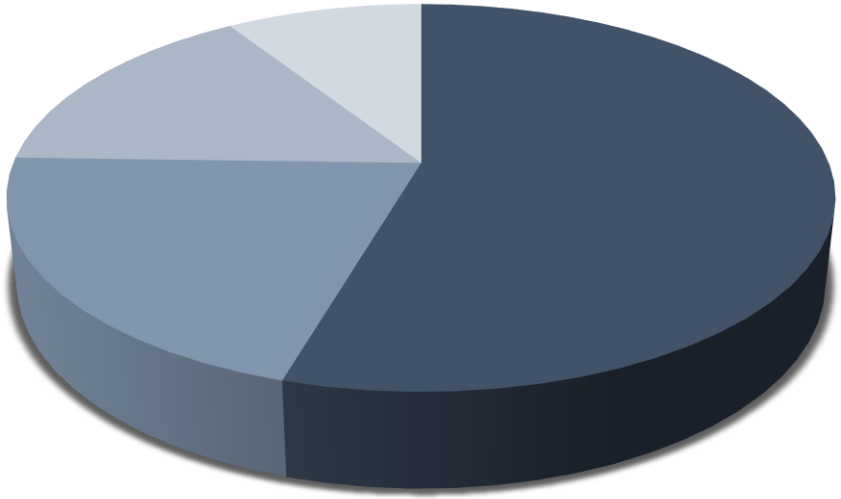
통계학 실습

13. 확률분포 (4)

Contents



Student–t Distribution

- Student–t Distribution은 직접 확률을 구하기보다 신뢰구간 혹은 가설검정 시에 사용되는 분포이다. 따라서 해당 분포의 t값을 구하는 것이 중요하다.

- 표본의 수가 적은 경우 신뢰도가 낮기 때문에 정규분포보다 한 단계 예측 범위가 넓은 분포를 사용하기 위해 Student–t Distribution을 이용한다. 따라서 넓은 종 형태의 대칭 그래프를 가지게 된다.

- Student–t Distribution에는 ‘자유도’가 존재한다.
  - 작은 표본 집단에서 하나를 제외하고 값들이 정해지면, 해당 하나의 값은 자동적으로 정해지는 것을 피하기 어렵다. 따라서 통상적으로 ‘자유도’는 (N–1)의 값을 가진다.

Student–t Distribution 내장 함수

함수 사용	설명	
dt(x, df)	‘D’ ensity	확률 밀도 함수 결과값 구하기 ex) P[X=x] = ?
pt(q, df, lower.tail)	‘P’ robability	누적 분포 함수의 누적확률 구하기 ex) P[X<q] = ?
qt(p, df, lower.tail)	‘Q’ uantile	누적 확률에 해당하는 분위값 구하기 ex) P[X<?] = p
rt(n, df)	‘R’ andom	분포 함수를 따르는 난수 생성

- x, q 분위수 벡터
- p 확률 벡터
- n 추출 난수 개수

- df 자유도
- lower.tail TRUE: P[X≤x], FALSE: P[X>x]

Student–t Distribution 문제 해결

- **어떤 10개의 표본을 조사하였을 때, 상위 1%가 나오는 t값을 구하여라.**

  Ⅰ. 표본이 10개이므로, 자유도는 90이다. (df = 9)

  Ⅱ. 상위 1%이므로 구하고자 하는 식은 P[X〉?] = 0.01 이다. (p = 0.01)

  Ⅲ. 누적 확률 값을 구해야 하므로 ‘qt’ 함수를 이용한다.

Student–t Distribution 함수 사용

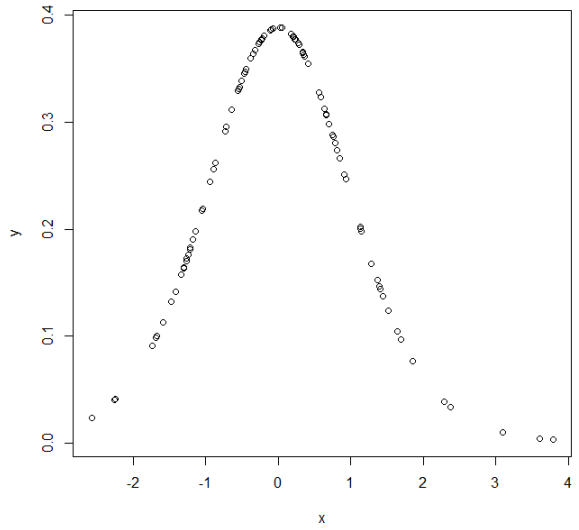
- **Q1. 어떤 10개의 표본을 조사하였을 때, 상위 1%가 나오는 t값은?**
  - abs(qt(p=0.01, df=9))
  - 상위의 t값이어야 하므로 절대값으로 양수를 취해주어야 한다.

- **Q2. 자유도가 7이고, 그래프의 면적 90%가 어떤 두 t값 사이에 존재할 때, 양쪽 t값은?**
  - c(qt(p=0.05, df=7), abs(qt(p=0.05, df=7)))
  - 면적 90%가 두 t값 사이에 존재하므로 각각 하위 5%, 상위 5% 지점이 된다.

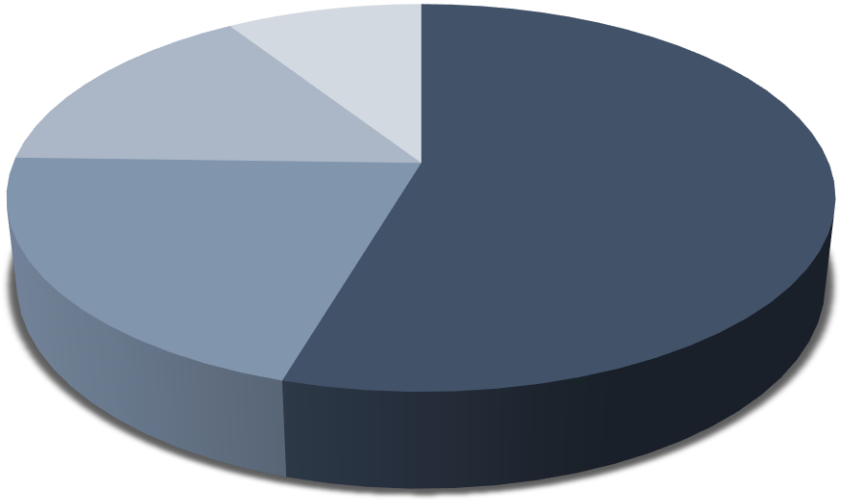
실습 문제

- 무작위 생성 함수를 통하여 자유도가 10인 Student–t Distribution을 따르는 수 100개를 생성한다.

- 생성된 데이터를 Student–t Distribution 확률 밀도 함수를 이용하여 그래프로 플로팅 해보자.



Contents



Hypothesis Testing

- Hypothesis Testing은 통계적 추측으로, 표본의 정보를 이용하여 모집단을 추측하고 해당 가설의 당위성 여부를 판정하는 과정이다.

- 통계적 가설에는 ‘귀무가설’과 ‘대립가설’이 존재한다.
  - 귀무가설 : 해당 집단 간의 차이가 없거나 의미가 존재하지 않는다.
  - 대립가설 : 해당 집단 간의 차이에는 의미가 존재한다.

Hypothesis Testing 예시

- Q. 어떤 단체에서 '학생들의 담배습관이 운동빈도와 관련이 있는가?' 에 대해 조사하기 위하여 학생들의 담배습관과, 운동습관을 조사하여 아래와 같이 데이터를 수집하였다. 학생들의 담배습관과 운동빈도가 독립인지 아닌지 가설을 통하여 확인해보자.

조사 (100명)	흡연	비흡연	전체
자주 운동	30	45	75
운동 안 함	5	20	25
전체	35	65	100

Hypothesis Testing 문제 해결

- 두 변수가 독립인지 알아보기 위하여 Chi-squared Test를 이용한다.
  - Hypothesis Testing은 원하는 분포의 'test' 함수를 이용한다.
  - 검사를 하기 위하여 주어진 표를 행렬로 저장한다.

```
> chisq.test(c)

        Pearson's Chi-squared test with Yates' continuity correction

data:  c
X-squared = 2.4762, df = 1, p-value = 0.1156
```

I. 결과를 보면, p-value가 0.1156으로 0.05보다 크다.
II. p-value가 0.05보다 큰 경우, 귀무가설을 기각할 수 없다.
III. 따라서 학생들의 담배습관과 운동빈도는 연관이 없는 독립이다.

실습 문제

- Q. ‘노란색을 좋아하는 사람이 녹색을 같이 좋아할까?’ 라는 의문에 조사를 시행하여 아래와 같이 수집하였다. Chi-squared Test를 이용하여 두 변수의 연관성에 대하여 확인해보자.

조사 (20명)	노란색	녹색	전체
좋아함	5	4	9
안 좋아함	3	8	11
전체	8	12	20

숙제

- SAS에서 숙제로 했던 레포트와 같이 가설검정 레포트를 작성합니다.

- MASS 패키지의 Cars93 데이터셋을 이용합니다.
  - 가설 : 차의 Price와 Length는 연관이 있다.

  - HINT) 데이터에 대해 행렬 등의 자료구조를 이용하여 테스트에 넣을 수 있도록 만들어 chisq.test() 함수를 이용하여 문제를 해결한다.

제출

- 아래 항목을 포함하여 작성 후 PDF로 만들어 제출합니다.
  - 숙제 페이지에 적힌 가설
  - 가설을 확인하기 위해 작성한 코드
  - 코드의 결과 화면
  - 코드를 통해 알 수 있는 가설에 대한 판단 결과

- 서식
  - [13주차][학번][이름]통계학실습

- 제출 이메일
  - gtsk623@gmail.com