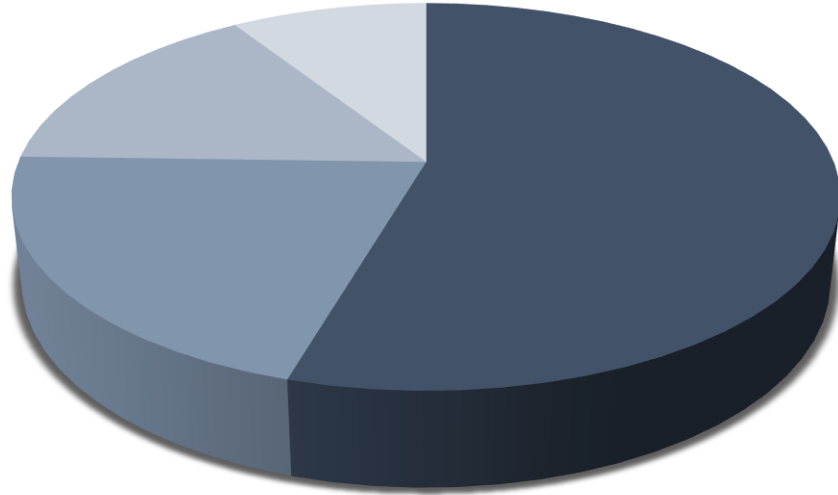




# 통계학 실습

## 12. 확률분포 (3)



- I. Chi-squared Distribution
- II. Normal Distribution

# Chi-squared Distribution

- Chi-squared Distribution은  $k$ 개의 서로 독립적이면서 표준정규분포를 따르는 확률변수들을 제공하고 합하였을 때의 분포이다.
- Chi-squared 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.

# Chi-squared Distribution 내장 함수

함수 사용	설명	
dchisq(x, df)	'D'ensity	확률 밀도 함수 결과값 구하기 ex) $P[X=x] = ?$
pchisq(q, df, lower.tail)	'P'robability	누적 분포 함수의 누적확률 구하기 ex) $P[X \leq q] = ?$
qchisq(p, df, lower.tail)	'Q'uantile	누적 확률에 해당하는 분위값 구하기 ex) $P[X \leq ?] = p$
rchisq(n, df)	'R'andom	분포 함수를 따르는 난수 생성

- x, q      분위수 벡터
- p      확률 벡터
- n      추출 난수 개수

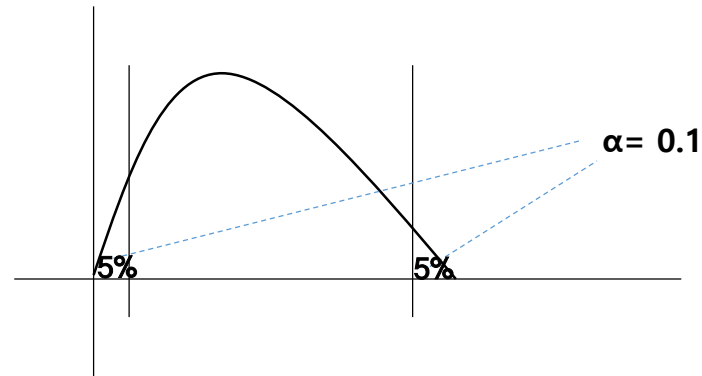
- df      자유도
- lower.tail      TRUE:  $P[X \leq x]$ , FALSE:  $P[X > x]$

## Chi-squared Distribution 문제 해결

- 자유도가 7인 Chi-squared Distribution에서 95% 백분위수를 구하여라.
  - I. 백분위수가 95%이면, 하위로부터 95%를 차지하는 위치이다.
  - II. 따라서 구하고자 하는 식은  $P[X \leq ?] = 0.95$  이다.
  - III. 분위값을 구해야 하므로 'qchisq' 함수를 이용한다.

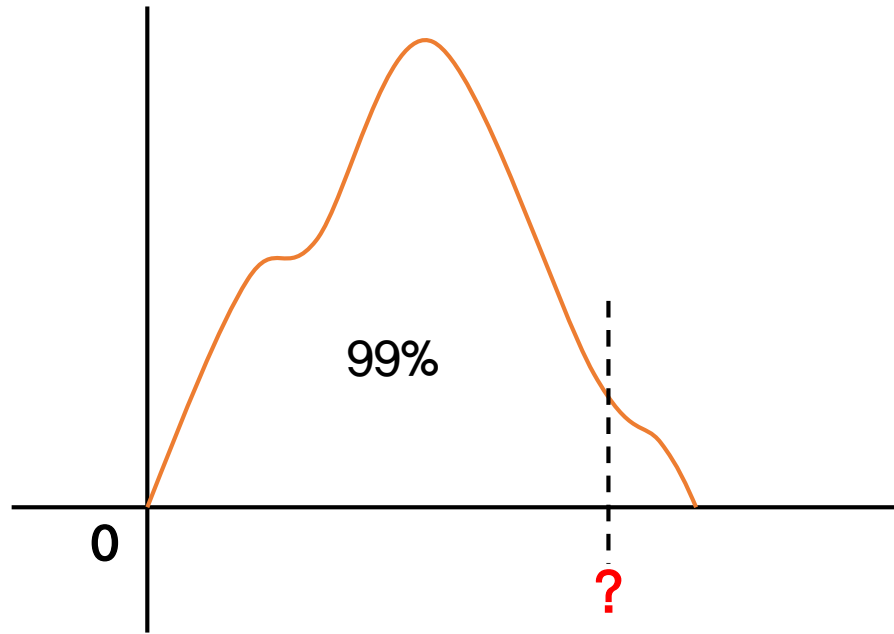
# Chi-squared Distribution 함수 사용

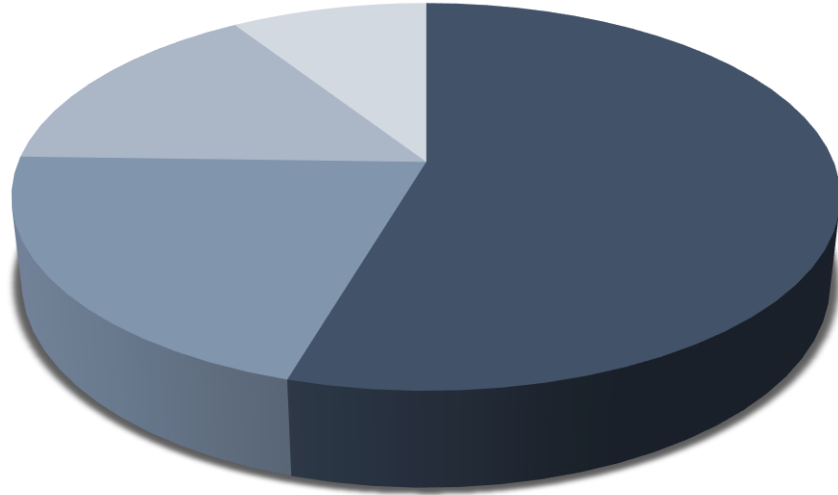
- Q1. 자유도가 7인 Chi-squared Distribution의 백분위수 95%는?
  - `qchisq(p=0.95, df=7)`
  - 결과 값 이하의 값은 전체의 하위 95%에 속하는 값이다.
- Q2. 표본이 8개이고,  $\alpha = 0.1$  이다. 그래프의 양쪽  $X^2$  값은?
  - `qchisq(c(0.05, 0.95), df=7)`
  - 표본이 8개이므로, 자유도는  $8-1=7$ 이다.
  - 양쪽의 합이 10%이므로, 그래프에서 하위 5%와 상위 5% 지점이 된다.



## 실습 문제

- 어느 공장에서 새로운 제품을 생산하고 표본으로 30개를 추출하였다. 생산된 제품의 결함도가 Chi-squared Distribution을 이룬다고 가정할 때, 이 Chi-squared Distribution의 백분위수 99%는?





- I. Chi-squared Distribution
- II. Normal Distribution



# Normal Distribution

- 어떤 분포든 수많이 반복하면, 표본 평균의 분포가 결국 Normal Distribution에 가까워진다.
- 표준정규분포는 평균이 0이고 표준편차가 1인 대칭된 종 모양 그래프를 띈다.

# Normal Distribution 내장 함수

함수 사용	설명	
<code>dnorm(x, mean=0, sd=1)</code>	'D'ensity	확률 밀도 함수 결과값 구하기 ex) $P[X=x] = ?$
<code>pnorm(q, mean=0, sd=1, lower.tail)</code>	'P'robability	누적 분포 함수의 누적확률 구하기 ex) $P[X \leq q] = ?$
<code>qnorm(p, mean=0, sd=1, lower.tail)</code>	'Q'uantile	누적 확률에 해당하는 분위값 구하기 ex) $P[X \leq ?] = p$
<code>rnorm(n, mean=0, sd=1)</code>	'R'andom	분포 함수를 따르는 난수 생성

- `x, q`      분위수 벡터
- `p`      확률 벡터
- `n`      추출 난수 개수

- `mean`      평균 (초기값 = 0)
- `sd`      표준편차 (초기값 = 1)
- `lower.tail`      TRUE:  $P[X \leq x]$ , FALSE:  $P[X > x]$

## Normal Distribution 문제 해결

- 어떤 회사에서 생산하는 기계의 수명이 평균 900일이고, 표준편차는 70일인 Normal Distribution을 따를 때, 기계의 수명이 1000일을 넘길 확률은?
  - I. 평균은 900일, 표준편차는 70일이다. (mean = 900, sd = 70)
  - II. 1000일 이상이므로 구하고자 하는 식은  $1 - P[X \leq 1000]$  이다.
  - III. 누적 확률 값을 구해야 하므로 'pnorm' 함수를 이용한다.

# Normal Distribution 함수 사용

- Q1. 평균 900, 표준편차 70인 정규분포를 따르는 기계의 수명이 1000일을 넘길 확률은?
  - `pnorm(q=1000, mean=900, sd=70, lower.tail=FALSE)`
- Q2. 어느 서비스의 가입자들의 가입 기간이 평균 5년, 분산 4년의 정규분포를 가진다고 한다. 이 때, 해당 서비스에 10년 이상 가입해 있는 가입자의 비율은?
  - `pnorm(q=10, mean=5, sd=sqrt(4), lower.tail=FALSE)`
  - 분산은 표준편차의 제곱이다.

## 실습 문제

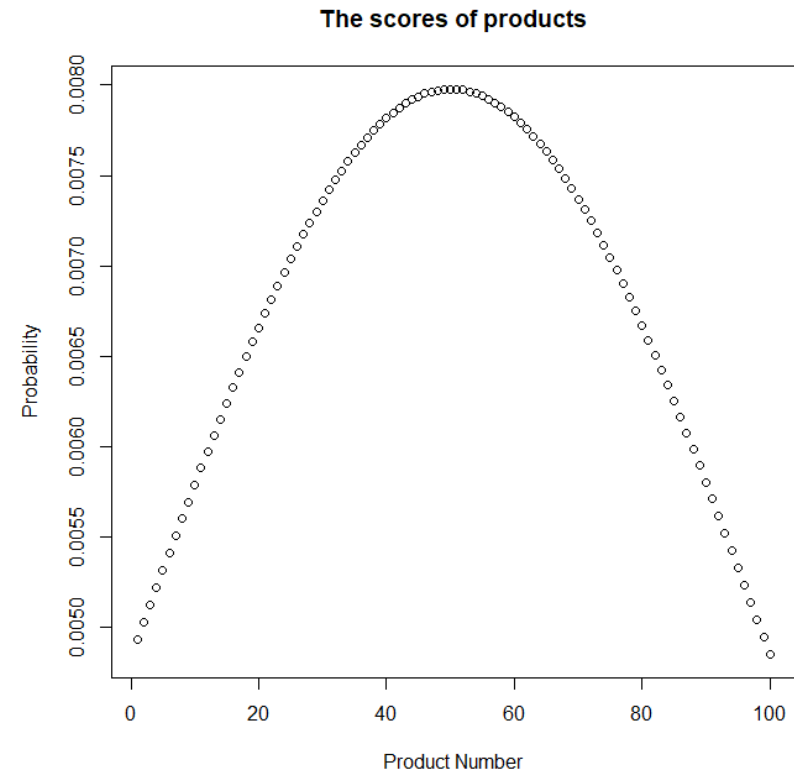
- Q. 아래는 어느 고등학교 3학년 학생들의 수학 성적표이다. 해당 학생들의 점수가 표준편차 10을 가지는 정규분포를 따른다고 한다. 그래프를 통하여 확인하자.
- HINT) 점수에 해당하는 확률 값은 'dnorm' 함수를 통하여 구한다.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
79	48	35	55	36	42	63	31	54	54	47	53	53	62	50
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
54	48	57	43	42	63	44	52	53	52	65	56	57	47	67

- $x \leftarrow c(79, 48, 35, 55, 36, 42, 63, 31, 54, 54, 47, 53, 53, 62, 50, 54, 48, 57, 43, 42, 63, 44, 52, 53, 52, 65, 56, 57, 47, 67)$

- 실생활의 대부분의 상황 혹은 수집된 결과 등은 정규분포로 이루어져 있다. 예를 들어 학생들의 시험 점수나, 물건의 등급 분포 등에서 많은 표본이 모이면, 결국 정규분포로 향하게 된다.
- 정규분포의 예시를 찾아서, 해당 예시에 대한 데이터를 만들어 분석하고 그래프를 만들어보자.
  - 데이터 표본 개수는 최소 1000 개 이상으로 할 것.
  - 무작위 추출을 이용하여도 된다.
- Example) 한 공장에서 물품을 1000 개를 찍어내었다. 각 제품의 상태에 따라 100점 만점으로 점수를 매겼다. 해당 점수를 매긴 분포가 정규 분포를 따르는지 확인한다.

- 아래의 항목을 포함하여 작성 후 제출
  - 1. 선택한 정규분포의 예시
  - 2. 해당 예시에 대한 예시 데이터
  - 3. 예시 데이터를 이용하여 그린 그래프
  - 4. 작성한 코드
- 제출 서식
  - [12주차][학번][이름]통계학실습
- 제출 이메일
  - gtsk623@gmail.com



〈그래프 예시〉