분포와 검정

데이터과학을 위한 통계(1~3장)

주요개념

- 평균
- 분산
- 표준편차
- 표본분포
  - Student-분포
  - 카이 제콥분포
  - F-분포
- 검정
  - 우연이다, 아니다.

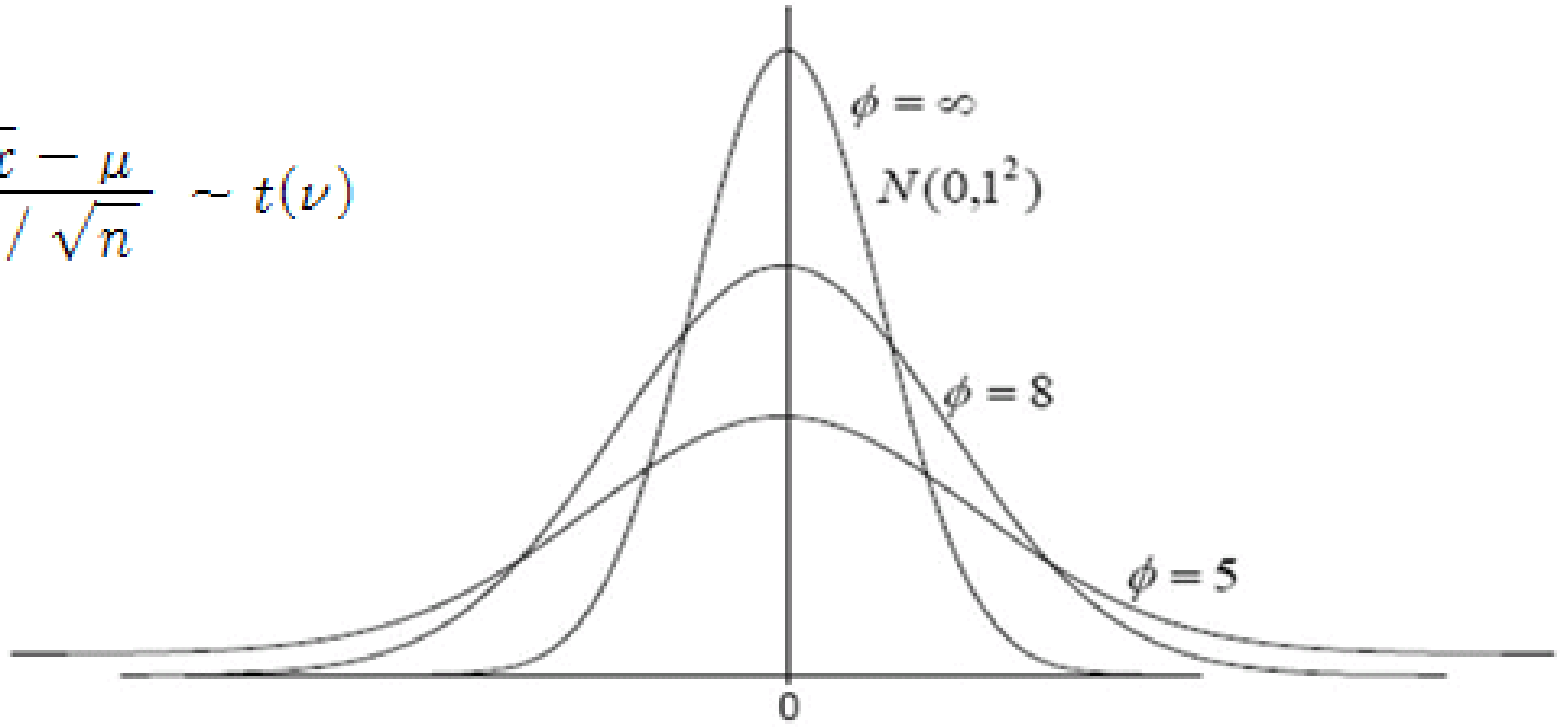
평균

분산

표준편차

t-분포

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t(\nu)$$



자유도(ν)가 ∞ 이면 t분포는 정규분포와 일치한다.

t-검정

- 독립검정
- 두개의 그룹 간의 통계량 차이분석
- 동일한 집단의 전-후 통계량 차이분석

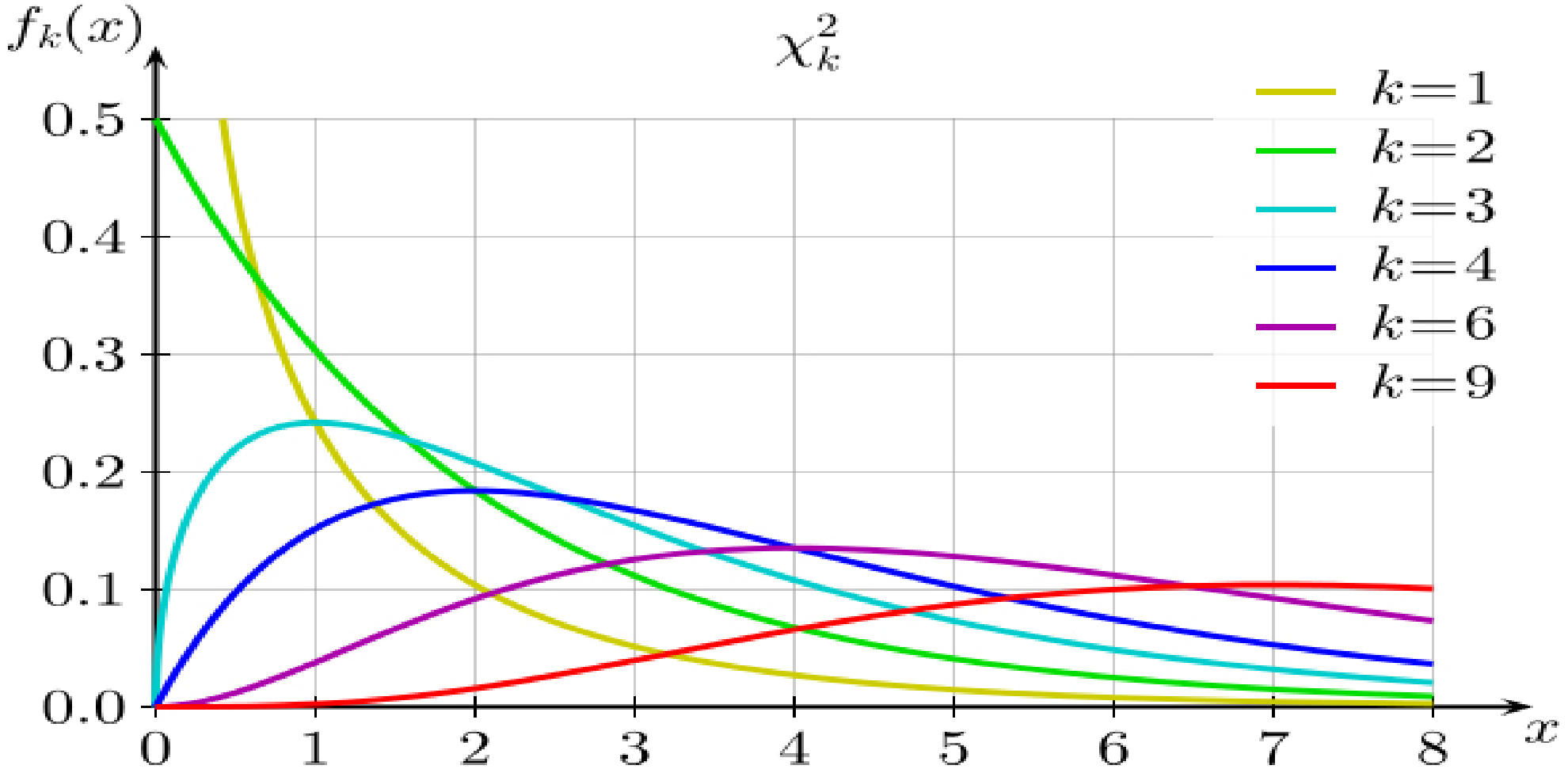
독립표본

- mtcars 통계량 확인
- 자동차 기어 종류에 따라 mpg 차이 확인
  - 오토 (0) , 수동(1)
- t.test(mpg~am,data=mtcars)
-

대응표본

- 중간/기말 고사
- mid = c(16,20,21,22,23,22,27,25,27,28)
- final = c(19,22,24,24,25,25,26,26,28,32)
- t.test(mid,final,paired = TRUE)

카이 제곱분포



	남	여
갤럭시	72	28
아이폰	55	67

$$\chi^2 = \sum\sum \frac{(\text{관측된 도수} - \text{기대도수})^2}{\text{기대도수}}$$

기대도수	남	여
갤럭시	$\left(\frac{127}{222} \times \frac{100}{222}\right) \times 222 = 57.2$	$\left(\frac{95}{222} \times \frac{100}{222}\right) \times 222 = 42.8$
아이폰	$\left(\frac{127}{222} \times \frac{122}{222}\right) \times 222 = 69.8$	$\left(\frac{95}{222} \times \frac{122}{222}\right) \times 222 = 52.2$

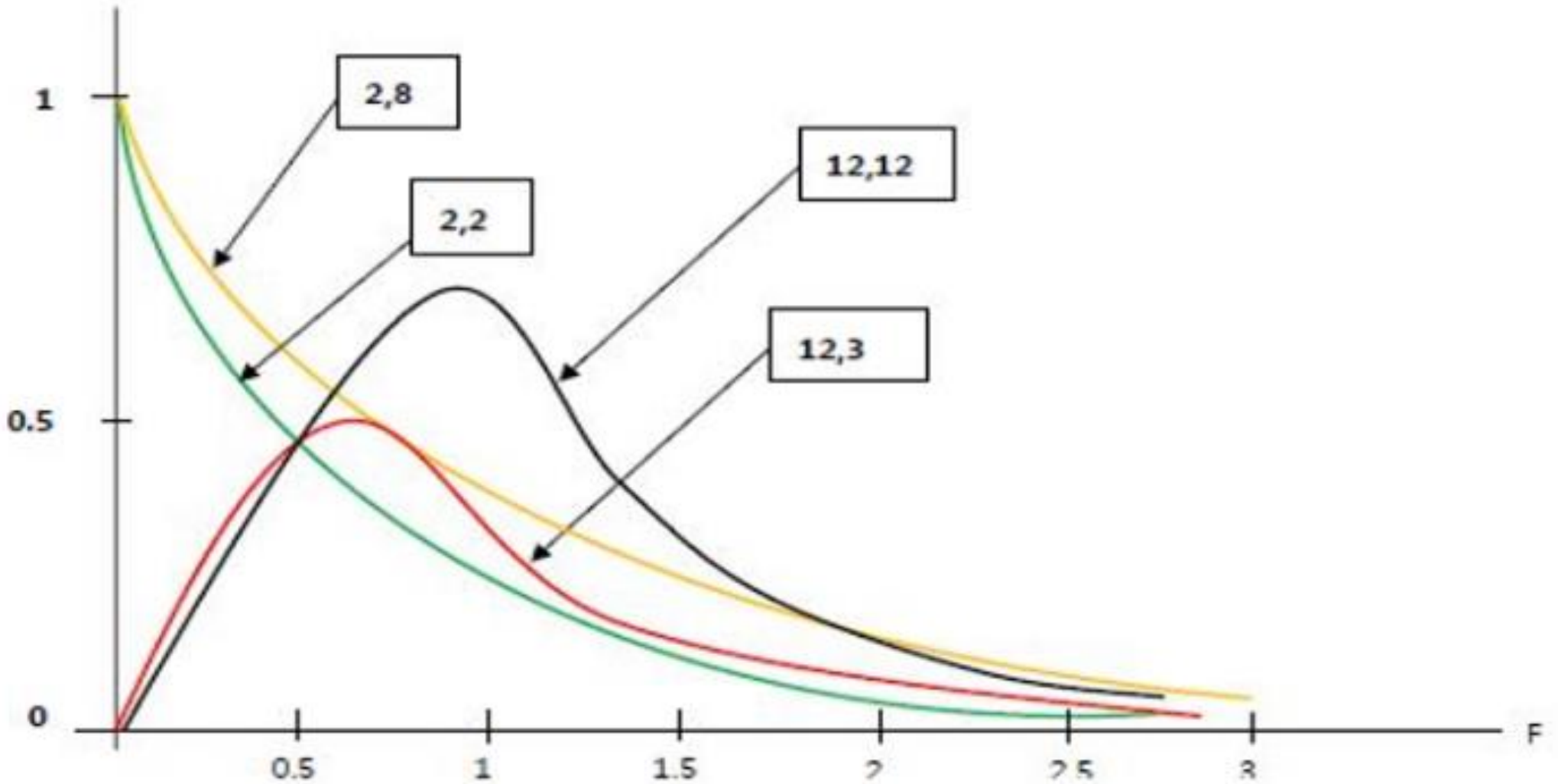
카이제곱 검정

- 두 요인간에 관계가 있는지 확인
- 독립성 검정
- 범주형 데이터 차이 분석

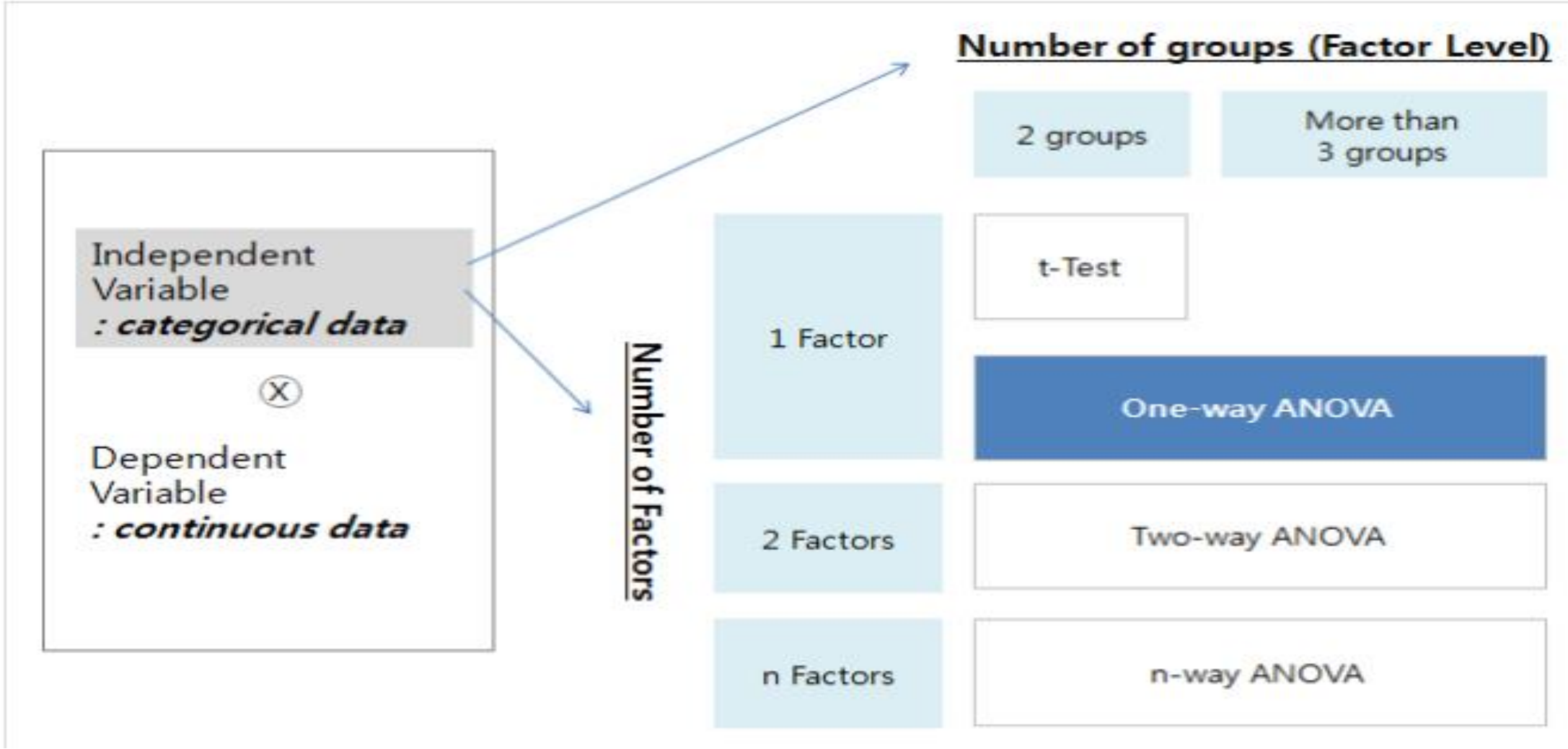
예제

- a = read.csv('./data/survey.csv')
- names(a) = c('position','ans6','ans7’)
- t1 = table(a$position,a$ans6)
- names(t1) = c('position','ans6','freq')
- t2= as.data.frame(t1)
- library(ggplot2)
- ggplot(t2,aes(position,freq,fill=ans6)) + geom_bar(stat='identity',position = 'fill')
- chisq.test(t1)

F-분포



ANOVA 분석 (분산분석)



ANOVA 분석 (분산분석)

n-way ANOVA	Model	Description
one-way ANOVA	y ~ x1	y is explained by x1 only
two-way ANOVA	y ~ x1 + x2	y is explained by x1 and x2
two-way ANOVA	y ~ x1 * x2	y is explained by x1, x2 and the interaction between them

F 검정

- F값 : 설명할 수 있는 부분의 평균 제곱 합 / 설명할 수 없는 부분의 평균 제곱 합의 비

- 전체 평균에 대한 실측치의 편차 제곱 합 A
- 전체 평균에 대한 그룹평균치의 편차 제곱 합 B
- 그룹평균에 대한 실측치의 편차 제곱 합 C
- $A^2 = B^2 + C^2$

일원배치 분산분석

- iris
- names(iris)

- result = aov(iris$Sepal.Width~iris$Species,iris)
- summary(result)
- bartlett.test(iris$Sepal.Width~iris$Species,iris)

이원배치 분산분석

- attach(score.df)
- r = aov(score_stats~gender.fac*class.fac)
- summary(r)
- plot(score_stats ~ gender.fac, main="box plot by gender")
- plot(score_stats ~ class.fac, main="box plot by class")
- interaction.plot(gender.fac, class.fac, score_stats, main="interaction effect plot")
- interaction.plot(class.fac, gender.fac, score_stats, main="interaction effect plot")

상관관계

- df = read.csv('http://goo.gl/HKnl74')
- plot(overall~rides)
- 상관계수 :
- cor(overall,rides,use='complete.obs',method='pearson')
- cor(df[,4:8])
- cor.test(overall,rides)
- x = cor(df[,4:8])
- plot(df[,4:8])
- corrplot(x)