2. 정규성검정

표본의 크기가 충분히 크지 않은데, 모집단이 정규분포를 따르는지 모를 때가 있습니다. 어떤 상황이죠? 표본평균이 정규분포를 따른다고 할 수 없는 상황입니다. 이 때 사용하는 검정이 정규성검정입니다. 내가 뽑은 표본이, 정규분포를 따르는 모집단에서 나온 것인지 아닌지를 판단해주는 검정방법입니다. 만약 정규성검정을 통해 정규성이 입증되면, z검정이나 t검정을 사용할 수 있습니다. 여기서 어떤 검정을 쓸지는 모분산을 아는지 여부에 따라 결정되구요.

1. 정규성 검정 (Normality Test) 이란?

데이터셋의 분포가 정규분포 (Normal Distribution)를 따르는지를 검정하는 것이다.

Statistics의 여러 검정법들이 데이터의 정규분포를 가정하고 (예: t-test) 수행되기 때문에,
데이터 자체의 정규성을 확인하는 검정 과정이 필수적이겠다.

~~중심극한정리~~에 의해 표본수(n)가 30이 넘어가면 데이터셋이 정규분포에 가까워진다.

그러나, 경우에 따라 30이 넘어감에도 데이터 특이성에 따라 정규분포를 반드시 따르지 않을 수도 있기에,
Normality Test를 통해 데이터의 정규분포를 확인해보자.

2. 정규성 검정 종류

ㄱ) **Shaprio-Wilks test**

-표본수(n)가 2000 미만인 데이터셋에 적합한 정규성 검정

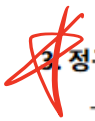
ㄴ) **Kolmogorove-Smirnov test**

-표본수(n)가 2000 초과인 데이터셋에 적합한 정규성 검정

ㄷ) **Quantile-Quantile plot (Graphic test)**

-데이터셋이 정규분포를 따르는지 판단하는 시각적 분석 방법

-분석할 데이터 종류가 많지 않다면, QQplot을 통해 시각적으로 확인해보는게 가장 간단하며 직관적이다.



3. 정규성 검정의 H0, H1

-귀무가설(H0) : 데이터셋이 정규분포를 따른다.

-대립가설(H1) : 데이터셋이 정규분포를 따르지 않는다.

-귀무가설을 기각하고 대립가설이 채택된다면 ($p<0.01$ or 0.05) 해당 데이터셋은 정규분포를 따르지 않는 것이다.

4. 정규성 검정 예제 (R)

-Pima Indian: 9~13세기에 걸쳐 아메리카로 이주해온 몽골리언계

-주식: 식물성. (나무의 순, 잡초, 밀, 콩, 호박 등)

-1960년대 이후 고지방/고칼로리 식습관으로 당뇨환자 증가.

#Pima.tr data (8개의 변수)

npreg: number of pregnancies.

glu: plasma glucose concentration in an oral glucose tolerance test.

bp: diastolic blood pressure (mm Hg).

skin:triceps skin fold thickness (mm).

bmi: body mass index (weight in kg/(height in m)\^2).

ped: diabetes pedigree function.

age: age in years.

type: Yes or No, for diabetic according to WHO criteria.

 # Pima.tr 이 들어있는 라이브러리

```
> library(MASS)
```

 # Pima.tr$bmi 로 접근하지 않더라도, 바로 bmi로 접근 가능하다.

```
> attach(Pima.tr)

> head(Pima.tr)
 npreg glu bp skin  bmi   ped age type
1     5  86 68   28 30.2 0.364  24   No
2     7 195 70   33 25.1 0.163  55  Yes
3     5  77 82   41 35.8 0.156  35   No
4     0 165 76   43 47.9 0.259  26   No
5     0 107 60   25 26.4 0.133  23   No
6     5  97 76   27 35.6 0.378  52  Yes
```

- type

Yes: 당뇨병을 가진 환자

No: 당뇨병이 없는 환자

#정규성 검정에 들어가기 전에 다시한 위에 가설 검정을 상기하자.

-H0 (귀무가설): 주어진 데이터의 분포는 정규분포를 따른다.

-H1 (대립가설): 주어진 데이터의 분포는 정규분포를 따르지 않는다.

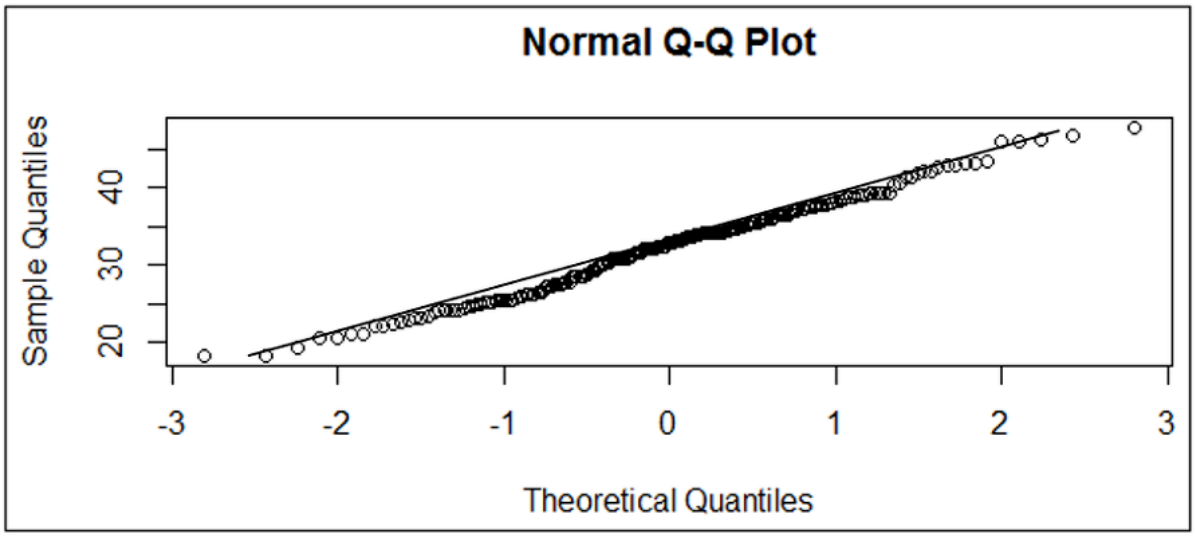
```
>shapiro.test(bmi)
    Shapiro-Wilk normality test

data:  bmi
W = 0.991, p-value = 0.2523
```

 # p-value를 통해 귀무가설을 기각할 수 없으므로 정규분포를 따른다고 할 수 있다.

#정규성 검정을 통해 데이터의 정규성 확인

#그러면, QQplot 을 통해, 시각적으로도 확인해보자.

```
>qqnorm(bmi)
>qqline(bmi)
```



위 QQplot 은 데이터의 quantile(분위수)과 특정 이론적 분포의 quantile 각각 구하여 산점도로 나타낸 그림이다.

QQplot 의 점들이 기울기의 직선상에 놓이면 자료가 해당 분포를 잘 따르거나 두 모집단 분포가 같다고 해석할 수 있다.

쉽게말해, 데이터셋의 점들이 라인을 따라서 잘 붙어있으므로 정규성을 따른다고 말할 수 있겠다.

■ kolmogorov-smirnov는 표본수가 많을 때 사용하는 검정 기법입니다.
(50 개 이상 - 중심극한의 정리를 감안하면 30개 이상)

보통 표본수가 30개 이상이면, 해당 확률변수가 '정규분포'를 따른다 가정한다
그래서 이 정규성 검정은 자주 사용되지 않는다.

■ shapiro-wilk는 표본수가 적을 때 적용할 수 있도록 개발된 검정 기법입니다.
(50개 미만 - 중심극한의 정리를 감안하면 30개 미만)

다만, 이때 주의하셔야 할 점은 위 두 검정의 귀무가설입니다.

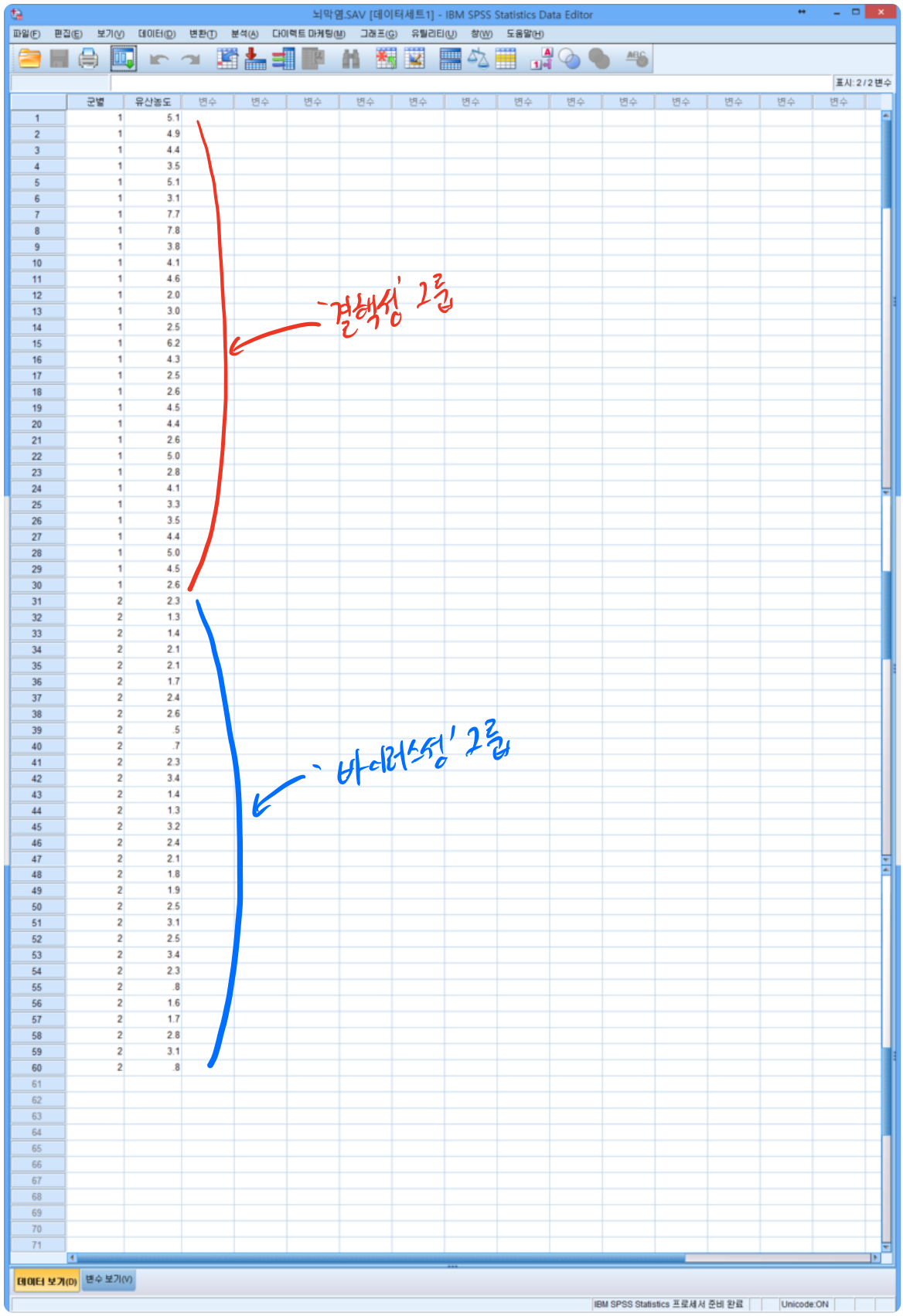
■ H0 : 표본의 분포는 정규성을 만족한다

■ H1 : 표본의 분포는 정규성을 만족하지 않는다
(정규분포가 아니다)

따라서 위에 표에서 '만나이'는 kolmogorov-smirnov와 shapiro-wilk 모두 5% 유의수준에서 귀무가설이 기각되었으므로 정규분포가 아닌 것으로 확인이 되었습니다.

즉, 비모수통계를 이용해서 분석을 해야하는 변수인 것이죠.

참고로 정규성 검정과 관련한 논문을 살펴보시면 shapiro-wilk의 결과값이 언급된 논문을 더욱 쉽게 보실 수 있으십니다. 왜냐하면 연구자들이 논문 작성시에는 표본수가 적어서 비모수 통계를 해야 할지 아니면 모수 통계를 해도 괜찮을지 판단을 위해 정규성 검정을 실시하기 때문입니다. '


뇌막염.SAV [데이터세트1] - IBM SPSS Statistics Data Editor
파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)
표시: 2 / 2 변수
군별
유산농도
변수
1 1 5.1
2 1 4.9
3 1 4.4
4 1 3.5
5 1 5.1
6 1 3.1
7 1 7.7
8 1 7.8
9 1 3.8
10 1 4.1
11 1 4.6
12 1 2.0
13 1 3.0
14 1 2.5
15 1 6.2
16 1 4.3
17 1 2.5
18 1 2.6
19 1 4.5
20 1 4.4
21 1 2.6
22 1 5.0
23 1 2.8
24 1 4.1
25 1 3.3
26 1 3.5
27 1 4.4
28 1 5.0
29 1 4.5
30 1 2.6
31 2 2.3
32 2 1.3
33 2 1.4
34 2 2.1
35 2 2.1
36 2 1.7
37 2 2.4
38 2 2.6
39 2 .5
40 2 .7
41 2 2.3
42 2 3.4
43 2 1.4
44 2 1.3
45 2 3.2
46 2 2.4
47 2 2.1
48 2 1.8
49 2 1.9
50 2 2.5
51 2 3.1
52 2 2.5
53 2 3.4
54 2 2.3
55 2 .8
56 2 1.6
57 2 1.7
58 2 2.8
59 2 3.1
60 2 .8
'결핵성' 그룹
'바이러스성' 그룹
데이터 보기(D)
변수 보기(V)
IBM SPSS Statistics 프로세서 준비 완료
Unicode:ON

케이스 처리 요약

		케이스					
		유효함		결측값		총계	
	그룹구분	N	퍼센트	N	퍼센트	N	퍼센트
유산농도	결핵성	30	100.0%	0	0.0%	30	100.0%
	바이러스성	30	100.0%	0	0.0%	30	100.0%

기술통계

그룹구분				통계	표준 오류
유산농도	결핵성	평균		4.130	.2578
		평균의 95% 신뢰구간	하한	3.603	
			상한	4.657	
		5% 잘린 평균		4.037	
		중앙값		4.200	
		분산		1.993	
		표준 편차		1.4118	
		최소값		2.0	
		최대값		7.8	
		범위		5.8	
		사분위수 범위		2.0	
		왜도		.953	.427
		첨도		1.183	.833
	바이러스성	평균		2.050	.1467
		평균의 95% 신뢰구간	하한	1.750	
			상한	2.350	
		5% 잘린 평균		2.057	
		중앙값		2.100	
		분산		.646	
		표준 편차		.8038	
		최소값		.5	
		최대값		3.4	
		범위		2.9	
		사분위수 범위		1.1	
		왜도		-.144	.427
		첨도		-.654	.833

정규성 검정

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	그룹구분	통계	df	유의수준	통계	df	유의수준
유산농도	결핵성	.146	30	.102	.915	30	.020
	바이러스성	.091	30	.200*	.970	30	.532

*. 실질적인 유의수준의 하한입니다.

a. Lilliefors 유의수준 정정

지난시간에 정규성 검정은 "Kolmogorov-Smirnov"와 "Shapiro-Wilk" 을 통해 검정할 수 있다고 하였다. 위의 표에서 "정규성 검정" 표를 확인하도록 하자.
왼쪽의 Kolmogorov-Smirnov는 결핵성, 바이러스성 모두 "p > 0.05"을 만족하였고, 오른쪽으로 "Shapiro-Wilk"는 결핵성이 만족하지 못하였다. 양족 모두 만족하면 좋겠지만 둘 중 하나만 만족해도 "정규성 검정"에는 문제없다고 하니 본격적으로 t-test 과정으로 이동한다.

1. 표본으로 정규성검정을 하여 정규성을 위반한다는 것이 증명되지 않는 경우 모집단이 정규성을 따르고 있다고 봅니다
왜냐하면 표본이 모집단을 잘 반영하고 있다고 보기 때문입니다