
R Recipes 데이터 분석

강사 : 문성민

NA(Missing Value) Handling

NA Handling

- NA(Not available)

- 값이 누락되거나 값이 없는 값을 나타내는 문자

- 예제1

- 변수 생성

```
> X<-c(1,2,3,4,5,6,7,8,NA)
> X
[1] 1 2 3 4 5 6 7 8 NA
```

- NA값으로 변환

```
> X[X==2]<-NA
> X
[1] 1 NA 3 4 5 6 7 8 NA
```

- 변수 요약

```
> summary(X)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's 
  1.00   2.75   4.50   4.50   6.25   8.00     1
```

- 변수 연산하기

```
> sum(X)
[1] NA
> mean(X)
[1] NA
> sum(X,na.rm=T)
[1] 34
> mean(X,na.rm=T)
[1] 4.857143
```

NA Handling

● 예제2

- 남녀간의 영어,수학 점수를 나타내는 데이터셋 생성

```
> Eng<-c(34,45,56,67,78,89,NA)
> Math<-c(98,NA,87,76,65,54,43)
> Gender<-c("M","F","M","F","M","M","M")
> Test<-data.frame(Eng=Eng,Math=Math,Gender,Gender)
> Test
```

	Eng	Math	Gender	Gender.1
1	34	98	M	M
2	45	NA	F	F
3	56	87	M	M
4	67	76	F	F
5	78	65	M	M
6	89	54	M	M
7	NA	43	M	M

- 데이터 확인

```
> str(Test)
'data.frame': 7 obs. of 4 variables:
 $ Eng      : num  34 45 56 67 78 89 NA
 $ Math     : num  98 NA 87 76 65 54 43
 $ Gender   : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2
 $ Gender.1 : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2
```

NA Handling

- NA를 포함한 행을 제거한 데이터 세트 생성

```
> na.omit(Test)
  Eng Math Gender Gender.1
1  34   98      M        M
3  56   87      M        M
4  67   76      F        F
5  78   65      M        M
6  89   54      M        M
```

- Test 데이터 요약

```
> summary(Test)
      Eng      Math      Gender
Min.   :34.00  Min.   :43.00  F:2
1st Qu.:47.75  1st Qu.:56.75  M:5
Median :61.50  Median :70.50
Mean   :61.50  Mean   :70.50
3rd Qu.:75.25  3rd Qu.:84.25
Max.   :89.00  Max.   :98.00
NA's   :1      NA's   :1
```

- 평균치 삽입법을 활용한 NA데이터 조작

```
> install.packages("gam")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/gam_1.09.1.tgz'을 시도합니다
Content type 'application/x-gzip' length 304040 bytes (296 Kb)
URL을 열었습니다
=====
downloaded 296 Kb

The downloaded binary packages are in
/var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmpv2osnU/downloaded_packages
> library(gam)
> na.gam.replace (Test)
  Eng Math Gender
1 34.0 98.0      M
2 45.0 70.5      F
3 56.0 87.0      M
4 67.0 76.0      F
5 78.0 65.0      M
6 89.0 54.0      M
7 61.5 43.0      M
```

NA Handling

- 영어와 수학 점수만으로 구성된 데이터 세트 생성

```
> Test2<-Test[,c("Eng","Math")]
> Test2
  Eng Math
1  34   98
2  45   NA
3  56   87
4  67   76
5  78   65
6  89   54
7  NA   43
```

- 데이터 세트 연산하기

```
> apply(Test2,2,mean)
  Eng Math
  NA   NA
> apply(Test2,2,mean,na.rm=TRUE)
  Eng Math
61.5 70.5
1
```

Outliers Handling

Outliers Handling

- Outliers

- 데이터 안에 존재하는 이상치로 데이터의 성질에 큰 영향을 미친다.

- 예제1

- 라이브러리 설치

```
> install.packages("outliers")
```

```
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/outliers_0.14.tgz'을 시도합니다
```

```
Content type 'application/x-gzip' length 50370 bytes (49 Kb)
```

```
URL을 열었습니다
```

```
=====
```

```
downloaded 49 Kb
```

```
The downloaded binary packages are in
```

```
  /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp8qGMiY/downloaded_packages
```

```
> library(outliers)
```

- 수치형 변수 추출

```
> Test_1=Test[,c(1,2)]
```

```
>
```


Outliers Handling

- 표준화 관련 수식(Formula)

$$Z_i = \frac{W_i - \bar{W}}{S_W} \quad \bar{W} = \frac{\sum_{i=1}^n W_i}{n} \quad S_W = \sqrt{\frac{\sum (W_i - \bar{W})^2}{n}}$$

- NA값에 평균치 삽입

```
> library(gam)
필요한 패키지를 로딩중입니다: splines
Loaded gam 1.09.1
```

```
>
> na.gam.replace (Test_1)
      Eng Math
1 34.0 98.0
2 45.0 70.5
3 56.0 87.0
4 67.0 76.0
5 78.0 65.0
6 89.0 54.0
7 61.5 43.0
```

- 데이터 표준화 시키기

```
> X <- scores(na.gam.replace (Test_1), type=c("z"))
>
> b=scale(na.gam.replace (Test_1))
>
> as.data.frame(b)
      Eng      Math
1 -1.4638501  1.4638501
2 -0.8783101  0.0000000
3 -0.2927700  0.8783101
4  0.2927700  0.2927700
5  0.8783101 -0.2927700
6  1.4638501 -0.8783101
7  0.0000000 -1.4638501
```

Outliers Handling

- 데이터 필터링 하기

```
> install.packages("dplyr")
```

```
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/dplyr_0.4.1.tgz'을 시도합니다
```

```
Content type 'application/x-gzip' length 3781115 bytes (3.6 Mb)
```

```
URL을 열었습니다
```

```
=====
downloaded 3.6 Mb
```

```
The downloaded binary packages are in
```

```
  /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp8qGMiY/downloaded_packages
```

```
>
```

```
> library(dplyr)
```

```
다음의 패키지를 부착합니다: 'dplyr'
```

```
The following object is masked from 'package:stats':
```

```
  filter
```

```
The following objects are masked from 'package:base':
```

```
  intersect, setdiff, setequal, union
```

```
>
```

```
> filter(X, Eng <= 1, Math <= 1)
```

	Eng	Math
1	-0.8783101	0.0000000
2	-0.2927700	0.8783101
3	0.2927700	0.2927700
4	0.8783101	-0.2927700
5	0.0000000	-1.4638501

Data Input & Output)

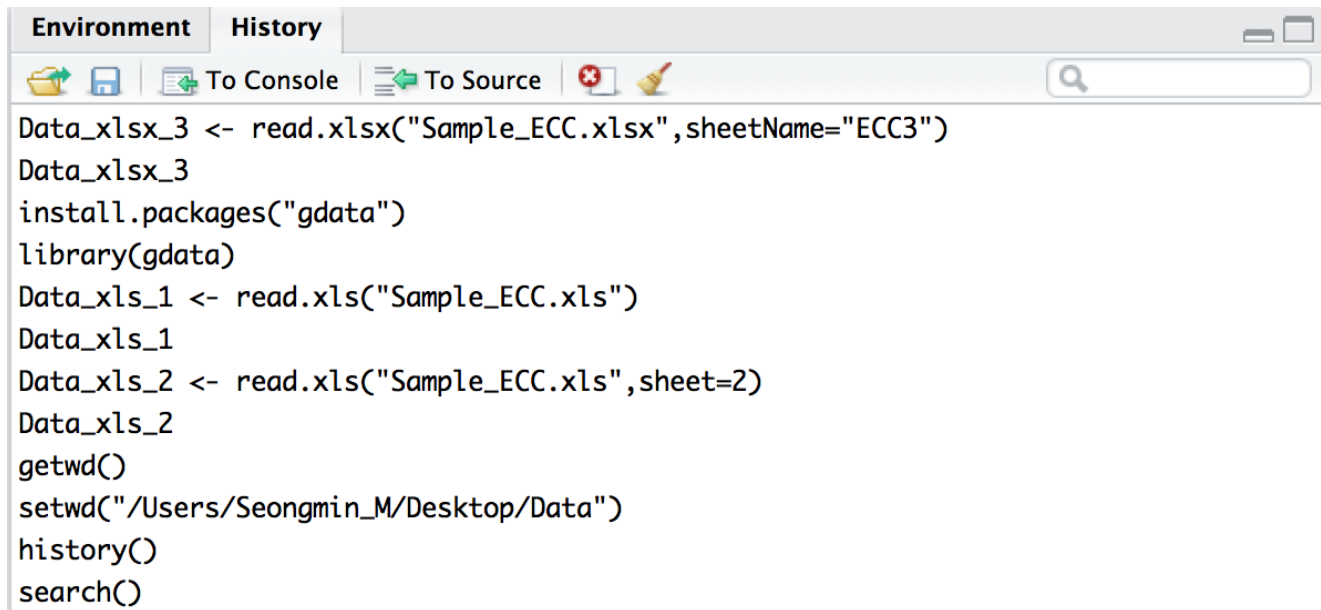
Data Handling(Input & Output)

- 경로 확인 및 지정

```
> getwd()  
[1] "/Users/Seongmin_M/Desktop/Data"  
>  
> setwd("/Users/Seongmin_M/Desktop/Data")
```

- 작업 내역 확인

```
> history()  
>
```



The screenshot shows the 'History' pane in R Studio. The pane has tabs for 'Environment' and 'History', with 'History' selected. Below the tabs are icons for file operations and buttons for 'To Console' and 'To Source'. The history list contains the following commands:

```
Data_xlsx_3 <- read.xlsx("Sample_ECC.xlsx",sheetName="ECC3")  
Data_xlsx_3  
install.packages("gdata")  
library(gdata)  
Data_xls_1 <- read.xls("Sample_ECC.xls")  
Data_xls_1  
Data_xls_2 <- read.xls("Sample_ECC.xls",sheet=2)  
Data_xls_2  
getwd()  
setwd("/Users/Seongmin_M/Desktop/Data")  
history()  
search()
```

Data Handling(Input & Output)

- 설치된 패키지 확인

> search()

[1] ".GlobalEnv"	"package:gdata"	"package:xlsx"
[4] "package:xlsxjars"	"package:XML"	"package:devtools"
[7] "package:RColorBrewer"	"package:ROAuth"	"package:twitter"
[10] "package:RJSONIO"	"package:RCurl"	"package:bitops"
[13] "package:tm"	"package:NLP"	"package:KoNLP"
[16] "package:Sejong"	"package:tau"	"package:hash"
[19] "package:stringr"	"package:rJava"	"tools:rstudio"
[22] "package:stats"	"package:graphics"	"package:grDevices"
[25] "package:utils"	"package:datasets"	"package:methods"
[28] "Autoloads"	"package:base"	

- 데이터 생성

```
> Sample_data = rbind(  
+   c("Anakin", "male", "Tatooine", "41.9BBY", "yes"),  
+   c("Amidala", "female", "Naboo", "46BBY", "no"),  
+   c("Luke", "male", "Tatooine", "19BBY", "yes"),  
+   c("Leia", "female", "Alderaan", "19BBY", "no"),  
+   c("Obi-Wan", "male", "Stewjon", "57BBY", "yes"),  
+   c("Han", "male", "Corellia", "29BBY", "no"),  
+   c("Palpatine", "male", "Naboo", "82BBY", "no"),  
+   c("R2-D2", "unknown", "Naboo", "33BBY", "no"))
```

Data Handling(Input & Output)

- Data.Frame으로 형태 변환

```
> Sample_df = data.frame(Sample_data)
>
```

- 열 이름 지정

```
> names(Sample_df) = c("Name", "Gender", "Homeworld", "Born", "Jedi")
>
> Sample_df
```

	Name	Gender	Homeworld	Born	Jedi
1	Anakin	male	Tatooine	41.9BBY	yes
2	Amidala	female	Naboo	46BBY	no
3	Luke	male	Tatooine	19BBY	yes
4	Leia	female	Alderaan	19BBY	no
5	Obi-Wan	male	Stewjon	57BBY	yes
6	Han	male	Corellia	29BBY	no
7	Palpatine	male	Naboo	82BBY	no
8	R2-D2	unknown	Naboo	33BBY	no

Data Handling(Input & Output)

- 행 이름 지정

```
> row.names(Sample_df) = c("#1", "#2", "#3", "#4", "#5", "#6", "#7", "#8")
```

```
>
```

```
> Sample_df
```

	Name	Gender	Homeworld	Born	Jedi
#1	Anakin	male	Tatooine	41.9BBY	yes
#2	Amidala	female	Naboo	46BBY	no
#3	Luke	male	Tatooine	19BBY	yes
#4	Leia	female	Alderaan	19BBY	no
#5	Obi-Wan	male	Stewjon	57BBY	yes
#6	Han	male	Coreellia	29BBY	no
#7	Palpatine	male	Naboo	82BBY	no
#8	R2-D2	unknown	Naboo	33BBY	no

- 데이터 속성 확인

```
> str(Sample_df)
```

```
'data.frame': 8 obs. of 5 variables:
```

```
$ Name : Factor w/ 8 levels "Amidala","Anakin",...: 2 1 5 4 6 3 7 8
```

```
$ Gender : Factor w/ 3 levels "female","male",...: 2 1 2 1 2 2 2 3
```

```
$ Homeworld: Factor w/ 5 levels "Alderaan","Coreellia",...: 5 3 5 1 4 2 3 3
```

```
$ Born : Factor w/ 7 levels "19BBY","29BBY",...: 4 5 1 1 6 2 7 3
```

```
$ Jedi : Factor w/ 2 levels "no","yes": 2 1 2 1 2 1 1 1
```

Data Handling(Input & Output)

- 상위 6개 데이터 확인

```
> head(Sample_df)
```

	Name	Gender	Homeworld	Born	Jedi
#1	Anakin	male	Tatooine	41.9BBY	yes
#2	Amidala	female	Naboo	46BBY	no
#3	Luke	male	Tatooine	19BBY	yes
#4	Leia	female	Alderaan	19BBY	no
#5	Obi-Wan	male	Stewjon	57BBY	yes
#6	Han	male	Corellia	29BBY	no

- 저장된 데이터 확인

```
> ls()
```

```
[1] "data"          "Data_xls_1"    "Data_xls_2"    "Data_xlsx_2"   "Data_xlsx_3"
[6] "Sample_csv_1"  "Sample_csv_2"  "Sample_data"    "Sample_df"     "Sample_txt_1"
```

- 데이터 지정하여 삭제하기

```
> rm(data)
```

```
> ls()
```

```
[1] "Data_xls_1"    "Data_xls_2"    "Data_xlsx_2"   "Data_xlsx_3"   "Sample_csv_1"
[6] "Sample_csv_2"  "Sample_data"    "Sample_df"     "Sample_txt_1"
```

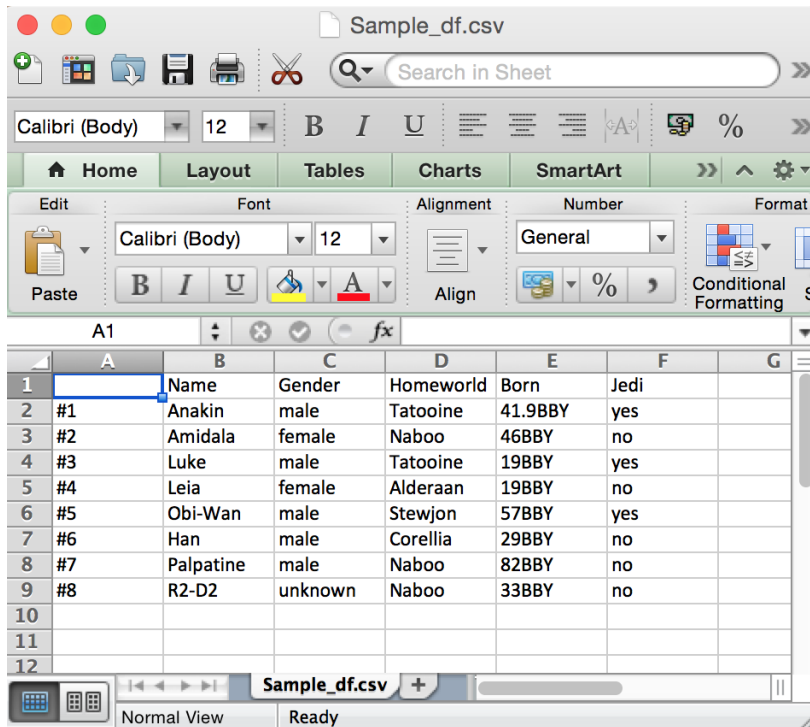

Data Handling(Input & Output)

- 전체 데이터 삭제하기

```
> rm(list=ls())  
> ls()  
character(0)
```

- 데이터 CSV형태로 내보내기(Output)

```
> write.csv(Sample_df,file="Sample_df.csv")
```



The screenshot shows a Microsoft Excel spreadsheet titled 'Sample_df.csv'. The data is organized into columns: Name, Gender, Homeworld, Born, and Jedi. The rows list characters from the Star Wars franchise, including Anakin, Amidala, Luke, Leia, Obi-Wan, Han, Palpatine, and R2-D2. The interface includes the standard Excel ribbon with tabs for Home, Layout, Tables, Charts, and SmartArt. The 'Home' tab is active, showing options for font, alignment, and number formatting. The status bar at the bottom indicates 'Normal View' and 'Ready'.

	A	B	C	D	E	F	G
1		Name	Gender	Homeworld	Born	Jedi	
2	#1	Anakin	male	Tatooine	41.9BBY	yes	
3	#2	Amidala	female	Naboo	46BBY	no	
4	#3	Luke	male	Tatooine	19BBY	yes	
5	#4	Leia	female	Alderaan	19BBY	no	
6	#5	Obi-Wan	male	Stewjon	57BBY	yes	
7	#6	Han	male	Corellia	29BBY	no	
8	#7	Palpatine	male	Naboo	82BBY	no	
9	#8	R2-D2	unknown	Naboo	33BBY	no	
10							
11							
12							

Data Handling(Input & Output)

- CSV형태의 데이터 읽어 들이기(Input)

```
> Sample_csv_1 <- read.csv("Sample_df.csv",head=T)
```

```
>
```

```
> Sample_csv_1
```

	X	Name	Gender	Homeworld	Born	Jedi
1	#1	Anakin	male	Tatooine	41.9BBY	yes
2	#2	Amidala	female	Naboo	46BBY	no
3	#3	Luke	male	Tatooine	19BBY	yes
4	#4	Leia	female	Alderaan	19BBY	no
5	#5	Obi-Wan	male	Stewjon	57BBY	yes
6	#6	Han	male	Corellia	29BBY	no
7	#7	Palpatine	male	Naboo	82BBY	no
8	#8	R2-D2	unknown	Naboo	33BBY	no

- 원하는 열 데이터 추출하기

```
> Sample_csv_2 = Sample_csv_1[,2:6]
```

```
>
```

```
> Sample_csv_2
```

	Name	Gender	Homeworld	Born	Jedi
1	Anakin	male	Tatooine	41.9BBY	yes
2	Amidala	female	Naboo	46BBY	no
3	Luke	male	Tatooine	19BBY	yes
4	Leia	female	Alderaan	19BBY	no
5	Obi-Wan	male	Stewjon	57BBY	yes
6	Han	male	Corellia	29BBY	no
7	Palpatine	male	Naboo	82BBY	no
8	R2-D2	unknown	Naboo	33BBY	no

- 행이름 지정하기

```
> row.names(Sample_csv_2) = Sample_csv_1[,1]
```

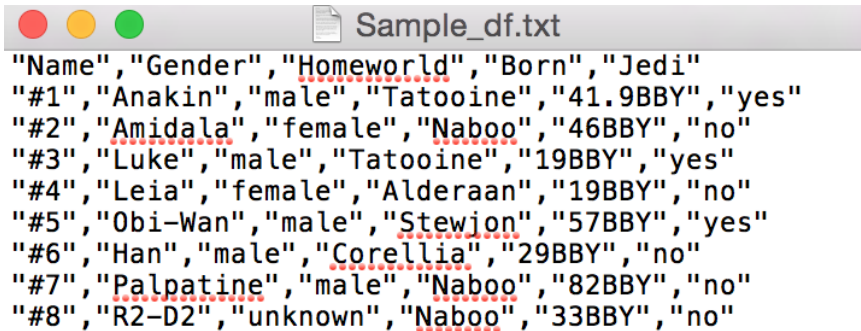
```
> Sample_csv_2
```

	Name	Gender	Homeworld	Born	Jedi
#1	Anakin	male	Tatooine	41.9BBY	yes
#2	Amidala	female	Naboo	46BBY	no
#3	Luke	male	Tatooine	19BBY	yes
#4	Leia	female	Alderaan	19BBY	no
#5	Obi-Wan	male	Stewjon	57BBY	yes
#6	Han	male	Corellia	29BBY	no
#7	Palpatine	male	Naboo	82BBY	no
#8	R2-D2	unknown	Naboo	33BBY	no

Data Handling(Input & Output)

- 데이터 TXT형태로 내보내기(Output)

```
> write.table(Sample_df,file="Sample_df.txt",sep=",")
```



```
"Name","Gender","Homeworld","Born","Jedi"
"#1","Anakin","male","Tatooine","41.9BBY","yes"
"#2","Amidala","female","Naboo","46BBY","no"
"#3","Luke","male","Tatooine","19BBY","yes"
"#4","Leia","female","Alderaan","19BBY","no"
"#5","Obi-Wan","male","Stewjon","57BBY","yes"
"#6","Han","male","Corellia","29BBY","no"
"#7","Palpatine","male","Naboo","82BBY","no"
"#8","R2-D2","unknown","Naboo","33BBY","no"
```

- TXT형태의 데이터 읽어 들이기(Input)

```
> Sample_txt_1 <- read.table("Sample_df.txt",header=TRUE,sep=",")
```

```
> str(Sample_txt_1)
```

```
'data.frame': 8 obs. of 5 variables:
```

```
$ Name : Factor w/ 8 levels "Amidala","Anakin",...: 2 1 5 4 6 3 7 8
```

```
$ Gender : Factor w/ 3 levels "female","male",...: 2 1 2 1 2 2 2 3
```

```
$ Homeworld: Factor w/ 5 levels "Alderaan","Corellia",...: 5 3 5 1 4 2 3 3
```

```
$ Born : Factor w/ 7 levels "19BBY","29BBY",...: 4 5 1 1 6 2 7 3
```

```
$ Jedi : Factor w/ 2 levels "no","yes": 2 1 2 1 2 1 1 1
```

```
> Sample_txt_1
```

	Name	Gender	Homeworld	Born	Jedi
#1	Anakin	male	Tatooine	41.9BBY	yes
#2	Amidala	female	Naboo	46BBY	no
#3	Luke	male	Tatooine	19BBY	yes
#4	Leia	female	Alderaan	19BBY	no
#5	Obi-Wan	male	Stewjon	57BBY	yes
#6	Han	male	Corellia	29BBY	no
#7	Palpatine	male	Naboo	82BBY	no
#8	R2-D2	unknown	Naboo	33BBY	no

One Sample T-test

One Sample T-test

● One Sample T-test

- 일 표본집단의 특성에 대한 가설을 검증하는 것으로 평균에 대한 가설과 비율에 대한 가설로 나뉜다.
- 표본 집단의 평균이 기존의 가설과 다르다는 것을 알고자 하면 양측 검증을 사용한다.
- 표본 집단의 평균이 기존의 가설 평균 값보다 작을 경우 좌측 단측 검증을 사용하고, 클 경우 우측 단측 검증을 사용한다.
- 계산된 T값(검정 통계량)이 T분포에서 문제 상황에 해당하는 T기준 값보다 크면 대립가설을 채택한다.

● 검정 통계량

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \quad S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

\bar{X} =표본평균, μ =모집단의 평균, S =표본 표준편차, n =표본의 수

One Sample T-test

- 평균에 대한 가설(Hypothesis)

- H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu = X$)
- H_1 (대립가설) : 기존의 평균값과 차이가 있다. (좌 : $\mu < X$, 우 : $\mu > X$, 양측 : $\mu \neq X$)

- 비율에 대한 가설(Hypothesis)

- H_0 (귀무가설) : 기존의 확률 값과 차이가 없다.
- H_1 (대립가설) : 기존의 확률 값과 차이가 있다.

One Sample T-test

- 신뢰구간(Confidence interval)
 - 실제 모수가 존재할 것으로 예측되는 구간으로 90%, 95%, 99%정도의 구간 추정이 가능하다.
 - 실제로는 95%신뢰 구간 추정이 통상적으로 사용된다.
 - Ex) 95%신뢰구간 : 예측된 구간 내에 실제 모평균이 있을 가능성이 95%라고 신뢰할 수 있는 구간

$$\text{모평균의 95\% 신뢰구간} = \bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \quad (\text{표본 평균 } \bar{X}, \text{표본 표준편차 } s, \text{표본의 크기 } n)$$

$$\text{모비율의 95\% 신뢰구간} = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}} \quad (\text{표본의 관심 사건의 비율 } p, \text{표본의 크기 } n)$$

One Sample T-test(ratio)

One Sample T-test(ratio)

- 예제를 활용한 모 비율 검증

- 어느 한 도시의 실업률은 5.5%로 알려져 있다.
- 어느 단체에서 이를 다시 조사한 결과 520명중 39명이 구직중인 것을 확인 할 수 있었다.
- 공표한 내용이 사실인지 신뢰성 95%로 검증하시오.

- 가설(Hypothesis)

- H_0 (귀무가설) : 작년 평균 실업률과 차이가 없다.
- H_1 (대립가설) : 작년 평균 실업률과 차이가 있다.

One Sample T-test(ratio)

- 검증
- H_0 (귀무가설) : 작년 평균 실업률과 차이가 없다.
- H_1 (대립가설) : 작년 평균 실업률과 차이가 있다.

```
> prop.test(39,520,0.055)
```

```
1-sample proportions test with continuity correction
```

```
data: 39 out of 520, null probability 0.055
X-squared = 3.6264, df = 1, p-value = 0.05687
alternative hypothesis: true p is not equal to 0.055
95 percent confidence interval:
 0.05452366 0.10197090
sample estimates:
      p 
0.075
```

- 모 비율 비교: 올해의 평균 실업률과 작년 평균 실업률은 차이가 없다.
- 대립가설(H_1 : 작년 평균 실업률과 차이가 있다.)을 기각, 귀무가설(H_0 : 작년 평균 실업률과 차이가 없다.)을 채택한다.

One Sample T-test(ratio)

- Q : 작년 평균 실업률이 0.5%였다면 결과 값은 어떠한가?

- 검증
- H0 (귀무가설) : 작년 평균 실업률과 차이가 없다.
- H1 (대립가설) : 작년 평균 실업률과 차이가 있다.

```
> prop.test(39,520,0.05)
```

```
1-sample proportions test with continuity correction
```

```
data: 39 out of 520, null probability 0.05
X-squared = 6.3259, df = 1, p-value = 0.0119
alternative hypothesis: true p is not equal to 0.05
95 percent confidence interval:
 0.05452366 0.10197090
sample estimates:
      p 
0.075
```

- 모 비율 비교: 올해의 평균 실업률과 작년 평균 실업률은 차이가 있다.
- 귀무가설(H0: 작년 평균 실업률과 차이가 없다.) 을 기각 , 대립가설(H1: 작년 평균 실업률과 차이가 있다.) 을 채택한다.

One Sample T-test(ratio)

- Q : 만약 신뢰구간의 수준이 99%라면 결과 값은 어떠한가?
- 검증
- H0 (귀무가설) : 작년 평균 실업률과 차이가 없다.
- H1 (대립가설) : 작년 평균 실업률과 차이가 있다.

```
> prop.test(39,520,0.05,conf.level=0.99)
```

1-sample proportions test with continuity correction

```
data: 39 out of 520, null probability 0.05
X-squared = 6.3259, df = 1, p-value = 0.0119
alternative hypothesis: true p is not equal to 0.05
99 percent confidence interval:
 0.04952988 0.11151740
sample estimates:
      p
0.075
```

- 신뢰구간의 값이 변경된 것을 확인 할 수 있다.

One Sample T-test(mean) and Bar chart

One Sample T-test(mean)

- 예제를 활용한 모 평균 검증

- 어느 수학 동아리 학생의 작년 IQ평균은 120이었고 올해 신입 동아리 학생들의 IQ는 아래와 같다.
- IQ = 127,125,110,115,130,123,135,140,120,105
- 올해 학생들과 작년 학생들간의 IQ차이가 있는지 신뢰수준 95%로 검증하시오.

- 가설(Hypothesis)

- H0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=120$)
- H1 (대립가설) : 기존의 평균값과 차이가 있다. (좌 : $\mu<120$, 우 : $\mu>120$, 양측 : $\mu\neq120$)

One Sample T-test(mean)

- 데이터 입력 및 확인

```
> y=c(127,125,110,115,130,123,135,140,120,105)
> y
[1] 127 125 110 115 130 123 135 140 120 105
```

- 좌측검증

- H0 (귀무가설) : 기존의 평균값과 차이가 없다.($\mu=120$)
- H1 (대립가설) : 기존의 평균값과 차이가 있다. (좌 : $\mu<120$)

```
> t.test(y,alternative = c("less"),mu=120,conf.level=0.95)
```

One Sample t-test

```
data: y
t = 0.8709, df = 9, p-value = 0.7968
alternative hypothesis: true mean is less than 120
95 percent confidence interval:
 -Inf 129.3147
sample estimates:
mean of x
123
```

- 평균 차 비교: 올해 학생들의 IQ는 작년 학생들의 IQ와 차이가 없다.
- 대립가설(H1: 기존의 평균값과 차이가 있다.)을 기각, 귀무가설(H0: 기존의 평균값과 차이가 없다.)을 채택한다.

One Sample T-test(mean)

- 우측검증
- H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=120$)
- H_1 (대립가설) : 기존의 평균값과 차이가 있다. (우 : $\mu>120$)

```
> t.test(y, alternative = c("greater"), mu=120, conf.level=0.95)
```

One Sample t-test

```
data: y
t = 0.8709, df = 9, p-value = 0.2032
alternative hypothesis: true mean is greater than 120
95 percent confidence interval:
 116.6853      Inf
sample estimates:
mean of x
    123
```

- 평균 차 비교: 올해 학생들의 IQ는 작년 학생들의 IQ와 차이가 없다.
- 대립가설(H_1 : 기존의 평균값과 차이가 있다.)을 기각, 귀무가설(H_0 : 기존의 평균값과 차이가 없다.)을 채택한다.

One Sample T-test(mean)

- 양측검증
- H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=120$)
- H_1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : $\mu \neq 120$)

```
> t.test(y,mu=120)
```

One Sample t-test

```
data: y
t = 0.8709, df = 9, p-value = 0.4065
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 115.2073 130.7927
sample estimates:
mean of x
    123
```

- 평균 차 비교: 올해 학생들의 IQ는 작년 학생들의 IQ와 차이가 없다.
- 대립가설(H_1 : 기존의 평균값과 차이가 있다.)을 기각, 귀무가설(H_0 : 기존의 평균값과 차이가 없다.)을 채택한다.

One Sample T-test(mean)

- Q : 만약 작년 학생들의 IQ 평균이 110이었다면 결과 값은 어떠한가?

- 양측검증
- H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=110$)
- H_1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : $\mu \neq 110$)

```
> t.test(y,mu=110)
```

One Sample t-test

```
data: y
t = 3.7738, df = 9, p-value = 0.004391
alternative hypothesis: true mean is not equal to 110
95 percent confidence interval:
 115.2073 130.7927
sample estimates:
mean of x
    123
```

- 평균 차 비교: 올해 학생들의 IQ는 작년 학생들의 IQ와 차이가 있다.
- 귀무가설(H_0 : 기존의 평균값과 차이가 없다.) 을 기각 , 대립가설(H_1 : 기존의 평균값과 차이가 있다.) 을 채택한다.

One Sample T-test(mean)

- Q : 만약 신뢰구간의 수준이 99%라면 결과 값은 어떠한가?
 - 양측검증
 - H0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=110$)
 - H1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : $\mu \neq 110$)

```
> t.test(y,mu=110,conf.level=0.99)
```

One Sample t-test

```
data: y
t = 3.7738, df = 9, p-value = 0.004391
alternative hypothesis: true mean is not equal to 110
99 percent confidence interval:
 111.805 134.195
sample estimates:
mean of x
    123
```

- 신뢰구간의 값이 변경된 것을 확인 할 수 있다.

One Sample T-test(mean)

- Q : 결과를 심층적으로 해석하면 어떠한가?

- 양측검증
 - H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=120$)
 - H_1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : $\mu \neq 120$)

```
> t.test(y,mu=120,conf.level=0.95)
```

One Sample t-test

```
data: y
t = 0.8709, df = 9, p-value = 0.4065
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 115.2073 130.7927
sample estimates:
mean of x
    123
```

- 평균은 123이고, T값은 0.8709, 자유도는 9(n-1), p값은 0.4065로 H_0 (귀무가설)를 채택한다. 따라서 기존의 평균과 차이가 없다고 할 수 있다.
- 또한 자료가 표본 일 경우 전체 집단의 평균 값으로 추정되는 신뢰구간은 $115.2073 < \bar{X} < 130.7927$ 이다.

One Sample T-test(mean)

- Q : 결과를 심층적으로 해석하면 어떠한가?

- 양측검증
 - H_0 (귀무가설) : 기존의 평균값과 차이가 없다. ($\mu=120$)
 - H_1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : $\mu \neq 120$)

- 분산 구하기

```
> var(y)
[1] 118.6667
```

- 표준편차 구하기

```
> sd(y)
[1] 10.89342
```

- 평균오차(표본 평균의 표준 편차) 구하기

```
> 10.89342/sqrt(10)
[1] 3.444802
>
```

- T값 구하기

```
> (123-120)/(10.89342/sqrt(10))
[1] 0.8708774
>
```

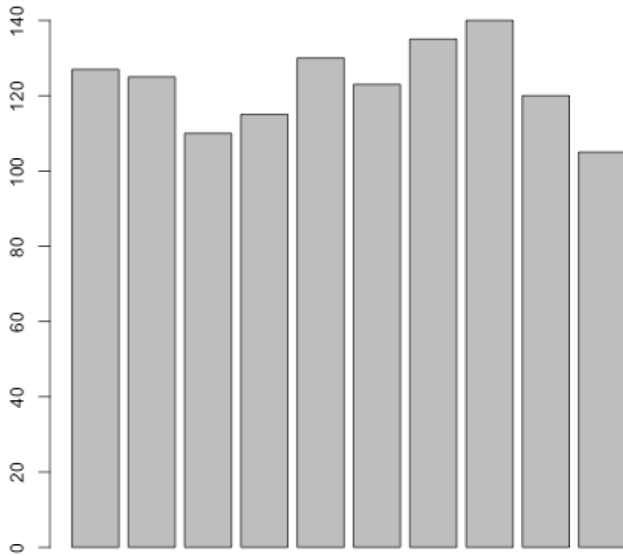
- 신뢰구간 구하기

```
> 123+1.96*(10.89342/sqrt(10))
[1] 129.7518
>
> 123-1.96*(10.89342/sqrt(10))
[1] 116.2482
>
```

Bar chart

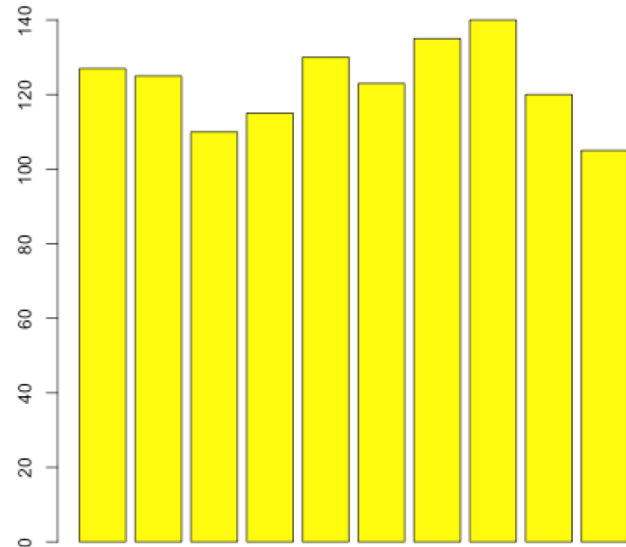
- 막대그래프(Bar chart)를 통한 데이터 분석

- 막대그래프 그리기



```
> barplot(y)
```

- 막대그래프에 색 추가하기

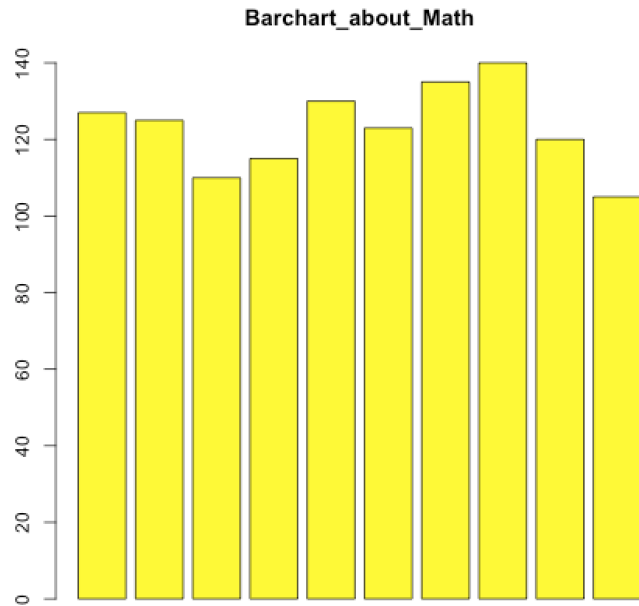


```
> barplot(y,col="yellow")
```

Bar chart

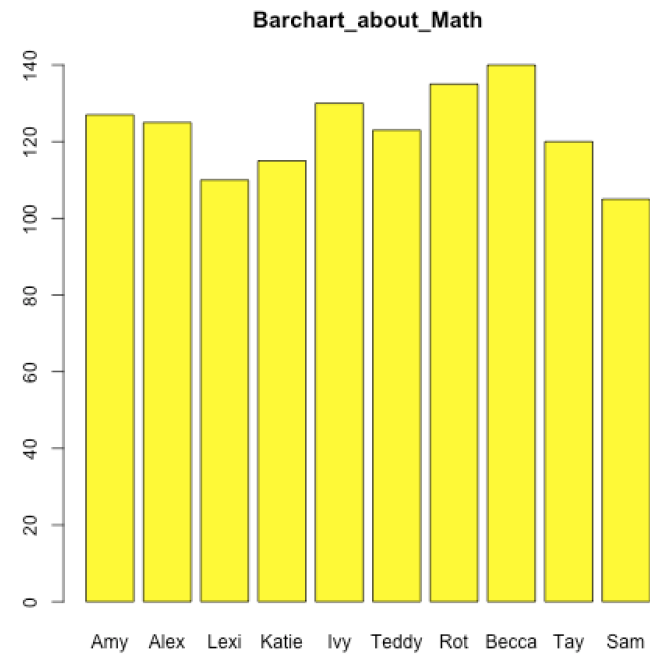
- 막대그래프(Bar chart)를 통한 데이터 분석

- 제목 추가하기



```
> barplot(y,col="yellow",main="Barchart_about_Math")
```

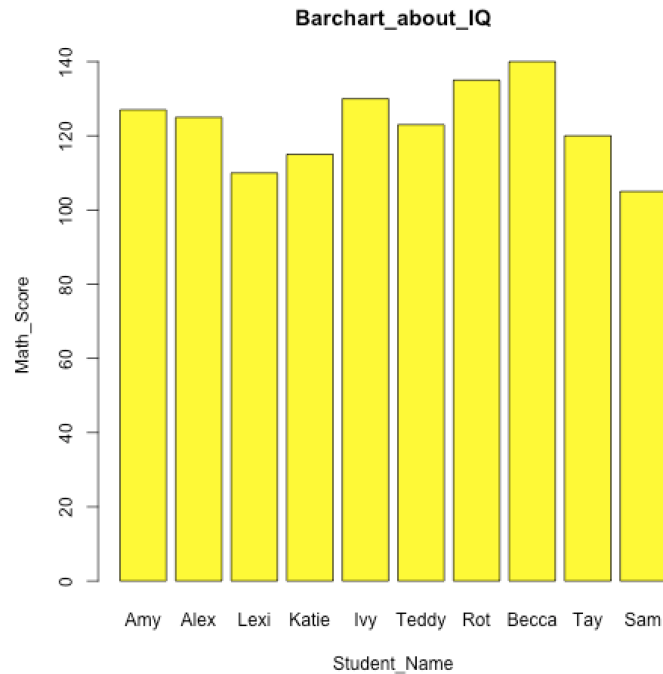
- 학생 이름 추가하기



```
>  
> Name<-c("Amy","Alex","Lexi","Katie","Ivy","Teddy","Rot","Becca","Tay","Sam")  
>  
> barplot(y,col="yellow",main="Barchart_about_Math",names.arg=Name)  
>
```

Bar chart

- 막대그래프(Bar chart)를 통한 데이터 분석
 - X축과 Y축의 이름을 지정하고 그림 파일로(png) 내보내기



```
> png("Barchart_about_IQ")#####  
>  
> barplot(y,col="yellow",main="Barchart_about_IQ",names.arg=Name,xlab="Student_Name",ylab="Math_Score")  
>  
> dev.off()
```

RStudioGD

NA Handling

- 과제
 - (1) One Sample T-test(mean)를 활용할 수 있는 예제를 만들고 99%신뢰수준으로 예제를 분석하고 결과를 해석하시오.
 - (2) 자신이 만든 예제를 Bar chart를 사용하여 분석하시오.