역사 데이터 시각화 분석

Descriptive Statistic

Descriptive Statistic

- 기술통계학(Descriptive statistics)

  - 관심의 대상이 되는 자료에 대해 그림 및 수치를 사용하여 정리하고 요약하는 방법
  - 추론통계를 위한 사전단계로 수집된 자료의 분석에 초점을 둔다.
  - 수치를 활용한 방법 : 비율, 지수, 평균, 분산 등
  - 그림을 활용한 방법 : 막대그래프, 히스토그램, 상자그림 등

- 수식(Formuler)

$$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n} \qquad S=\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}} \qquad S^2=\frac{\sum(X_i-\bar{X})^2}{n-1}$$

$\bar{X}$=평균, S=표준편차, S^2=분산

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 데이터 설명하기

Student Admissions at UC Berkeley

Description

Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

Usage

```
UCBAdmissions
```

Format

A 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

- 데이터 확인하기(head,str)

```
> head(DF)
     Admit Gender Dept Freq
1 Admitted   Male    A  512
2 Rejected   Male    A  313
3 Admitted Female    A   89
4 Rejected Female    A   19
5 Admitted   Male    B  353
6 Rejected   Male    B  207
> str(DF)
'data.frame':   24 obs. of  4 variables:
 $ Admit : Factor w/ 2 levels "Admitted","Rejected": 1 2 1 2 1 2 1 2 1 2 ...
 $ Gender: Factor w/ 2 levels "Male","Female": 1 1 2 2 1 1 2 2 1 1 ...
 $ Dept  : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 2 2 2 2 3 3 ...
 $ Freq  : num  512 313 89 19 353 207 17 8 120 205 ...
```

- Data.frame형태로 가져오기

```
> DF=as.data.frame(UCBAdmissior
> DF
      Admit Gender Dept Freq
1  Admitted   Male    A  512
2  Rejected   Male    A  313
3  Admitted Female    A   89
4  Rejected Female    A   19
5  Admitted   Male    B  353
6  Rejected   Male    B  207
7  Admitted Female    B   17
8  Rejected Female    B    8
9  Admitted   Male    C  120
10 Rejected   Male    C  205
11 Admitted Female    C  202
12 Rejected Female    C  391
13 Admitted   Male    D  138
14 Rejected   Male    D  279
15 Admitted Female    D  131
16 Rejected Female    D  244
17 Admitted   Male    E   53
18 Rejected   Male    E  138
19 Admitted Female    E   94
20 Rejected Female    E  299
21 Admitted   Male    F   22
22 Rejected   Male    F  351
23 Admitted Female    F   24
24 Rejected Female    F  317
```

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 데이터 적용하기

```
> attach(DF)
The following object is masked from Exam_1 (pos = 3):

    Gender

The following object is masked from Exam_1 (pos = 4):

    Gender
```

- 최대값 구하기

```
> max(Freq)
[1] 512
```

- 평균값 구하기

```
> mean(Freq)
[1] 188.5833
```

- 분산 구하기

```
> var(Freq)
[1] 19617.82
```

- 최소값 구하기

```
> min(Freq)
[1] 8
```

- 중앙값 구하기

```
> median(Freq)
[1] 170
```

- 표준편차 구하기

```
> sd(Freq)
[1] 140.0636
```

- 데이터 요약하기

```
> summary(DF)
      Admit       Gender   Dept       Freq
 Admitted:12   Male  :12   A:4   Min.   :  8.0
 Rejected:12   Female:12   B:4   1st Qu.: 80.0
                           C:4   Median :170.0
                           D:4   Mean   :188.6
                           E:4   3rd Qu.:302.5
                           F:4   Max.   :512.0
```

- 데이터 적용 해지하기

```
> detach(DF)
```

Descriptive Statistic

- 기술통계학(Descriptive statistics)을 사용하여 데이터 분석하기

- 변수생성

```
> a=c(1,2,3,1,2,3,4,1,2,3,4,1,2,1,3,4,1,1)
> b=c(5,6,7,5,6,5,6,7,8,5,6,8,5,5,6,7,5,5)
```

- 도수 분포표 만들기

```
> table(a)
a
1 2 3 4
7 4 4 3
> table(b)
b
5 6 7 8
8 5 3 2
```

- 도수 분할표 만들기

```
> table(a,b)
   b
a   5 6 7 8
  1 5 0 1 1
  2 1 2 0 1
  3 2 1 1 0
  4 0 2 1 0
```

- UCBAdmissions데이터를 활용하여 도수 분할표 만들기

```
> xtabs(Freq ~ Gender + Admit, DF)
        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278
```

Descriptive Statistic

- Pressure데이터를 활용한 선 그래프
- 데이터 확인

pressure {datasets} R Documentation

Vapor Pressure of Mercury as a Function of Temperature

Description

Data on the relation between temperature in degrees Celsius and vapor pressure of mercury in millimeters (of mercury).

Usage

```
pressure
```

Format

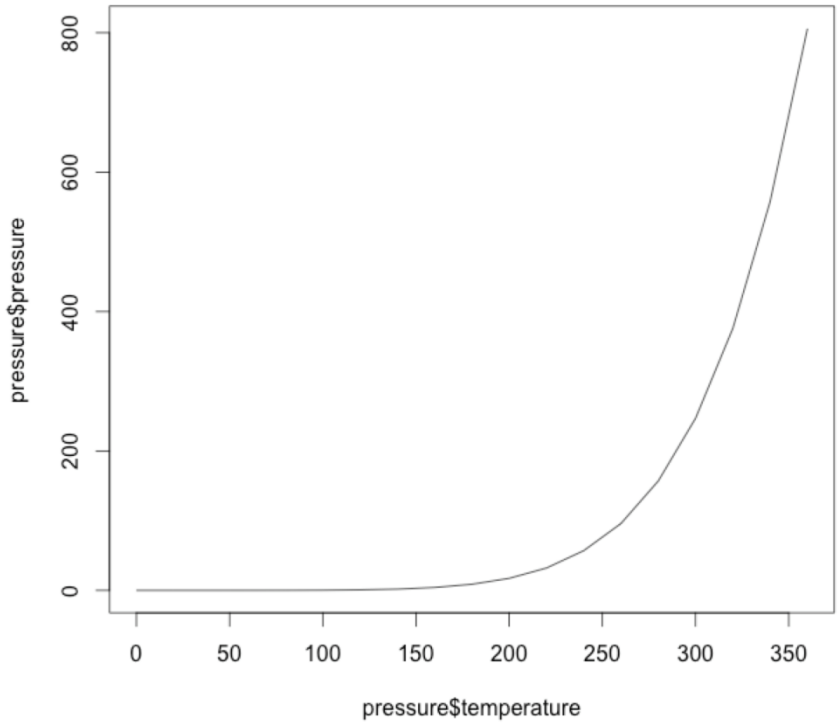
A data frame with 19 observations on 2 variables.

[, 1] temperature numeric temperature (deg C)
[, 2] pressure numeric pressure (mm)

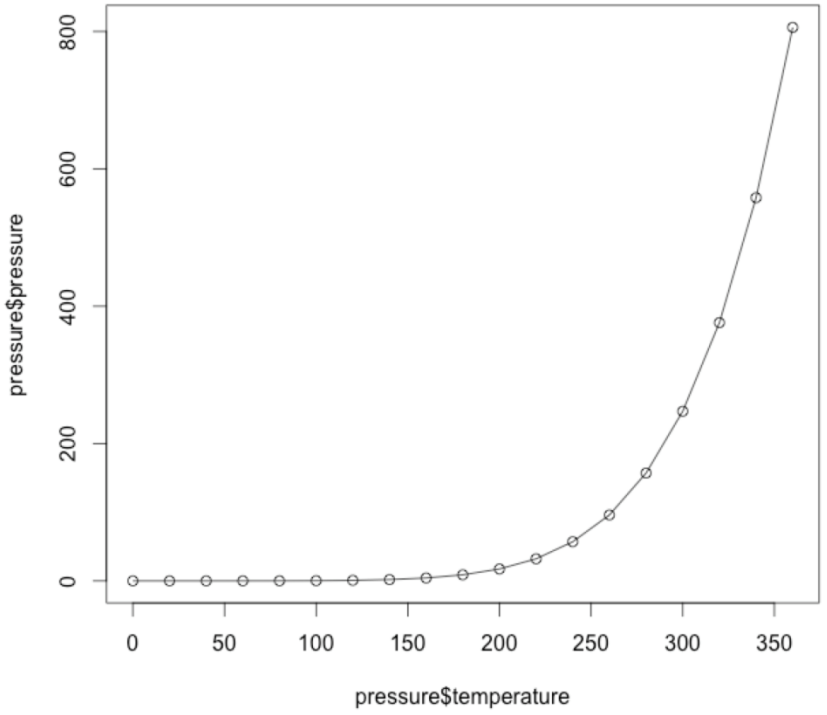
pressure데이터 설명 : 1 미리리터의 수은의 증기압과 섭씨온도 사이의 관계
변수설명
temperature = 섭씨온도
pressure = 1 미리리터의 수은의 증기압력

Descriptive Statistic

- Pressure데이터를 활용한 선 그래프




```
> plot(pressure$temperature,pressure$pressure,type="l")
```




```
> points(pressure$temperature,pressure$pressure)
```

Descriptive Statistic

- mtcars데이터를 활용한 산점도, 히스토그램
  - 데이터 확인

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 variables.

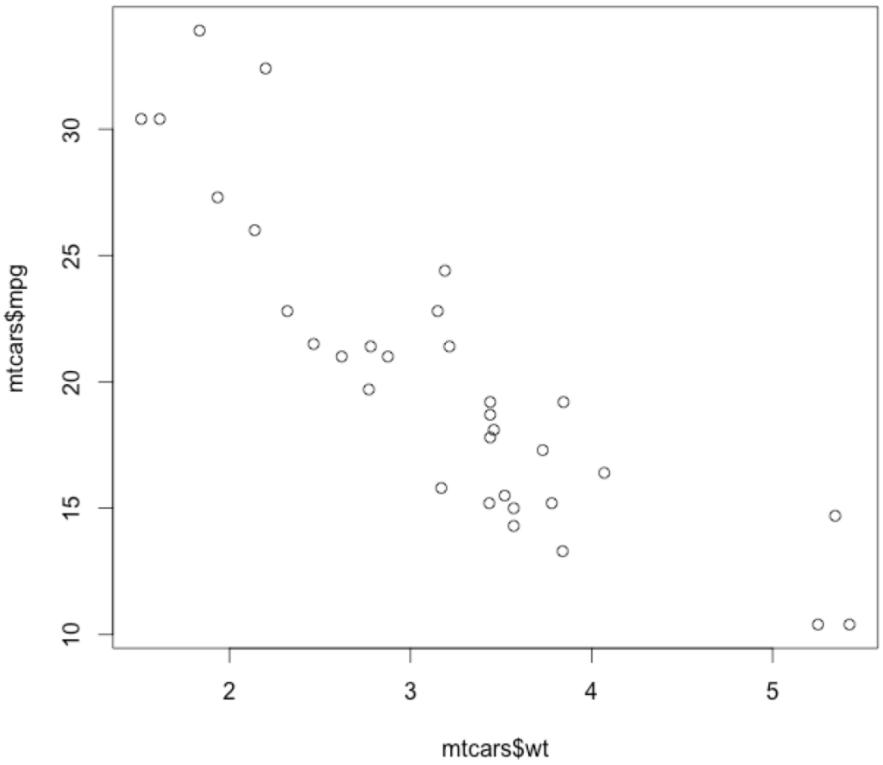
[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

mtcars데이터 설명
이 데이터는 1974 년 모터 트렌드 미국 잡지에서 추출하였으며 1973년-1974년도 모델의 32종의 자동차들의 연비등 자동차의 10가지 중요요소를 보여준다.

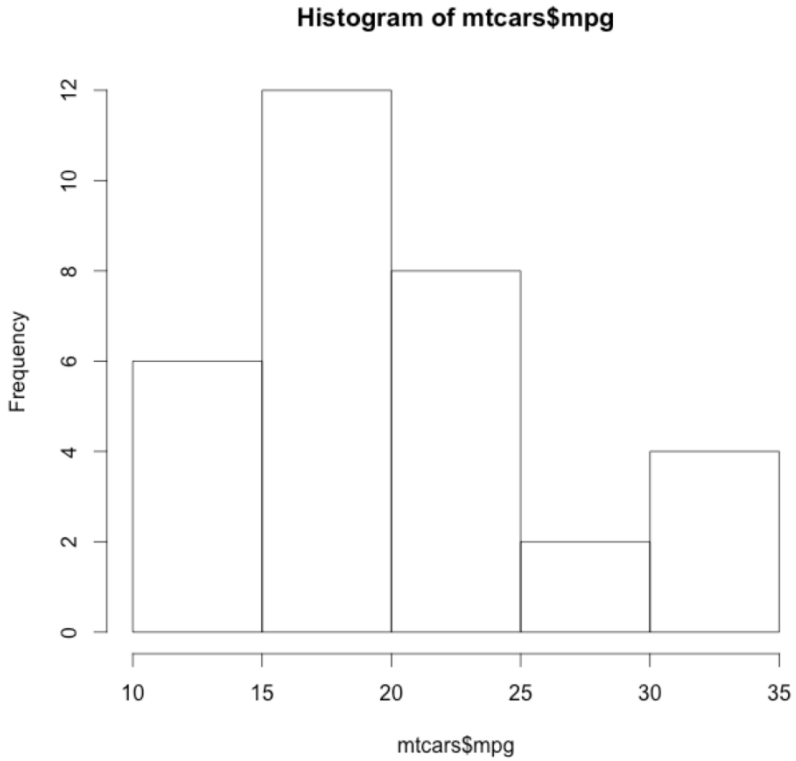
변수설명
mpg = 마일 / (US) 갤런
cyl = 실린더의 수
disp = 변위 (cu.in.)
hp = 총 마력
drat = 리어 액슬 비율
wt = 무게 (파운드 / 1000)
qsec = 1/4 마일 시간
vs = V / S
am = 변속기 (0 = 자동, 1 = 수동)
gear = 기어의 수
carb = 기화기의 수

Descriptive Statistic

- mtcars데이터를 활용한 산점도, 히스토그램




```
>
> plot(mtcars$wt,mtcars$mpg)
>
```




```
>
> hist(mtcars$mpg)
>
```

Descriptive Statistic

- BOD데이터를 활용한 막대 그래프
- 데이터 확인

BOD {datasets} R Documentation

Biochemical Oxygen Demand

Description

The BOD data frame has 6 rows and 2 columns giving the biochemical oxygen demand versus time in an evaluation of water quality.

Usage

```
BOD
```

Format

This data frame contains the following columns:

`Time`

A numeric vector giving the time of the measurement (days).

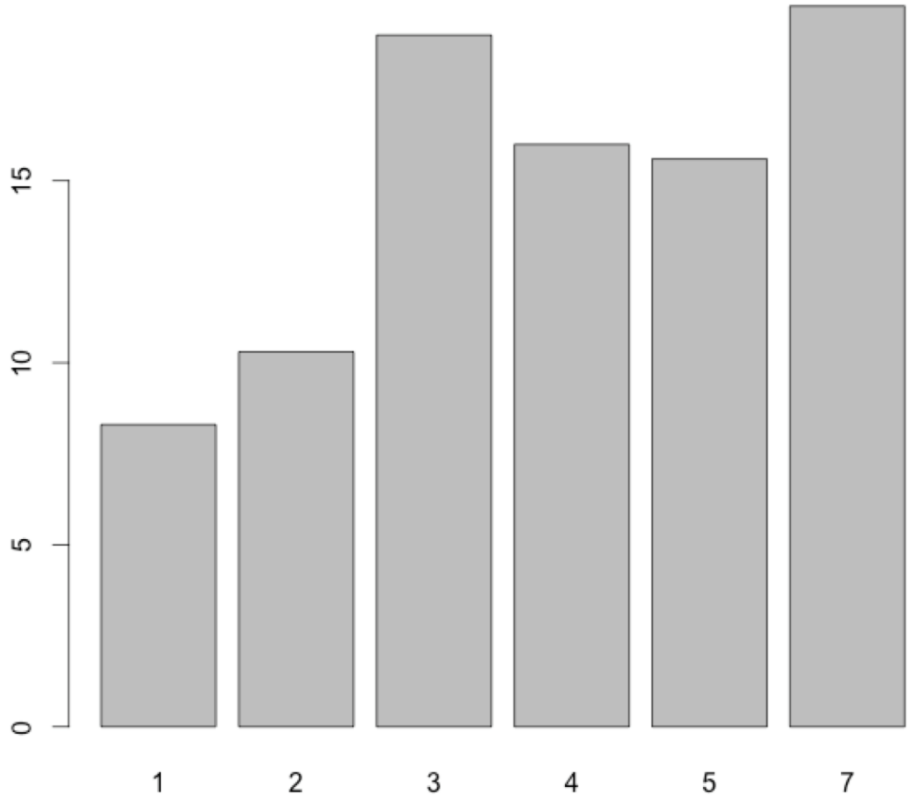
`demand`

A numeric vector giving the biochemical oxygen demand (mg/l).

BOD데이터 설명 : 수질평가 시간과 생화학적 산소요구량의 관계를 보여주는 자료
변수설명
time = 수질 평가 측정 시간 , A numeric vector giving the time of the measurement (days).
demand = 산소요구량 , A numeric vector giving the biochemical oxygen demand (mg/l).

Descriptive Statistic

- BOD데이터를 활용한 막대 그래프




```
>
> barplot(BOD$demand,names.arg=BOD$Time)
>
```

Descriptive Statistic

- ToothGrowth데이터를 활용한 상자그림

- 데이터 확인

ToothGrowth {datasets} R Documentation

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

Description

The response is the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg) with each of two delivery methods (orange juice or ascorbic acid).

Usage

```
ToothGrowth
```

Format

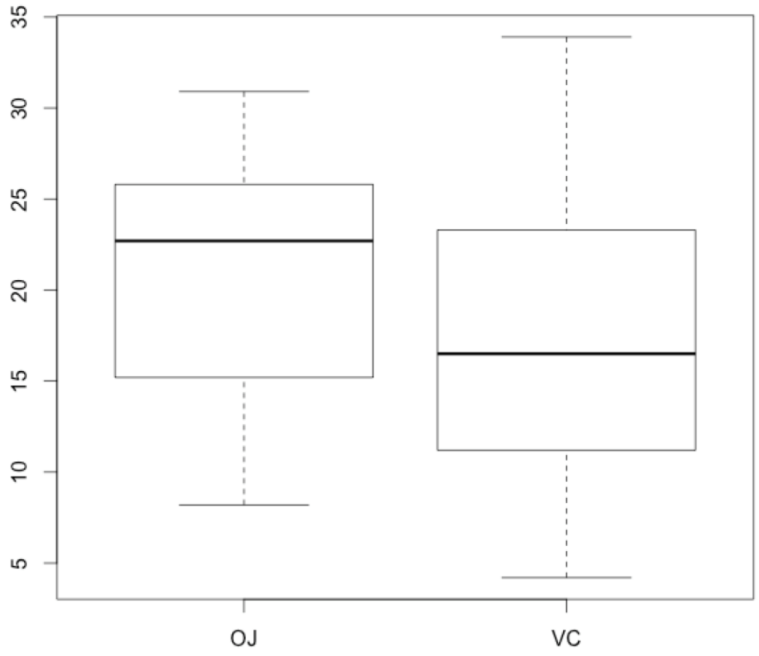
A data frame with 60 observations on 3 variables.

[,1] len numeric Tooth length
[,2] supp factor Supplement type (VC or OJ).
[,3] dose numeric Dose in milligrams.

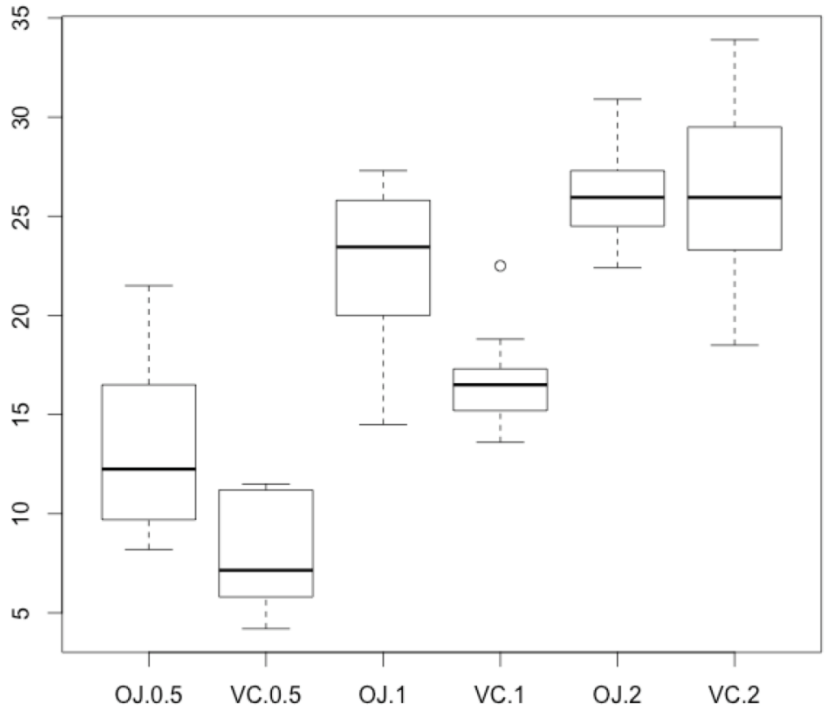
ToothGrowth데이터 설명 :데이터는 각 기니피그 아세포치아의 길이에 대해 두가지 전송방법(오렌지쥬스, 아스코르브산)과 비타민C의 세레벨(0.5mg,1mg,2mg)을 혼합하여 대입하였을때의 반응을 비교한 데이터이다.

변수설명
len 기니피그치아의 길이(Tooth length)
supp 두가지 전송방법(Supplement type (VC or OJ))
dose 비타민 C의 레벨(Dose in milligrams.)

- ToothGrowth데이터를 활용한 상자그림




```
>
> boxplot(len ~ supp, data = ToothGrowth)
>
```




```
>
> boxplot(len ~ supp + dose, data = ToothGrowth)
>
```

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.

- 검정 통계량

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

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

One Sample T-test

- 평균에 대한 가설(Hypothesis)

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

- 비율에 대한 가설(Hypothesis)

  - H0 (귀무가설) : 기존의 확률 값과 차이가 없다.
  - H1 (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

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

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

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

One Sample T-test(ratio)

- 검증
  - H0 (귀무가설) : 작년 평균 실업률과 차이가 없다.
  - H1 (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(H1: 작년 평균 실업률과 차이가 있다.)을 기각 , 귀무가설(H0: 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.(μ=120)
- H1 (대립가설) : 기존의 평균값과 차이가 있다. (좌 : μ<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

- H0 (귀무가설) : 기존의 평균값과 차이가 없다.(μ=120)
- H1 (대립가설) : 기존의 평균값과 차이가 있다. (우 : μ>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(H1: 기존의 평균값과 차이가 있다.)을 기각 , 귀무가설(H0: 기존의 평균값과 차이가 없다.)을 채택한다.

One Sample T-test(mean)

- 양측검증

- H0 (귀무가설) : 기존의 평균값과 차이가 없다.(μ=120)
- H1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : μ≠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(H1: 기존의 평균값과 차이가 있다.)을 기각 , 귀무가설(H0: 기존의 평균값과 차이가 없다.)을 채택한다.

One Sample T-test(mean)

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

  - 양측검증

  - H0 (귀무가설) : 기존의 평균값과 차이가 없다.(μ=110)
  - H1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : μ≠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(H0: 기존의 평균값과 차이가 없다.) 을 기각 , 대립가설(H1: 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.(μ=110)
  - H1 (대립가설) : 기존의 평균값과 차이가 있다. (양측 : μ≠110)

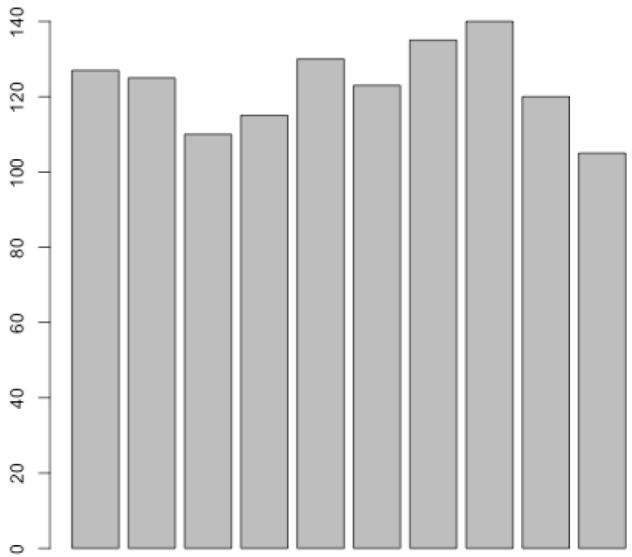
```
> t.test(y,mu=110,conf.level=0.99)

        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.

Bar chart

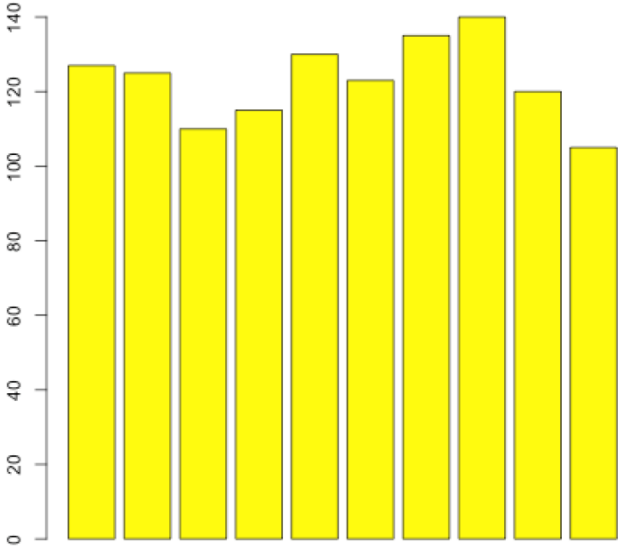
- 막대그래프(Bar chart)를 통한 데이터 분석

  - 막대그래프 그리기




```
> barplot(y)
```

  - 막대그래프에 색 추가하기

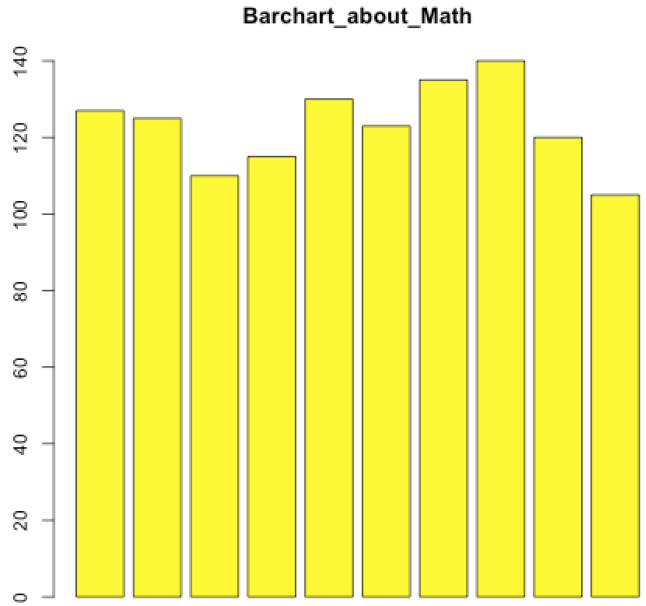



```
> barplot(y,col="yellow")
```

Bar chart

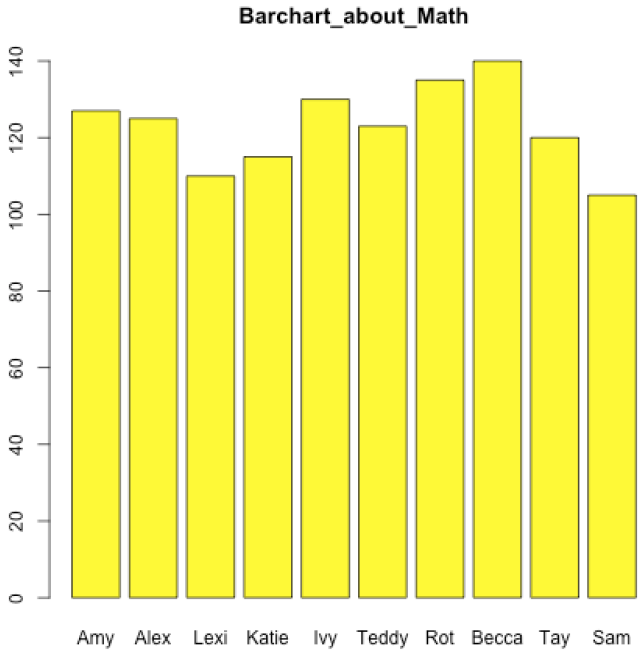
- 막대그래프(Bar chart)를 통한 데이터 분석

  - 제목 추가하기




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

  - 학생 이름 추가하기

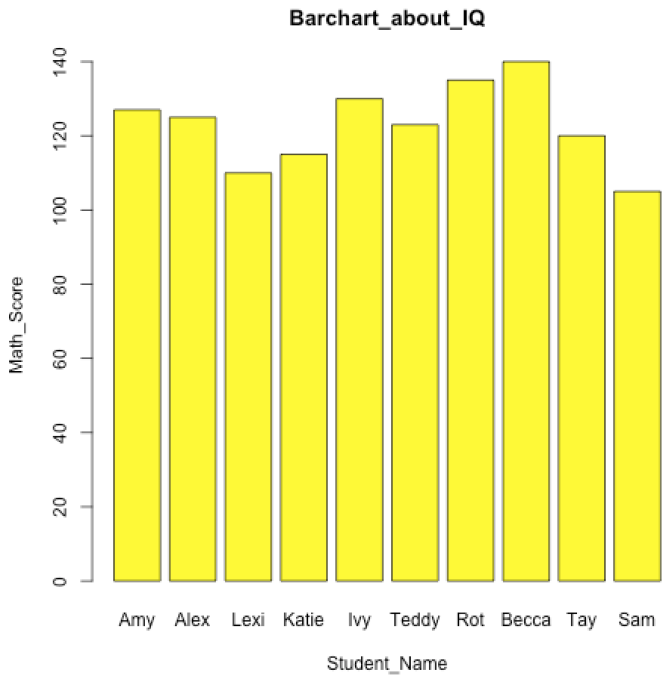



```
>
> Name<-c("Amy","Alex","Lexi","Katie","Ivy","Teddy","Rot","Becca","Tay","Sam")
>
> barplot(y,col="yellow",main="Barchart_about_Math",names.arg=Name)
>
```

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
        2
```

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.