1. t검정이란?

t검정(t-test)은 두 집단 간 평균의 차이를 통계적으로 검정하고자 할 때 사용된다.

2. t분포의 이해와 t검정의 유형

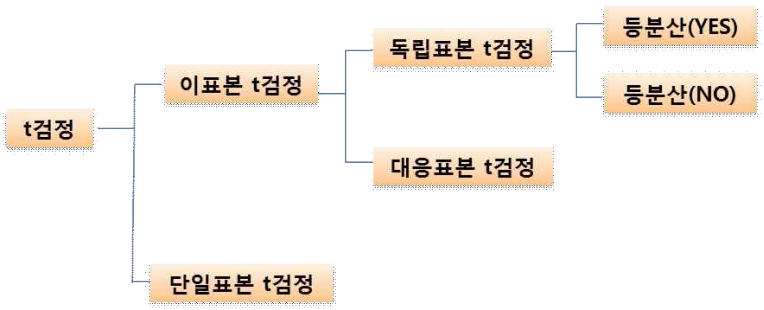
1) t분포

t분포는 모집단이 표준편차가 알려져 있지 않을 경우, 다시 말해 모집단이 정규분포를 따르지만 표준정규분포를 그릴 수 없는 경우 주로 사용되는 분포이다. 검정통계량 t는 다음의 식으로 구할 수 있다.

$$t = \frac{Z}{\sqrt{\frac{V}{v}}}$$

이 식에서 n은 표본의 수, Z는 표준정규분포의 확률을 의미하고, $V = \frac{(n-1)S^2}{\sigma^2}$을 의미하고, 마지막으로 v는 (n-1)이다. 통계학에서 (n-1)을 자유도라 한다. t값이 커질수록 귀무가설이 기각되고 대립가설이 채택될 확률이 커짐을 의미한다. t분포 역시 0을 중심으로 좌우 대칭인 그래프이며 자유도가 커질수록, 즉 표본의 수가 커질수록 표준정규분포에 가까워진다. 따라서 표본의 수가 작을 경우에 활용될 수 있다. 일반적으로 30개가 기준이 된다.

2) t분포의 유형



위 그림은 t검정의 유형을 도식화한 것이다.

3. R활용 일표본 t검정

1) 단일표본 t검정

단일표본 t검정은 표본이 하나일 경우 활용할 수 있는 t검정이다. 단일표본 t검정은 모집단에서 추출한 표본과 모집단의 차이를 비교하는 것이다. 따라서 단일표본 t검정을 하기 위해서는 반드시 모집단의 평균을 미리 알고 있어야 한다.

2) 가설의 설정

우리나라의 75개의 자치시(모집단) 중 20개의 자치시를 표본으로 추출하여 추출한 20개의 자치시가 모집단의 특성과 유사한지를 알아 보고자한다. 합계출산율에 초점을 맞추고자 하는데 모집단의 합계출산율 평균은 1.37812임을 알고 있다. 그렇다면 귀무가설과 대립가설은 다음과 같이 나타낼 수 있다.

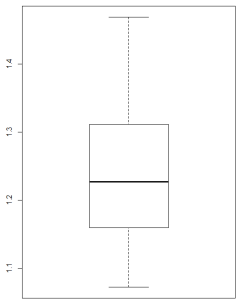
$$H_0 : \mu = 1.37812$$
$$H_1 : \mu \neq 1.37812$$

3) R활용 단일표본 t검정

```
setwd("C:/Users/USER/Desktop/교육원초급/data/xlsx 파일")
library(openxlsx)
mydata = read.xlsx("onesample.xlsx",1)
attach(mydata)
View(mydata)
```

cities	ID	birth_rate
경기 수원시	1	1.292
경기 성남시	2	1.159
경기 의정부시	3	1.104
경기 안양시	4	1.177
경기 부천시	5	1.072
경기 광명시	6	1.235
경기 평택시	7	1.469
경기 동두천시	8	1.292
경기 안산시	9	1.219
경기 고양시	10	1.161

```
> summary(birth_rate) # birth_rate의 기술통계량을 구함
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.072   1.159   1.227   1.246   1.308   1.469
> sd(birth_rate) # birth_rate(합계출산율)의 표준편차를 구함
[1] 0.1272301
> boxplot(birth_rate)
```



R에서 t검정은 t.test라는 명령어를 통해 검정통계량을 구할 수 있다. 전체 76개 자치시의 합계출산율 편균은 1.37812었다. 따라서 현재 R에 코딩되어 있는 20개 자치시의 평균과 모집단인 76개 자치시의 평균이 동일

한지를 분석한 것이다.

```
> t.test(birth_rate,mu=1.37812) # 단일표본 t-test의 시행(모집단의 평균 : 1.37812)

        One Sample t-test

data:  birth_rate
t = -4.6387, df = 19, p-value = 0.0001791
alternative hypothesis: true mean is not equal to 1.37812
95 percent confidence interval:
 1.186604 1.305696
sample estimates:
mean of x
  1.24615
```

먼저 t값과 p값을 확인한다. t는 검정통계량을 의미하고, p는 유의확률을 의미한다. 분석 결과 검정통계량이 높고 유의확률이 0.1이하임을 알 수 있다. 이와 같은 결과는 귀무가설이 기각되었고 대립가설이 채택되었음을 의미한다. 다시 말해 “20개 자치시의 평균이 모집단의 평균과 통계적으로 차이가 없다.”는 진술이 기각되었고, “20개 자치시의평균이 모집단의 평균과 차이가 있다.”는 진술이 채택된 것이다.

df=19는 자유도를 의미하는데 표본의 수에서 1을 뺀 값이다. 다음으로 신뢰구간이 표기되어있으며 ‘mean of x’는 birth_rate의 평균이 1.24615임을 알 수 있다.

4. R확용 독립표본 t검정

1) 독립표본 t검정

독립표본 t검정은 서로 독립적인 두 모집단으로부터 추출한 두 표본의 평균이 통계적으로 다른지 여부를 분석하는 것이다. 모집단이 독립적이므로 각각의 모집단에서 추출한 두 개의 표본 집단 역시 서로 독립적이다. 예

를 들어 여성과 남성의 입사 후 첫 연봉의 차이는 표본 여성과 표본 남성은 어떠한 영향도 주고받지 않고 상호독립적이므로 독립표본 t검정을 통해 검정할 수 있다.

독립표본 t검정을 위해 추출된 두 표본의 개수는 반드시 같을 필요는 없지만 두 집단의 분산은 동일한지 여부는 확인해야한다. 분산이 동일하지 않다면 통계적으로 보정해야한다.

2) 가설의 설정

예를 들어, 우리나라 시(자치시)와 군(자치군)의 합계출산율이 차이가 나는지를 분석하고자 한다면 독립표본 t검정이 가장 적합한 분석방법이고 가설은 다음과 같이 설정된다.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

여기서 μ_1은 자치시 합계출산율의 평균이고, μ_2는 자치군의 평균이다.

3) R 활용 독립표본 t검정

우리나라 자치시와 자치군의 합계출산율 차이를 통계적으로 규명하고자 한다. 차이가 나지 않는다는 진술이 귀무가설이고 차이가 난다는 가설이 귀무가설이다.

먼저 Rstudio에서 독립표본 t검정을 위한 데이터를 불러온다.

```
> mydata1 = read.xlsx("independent.xlsx",1) # 파일을 불러와 mydata1로 저장
> attach(mydata1) # R과 mydata1을 연동
> View(mydata1) # mydata1을 좌측 상단에서 확인
```

cities	birth_rate	dummy
경기 수원시	1.292	1
경기 성남시	1.159	1
경기 의정부시	1.104	1
경기 안양시	1.177	1
경기 부천시	1.072	1
경기 광명시	1.235	1
경기 평택시	1.469	1
경기 동두천시	1.292	1
경기 안산시	1.219	1
경기 고양시	1.161	1

```
> names(mydata1) # mydata1안에 포함되어 있는 변수의 이름을 확인
[1] "cities"     "birth_rate" "dummy"
```

자료는 자치시와 자치군의 데이터가 코딩되어 있다. dummy변수는 해당 표본이 자치시에 속하면 “1”로, 자치군에 속하면 “0”으로 코딩되었다.

독립표본 t검정은 등분산이 가정되는 경우와 아닌 경우가 있다. 먼저 등분산이 가정되는 경우는 살펴보자.

```
> t.test(birth_rate~dummy, mu=0, alt="two.sided",conf=0.95,paired=F, var.eq=T)

        Two Sample t-test

data:  birth_rate by dummy
t = 2.458, df = 155, p-value = 0.01507
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01846389 0.16961318
sample estimates:
mean in group 0 mean in group 1
       1.472159        1.378120
```

명령어에서 alt=“two-paired”는 양측 검정을 실시하라는 뜻이고 conf=0.95는 신뢰구간을 95%로 설정하라는 의미이다. paired=F는 대응

표본이 아니라는 의미이고, var.eq=T가 등분산이 가정된 경우라는 뜻이다.

분석 결과, 검정통계량이 높은 값을 가지고있고 유의확률 p값이 0.1이하임을 알 수 있다. 다시 말해 귀무가설이 기각되었으며 대립가설이 채택되었다. 이에 자치시와 자치군의 합계출산율 차이는 통계적으로 유의미한 것으로 나타났다.

다음은 등분산이 가정되지 않을 경우이다.

```
>     t.test(birth_rate~dummy,     mu=0,     alt="two.sided",conf=0.95,paired=F,
var.eq=F)

          Welch Two Sample t-test

data:  birth_rate by dummy
t = 2.5067, df = 137.59, p-value = 0.01335
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01985852 0.16821855
sample estimates:
mean in group 0 mean in group 1
       1.472159        1.378120
```

등분산이 가정되지 않을 경우, 즉 비교 대상인 두 집단의 분산이 다를 경우에 활용할 수 있는 “Welch Two Sample t-test”가 진행되었음을 알 수 있다. 결과는 일반적으로 t검정의 결과와 유사하게 유의수준 0.05이하에서 통계적으로 유의하였음을 알 수 있다.

F검정을 통해 t검정을 진행하기 전 등분산의 가정 여부를 확인할 수 있다. F검정은 분산의 차이를 통하여 집단 간 차이가 있는지 여부를 검정하는 방법이다. 다음은 F검정의 결과이다.

```
> var.test(birth_rate~dummy)
```

```
        F test to compare two variances

data:  birth_rate by dummy
F = 2.5541, num df = 81, denom df = 74, p-value = 6.118e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.626386 3.992555
sample estimates:
ratio of variances
          2.554097
```

대립가설은 “두 집단 간 분산의 비율이 1이 아니다.”이고 귀무가설은 “두 집단 간 분산의 비율이 1이다.”이다. 분석 결과 p값이 매우 낮음을 알 수 있다. 따라서 귀무가설이 기각되어 대립가설이 채택되었다. 이는 합계출산율에서 자치시와 자치군의 등분산 가정이 이루어지지 않음을 의미한다. 따라서 이 예에서는 “Welch Two Sample t-test”를 활용하여 분석하여야한다.

5. R 활용 대응표본 t검정

1) 대응표본 t검정

대응표본 t검정은 비교하고자 하는 표본이 하나의 모집단으로부터 추출되었을 경우 활용할 수 있는 통계적 검정방법이다. 일반적으로 대응표본 t검정은 시간적 선후 관계의 차이를 통계적으로 규명하고자 할 때 많이 활용된다. 따라서 같은 표본이 시간의 차이를 두고 비교되는 것이므로 두 표본의 구성은 반드시 동일하여야 한다. 예를 들어 K대학교 행정학과 학생들 중 일부를 선발하여 토익 클래스를 운영하여 클래스 전과 후의 점수를 비교하는 경우에 대응표본 t검정이 활용된다.

2) 가설의 설정

예를 들어, 우리나라 자치시의 합계출산율이 2010년과 2015년이 다른지를 비교하고자 한다. 이때 가설은 다음과 같다.

$$H_0 : \mu_{2010} = \mu_{2015}$$
$$H_1 : \mu_{2010} \neq \mu_{2015}$$

3) R 활용 대응표본 t검정

2010과 2015년도의 합계출산율이 통계적으로 동일한지를 분석하고자 한다. 분석단위는 우리나라 자치시이다.
분석을 하기 위해 데이터를 불러온다.

```
> mydata2 = read.xlsx("paired.xlsx",1)
> attach(mydata2)
> View(mydata2)
```

cities	birth_rate_2015	birth_rate_2010	ID
경기 수원시	1.292	1.226	1
경기 성남시	1.159	1.170	2
경기 의정부시	1.104	1.179	3
경기 안양시	1.177	1.186	4
경기 부천시	1.072	1.172	5
경기 광명시	1.235	1.336	6
경기 평택시	1.469	1.447	7
경기 동두천시	1.292	1.450	8
경기 안산시	1.219	1.251	9
경기 고양시	1.161	1.165	10

```
> names(mydata2)
[1] "cities"          "birth_rate_2015" "birth_rate_2010" "ID"
```

데이터는 자치시의 합계출산율을 2010년도와 2015년도의 값으로 구분되어있다.

```
> t.test(birth_rate_2010,birth_rate_2015,paired = T)

        Paired t-test

data:  birth_rate_2010 and birth_rate_2015
t = 0.66928, df = 72, p-value = 0.5055
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -0.01664141  0.03346333
sample estimates:
mean of the differences
              0.008410959
```

대응표본 t검정 결과 검정통계량의 값이 작고 p값도 0.1이상임을 알 수 있다. 이는 귀무가설이 채택되고 대립가설이 기각되어 두 집단 간 차이가 없음을 의미한다. 다시 말해, 2010년의 합계출산율과 2015년의 합계출산율은 통계적으로 동이라다는 것을 의미한다. 평균의 차이가 0.008410959로 도출되었지만 통계적으로 이러한 차이는 무시하여도 좋다는 결론이다.