want to have various summary statistics computed within groups; for example, a table of means and standard deviations.

```
> attach(red.cell.folate)
> head(red.cell.folate, 3)
  folate ventilation
1    243  N2O+O2,24h
2    251  N2O+O2,24h
3    275  N2O+O2,24h
> tapply(folate,ventilation,mean)
N2O+O2,24h  N2O+O2,op     O2,24h
  316.6250   256.4444   278.0000
> tapply(folate,ventilation,sd)
N2O+O2,24h  N2O+O2,op     O2,24h
  58.71709   37.12180   33.75648
> tapply(folate,ventilation,length)
N2O+O2,24h  N2O+O2,op     O2,24h
         8          9          5
```

```
> xbar <- tapply(folate, ventilation, mean)
> s <- tapply(folate, ventilation, sd)
> n <- tapply(folate, ventilation, length)
> cbind(mean=xbar, std.dev=s, n=n)
                mean   std.dev n
N2O+O2,24h 316.6250 58.71709 8
N2O+O2,op  256.4444 37.12180 9
O2,24h     278.0000 33.75648 5
```

3. Summary statistics by groups

- For the juul data

```
> tapply(igf1, tanner, mean)
 1  2  3  4  5
NA NA NA NA NA
> tapply(igf1, tanner, mean, na.rm=T)
       1        2        3        4        5
207.4727 352.6714 483.2222 513.0172 465.3344
```

- The functions aggregate and by are variations on the same topic.

```
> aggregate(juul[c("age","igf1")],
+                 list(sex=juul$sex), mean, na.rm=T)
  sex      age      igf1
1   1 15.38436 310.8866
2   2 14.84363 368.1006
```

3. Summary statistics by groups

- Using by function, you can for instance summarize the Juul data by sex as follows:

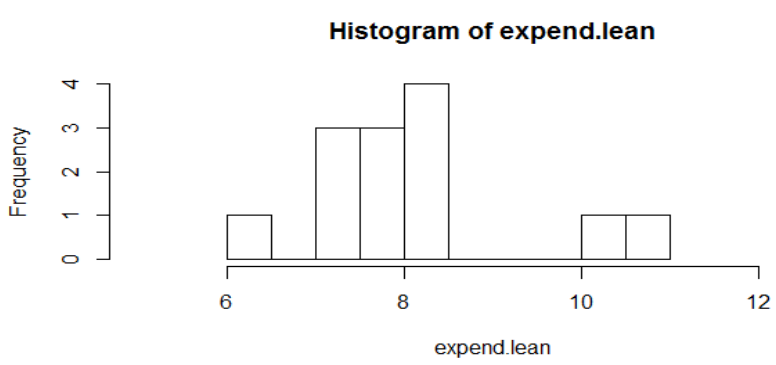
```
> by(juul, juul["sex"], summary)
sex: 1
      age            menarche          sex          igf1            tanner          testvol      
 Min.   : 0.17   Min.   : NA   Min.   :1   Min.   : 29.0   Min.   :1.000   Min.   : 1.000  
 1st Qu.: 8.85   1st Qu.: NA   1st Qu.:1   1st Qu.:176.0   1st Qu.:1.000   1st Qu.: 1.000  
 Median :12.38   Median : NA   Median :1   Median :280.0   Median :1.000   Median : 3.000  
 Mean   :15.38   Mean   :NaN   Mean   :1   Mean   :310.9   Mean   :2.361   Mean   : 7.896  
 3rd Qu.:16.77   3rd Qu.: NA   3rd Qu.:1   3rd Qu.:430.2   3rd Qu.:4.000   3rd Qu.:15.000  
 Max.   :83.00   Max.   : NA   Max.   :1   Max.   :915.0   Max.   :5.000   Max.   :30.000  
                 NA's   :621               NA's   :145     NA's   :76      NA's   :141     
------------------------------------------------------------------------------------------
sex: 2
      age            menarche          sex          igf1            tanner          testvol      
 Min.   : 0.25   Min.   :1.000   Min.   :2   Min.   : 25.0   Min.   :1.000   Min.   : NA  
 1st Qu.: 9.30   1st Qu.:1.000   1st Qu.:2   1st Qu.:233.0   1st Qu.:1.000   1st Qu.: NA  
 Median :12.80   Median :1.000   Median :2   Median :352.0   Median :3.000   Median : NA  
 Mean   :14.84   Mean   :1.476   Mean   :2   Mean   :368.1   Mean   :2.913   Mean   :NaN  
 3rd Qu.:16.93   3rd Qu.:2.000   3rd Qu.:2   3rd Qu.:483.0   3rd Qu.:5.000   3rd Qu.: NA  
 Max.   :75.12   Max.   :2.000   Max.   :2   Max.   :914.0   Max.   :5.000   Max.   : NA  
                 NA's   :9                   NA's   :176     NA's   :159     NA's   :713  
```

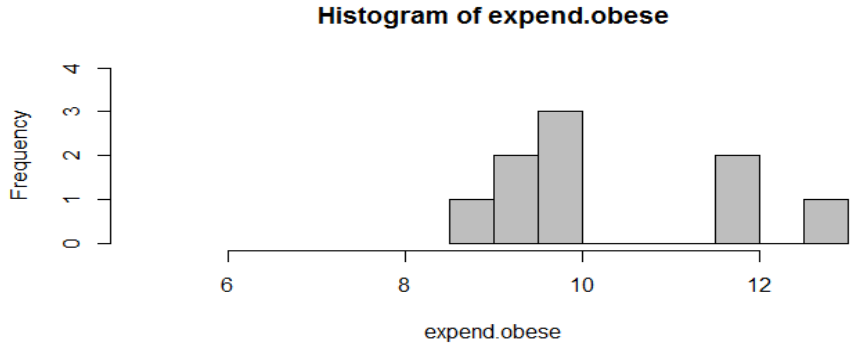
4. Graphics for grouped data

- In dealing with grouped data, it is important to be able not only to create plots for each group but also to compare the plots between groups.

Histograms

```
> attach(energy)
> expend.lean <- expend[stature=="lean"]
> expend.obese <- expend[stature=="obese"]
> par(mfrow=c(2,1))
> hist(expend.lean,breaks=10,xlim=c(5,13),ylim=c(0,4),col="white")
> hist(expend.obese,breaks=10,xlim=c(5,13),ylim=c(0,4),col="grey")
> par(mfrow=c(1,1))
```



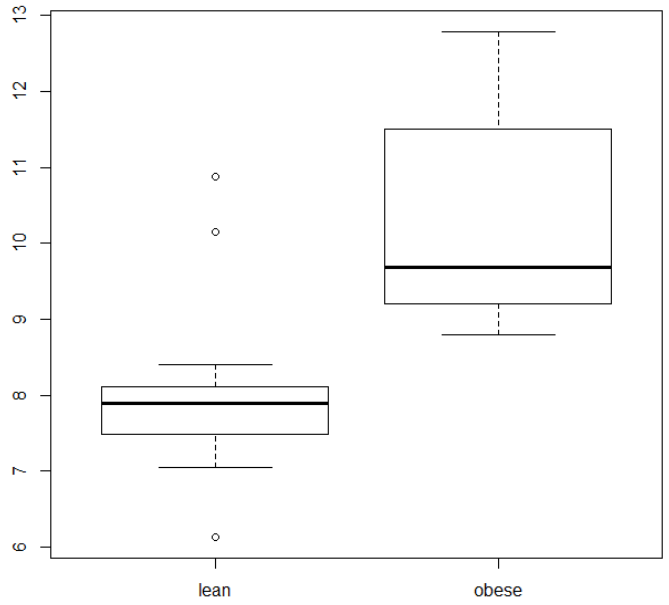


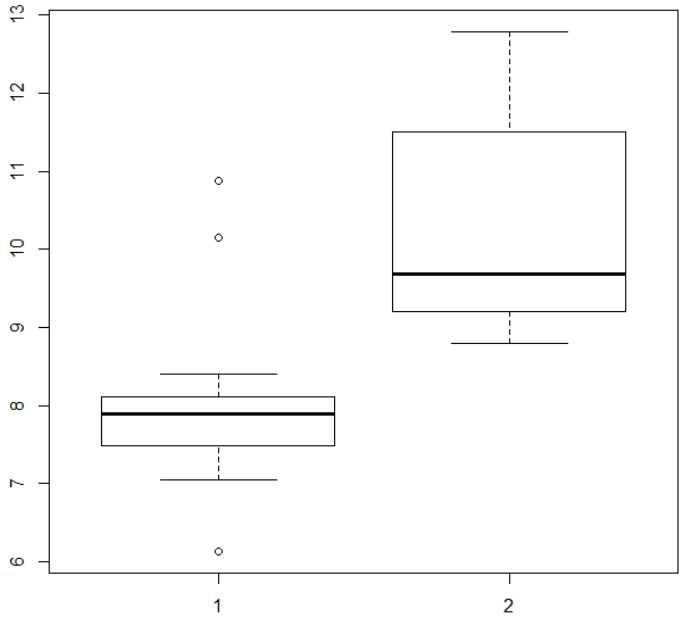
4. Graphics for grouped data

Parallel boxplots

- You might want a set of boxplots from several groups in the same frame.

```
> boxplot(expend ~ stature)  #  1
> boxplot(expend.lean,expend.obese)  # 2
```



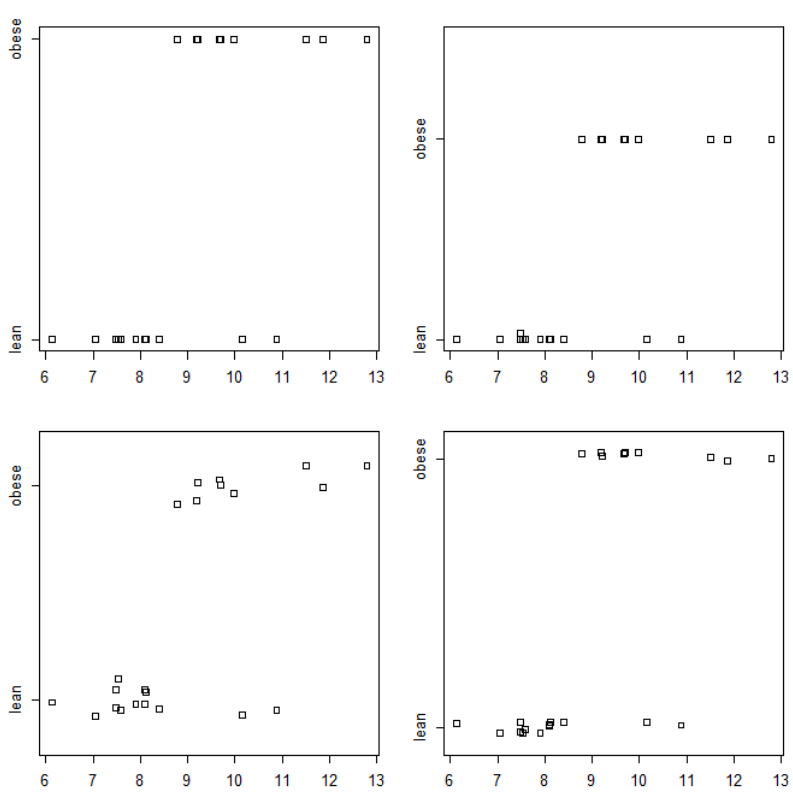


4. Graphics for grouped data

Stripcharts

- With groups as small as these, the quartiles will be quite inaccurately determined, and it may therefore be more desirable to plot the raw data.

```
> opar <- par(mfrow=c(2,2), mex=0.8, mar=c(3,3,2,1)+.1)
> stripchart(expend ~ stature)
> stripchart(expend ~ stature, method="stack")
> stripchart(expend ~ stature, method="jitter")
> stripchart(expend ~ stature, method="jitter", jitter=.03)
> par(opar)
```



다음시간 안내

04

One-sample and Two-sample test