제 1 장

빅분기 실기 유형 3 예상문제
t 검정 3종 세트정리

본 교재는 슬기로운 통계생활 빅분기 작업형 제 3유형 예상문제 요약 정리본입니다.

1.1 1 표본 t 검정

영화 스트리밍 평균시간

2005년에 설립된 SPstream은 영화 스트리밍 서비스를 제공하고 있습니다. 2020년에 SPstream의 보고서에 따르면, 5G 네트워크를 이용하는 한국 사용자들은 한 달에 평균 12.6시간 동안 스마트폰에서 영화를 스트리밍 한다고 합니다. 다음은 이번달 5G 사용자 고객 중 12명을 무작위로 선택 한 후, 시청 시간을 기록한 것입니다.

$$16, 12, 9, 8, 14, 10, 17, 12, 3, 19, 18, 9$$

데이터를 사용하여 2023년 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균 시간이 12.6시간보다 큰 지 유의수준 0.05 하에서 검정하세요.

1. 귀무가설과 대립가설을 설정하세요.
2. 데이터의 표본 평균과 검정통계량 값을 구하세요.
3. p-value값을 구하고, 통계적 판단을 내려보세요.
4. 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균 시간에 대한 95% 신뢰 구간을 고르시오.

풀이 답안

위의 문제를 코드 하나로 풀 수 있습니다. 중요 포인트는 귀무가설과 대립가설을 잘 세울 수 있는가와 이것을 코드에 잘 반영하여 돌릴 수 있는가입니다.

1. 귀무가설과 대립가설을 설정하세요.

귀무가설: 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균시간은 12.6시간이다. $\mu = 12.6$

대립가설: 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균시간은 12.6시간보다 크다. $\mu > 12.6$

2, 3, 4번을 위해서 다음과 같이 코드를 작성하세요.

```r
x <- c(16, 12, 9, 8, 14, 10, 17, 12, 3, 19, 18, 9)
t.test(x, alternative = "greater", mu = 12.6, conf.level = 0.95)
```

```
>
>   One Sample t-test
>
> data:  x
> t = -0.25522, df = 11, p-value = 0.5984
> alternative hypothesis: true mean is greater than 12.6
> 95 percent confidence interval:
>  9.78716     Inf
> sample estimates:
> mean of x
>     12.25
```

결과를 살펴보면, 다음의 답을 얻을 수 있습니다.

2. 표본평균: 12.25, t 검정통계량: -0.25522
3. p-value: 0.5984 따라서 p-value가 유의수준 0.05보다 크므로 귀무가설을 기각할 수 없다. 즉, 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균시간은 12.6시간보다 크다는 통계적 근거가 충분하지 않다.
4. 신뢰수준 95%하에서 한국 사용자들의 한 달 스마트폰 영화 스트리밍 평균시간이 9.78보다 크다고 말할 수 있다.

1.2 2 표본 t 검정 (독립)

남자-여자 그룹의 영화 스트리밍 평균시간

2005년에 설립된 SPstream은 영화 스트리밍 서비스를 제공하고 있습니다. 2020년에 SPstream의 보고서에 따르면, 5G 네트워크를 이용하는 한국 사용자들의 남자 그룹와 여자 그룹의 한 달에 평균 영화시청 시간은 같았다라고 합니다. 다음은 이번달 5G 사용자 고객 중 남자 사용자 13명과 여자 사용자 10명을 무작위로 선택 한 후, 시청 시간을 기록한 것입니다.

- 남자 그룹

$$13.3, 6.0, 20.0, 8.0, 14.0, 19.0, 18.0, 25.0, 16.0, 24.0, 15.0, 1.0, 15.0$$

- 여자 그룹

$$22.0, 16.0, 21.7, 21.0, 30.0, 26.0, 12.0, 23.2, 28.0, 23.0$$

데이터를 사용하여 2023년 남자 여자 사용자 집단 간 한 달 스마트폰 영화 스트리밍 평균 시간이 차이가 나는지 유의수준 0.05 하에서 검정하세요. 단, 각 그룹의 시청시간 분포는 정규분포를 따르고, 두 그룹의 분산은 다르다고 가정한다.

1. 귀무가설과 대립가설을 설정하세요.
2. 각 그룹의 평균차이와 검정통계량 값을 구하세요.
3. p-value값을 구하고, 통계적 판단을 내려보세요.
4. 두 그룹의 평균 시간 차이에 대한 95% 신뢰 구간을 고르시오.

풀이 답안

위의 문제를 코드 하나로 풀 수 있습니다. 중요 포인트는 귀무가설과 대립가설을 잘 세울 수 있는가와 이것을 코드에 잘 반영하여 돌릴 수 있는가입니다.

1. 귀무가설과 대립가설을 설정하세요.

귀무가설: 남자 그룹(1)과 여자 그룹(2)의 평균 시청 시간은 같다.

$$\mu_1 = \mu_2$$

대립가설: 남자 그룹(1)과 여자 그룹(2)의 평균 시청 시간은 같지 않다.

$$\mu_1 \neq \mu_2$$

2, 3, 4번을 위해서 다음과 같이 코드를 작성하세요. 단, 문제의 조건에 따라서 각 그룹 분산이 같은지 다른지에 대하여 언급이 있을 경우 `var.equal` 옵션을 바꿔줄 수 있습니다. `t.test` 기본은 각 그룹 분산이 다르다로 설정이 되어있습니다.

```
male <- c(13.3, 6, 20, 8, 14, 19, 18, 25, 16, 24, 15, 1, 15)
female <- c(22, 16, 21.7, 21, 30, 26, 12, 23.2, 28, 23)

# 각 그룹 분산이 다른 경우
t.test(x = male, y = female, alternative = "two.sided", var.equal = FALSE)
```

```
>
>   Welch Two Sample t-test
>
> data:  male and female
> t = -2.8958, df = 20.989, p-value = 0.00865
> alternative hypothesis: true difference in means is not equal to 0
> 95 percent confidence interval:
>  -12.618000  -2.069692
> sample estimates:
> mean of x mean of y
>  14.94615  22.29000
```

```
# 각 그룹 분산이 같은 경우 t.test(x = male, y = female,
# alternative = 'two.sided', var.equal = TRUE)
```

결과를 살펴보면, 다음의 답을 얻을 수 있습니다.

2. 평균차이: 22.29 - 14.94615 = 7.345, 검정통계량: -2.8958
3. p-value: 0.00865. 따라서 p-value가 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 두 그룹의 평균 시청 시간이 같이 않다는 통계적인 근거가 충분하다.
4. 신뢰수준 95%하에서 두 그룹의 스마트폰 영화 스트리밍 시간 차이 평균은 (-12.61, -2.06) 사이에 존재한다고 말할 수 있다.

주의사항

만약 데이터를 사용하여 2023년 여자 사용자 그룹의 평균 시청 시간이 남자 사용자 그룹의 평균 시청 시간보다 큰 지 유의수준 0.05 하에서 검정하세요.라고 나왔다면?

대립가설은 다음과 같이 작성되고,

$$\mu_1 < \mu_2$$

코드는 다음과 같이 작성하세요. alternative의 방향 기준은 x 그룹이 y 그룹보다 큰 지 작은 지에 따라 결정됩니다.

```
t.test(x = male, y = female, alternative = "less", var.equal = FALSE)
```

```
>
>   Welch Two Sample t-test
>
> data:  male and female
> t = -2.8958, df = 20.989, p-value = 0.004325
> alternative hypothesis: true difference in means is less than 0
> 95 percent confidence interval:
>       -Inf -2.979867
> sample estimates:
> mean of x mean of y
>  14.94615  22.29000
```

1.3 2 표본 t 검정 (대응표본)

새로운 추천시스템의 효과

2005년에 설립된 SPstream은 영화 스트리밍 서비스를 제공하고 있습니다. 2023년에 SPstream은 새로운 추천 시스템을 도입했습니다. 이번 추천 시스템이 사용자들의 평균 시청시간을 증가시켰는지

도입전	도입후	나이	성별
14.3	15.3	35	남
16.3	17.0	64	남
15.3	16.8	61	남
15.3	14.2	44	여
14.6	14.1	30	남
15.0	14.7	41	여
14.8	13.3	31	남
15.3	14.1	48	여
15.3	13.1	35	남
15.8	12.5	44	남
14.9	16.4	51	여
17.3	15.9	52	남
16.5	13.5	37	여
14.5	15.9	61	여
14.3	15.3	44	남
15.2	12.6	52	여
14.1	16.2	31	남
17.6	13.4	58	남
14.3	13.6	22	남

검정하기 위하여, 5G 사용자 고객 중 19명을 무작위로 선택 한 후, 추천 시스템 도입 전과 도입 후 시청 시간을 기록한 것입니다.

데이터를 사용하여 새로운 추천 시스템은 영화 평균 영화시청 시간을 증가시켰는지 유의수준 0.05 하에서 검정하세요. 단, 시청변화 시간의 분포는 정규분포를 따른다고 가정한다.

1. 귀무가설과 대립가설을 설정하세요.
2. 각 그룹의 평균차이와 검정통계량 값을 구하세요.
3. p-value값을 구하고, 통계적 판단을 내려보세요.
4. 두 그룹의 평균 시간 차이에 대한 95% 신뢰 구간을 고르시오.

풀이 답안

위의 문제를 코드 하나로 풀 수 있습니다. 중요 포인트는 귀무가설과 대립가설을 잘 세울 수 있는가와 이것을 코드에 잘 반영하여 돌릴 수 있는가입니다.

1. 귀무가설과 대립가설을 설정하세요.

귀무가설: 추천 시스템 도입 전과 후의 평균 시청 시간은 같다.

$$\mu_{before} = \mu_{after}$$

대립가설: 추천 시스템 도입 후의 평균 시청 시간은 증가하였다.

$$\mu_{before} < \mu_{after}$$

이것은 다음과 같은 귀무가설, 대립가설로 바꿀 수 있습니다.

- 귀무가설

$$\mu_d = \mu_{before} - \mu_{after} = 0$$

- 대립가설

$$\mu_d = \mu_{before} - \mu_{after} < 0$$

2, 3, 4 번을 풀기 위해서는 다음과 같은 두 가지 방법이 존재합니다.

```
before <- c(14.3, 16.3, 15.3, 15.3, 14.6, 15, 14.8, 15.3, 15.3,
    15.8, 14.9, 17.3, 16.5, 14.5, 14.3, 15.2, 14.1, 17.6, 14.3)
after <- c(15.3, 17, 16.8, 14.2, 14.1, 14.7, 13.3, 14.1, 13.1,
    12.5, 16.4, 15.9, 13.5, 15.9, 15.3, 12.6, 16.2, 13.4, 13.6)
```

1 표본으로 변환 입력 (추천)

```
d <- before - after
t.test(d, mu = 0, alternative = "less")
```

```
>
>   One Sample t-test
>
> data:  d
> t = 1.5879, df = 18, p-value = 0.9351
> alternative hypothesis: true mean is less than 0
> 95 percent confidence interval:
>      -Inf 1.409366
> sample estimates:
> mean of x
> 0.6736842
```

2 표본 입력 후 대응표본 옵션 사용

```
t.test(before, after, alternative = "less", paired = TRUE)
```

```
>
>   Paired t-test
>
> data:  before and after
> t = 1.5879, df = 18, p-value = 0.9351
> alternative hypothesis: true mean difference is less than 0
> 95 percent confidence interval:
>      -Inf 1.409366
```

```
> sample estimates:
> mean difference
>       0.6736842
```


주의사항

1번 방법을 추천하는 이유는 이번 빅분기 실기 예제에서 μ_d를 지정해줬기 때문이다. R의 기본 μ_d는 `before - after`로 정의가 되어있지만, 위의 예제에서

$$\mu_d = \mu_{after} - \mu_{before}$$

로 정의한 경우,

$$\mu_{before} < \mu_{after}$$

는 다음과 같이 바꿀 수 있으며, 코드도 맞춰서 바꿔주면 된다.

$$\mu_d = \mu_{after} - \mu_{before} > 0$$

```r
d <- after - before
t.test(d, mu = 0, alternative = "greater")
```

```
>
>   One Sample t-test
>
> data:  d
> t = -1.5879, df = 18, p-value = 0.9351
> alternative hypothesis: true mean is greater than 0
> 95 percent confidence interval:
>  -1.409366       Inf
> sample estimates:
>  mean of x
> -0.6736842
```