데이터로 배우는 통계학

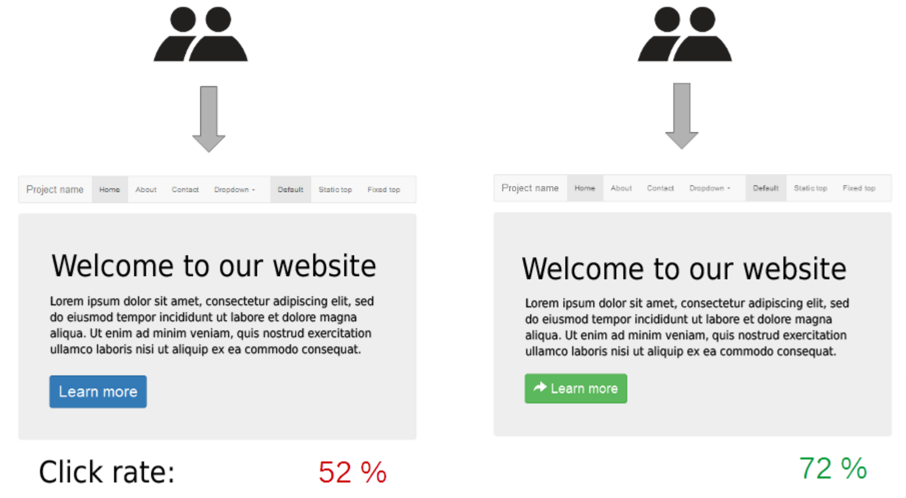
자연과학대학 통계학과
장원철 교수

알고 보면 쉬운 두 집단의 비교

1. A/B test란?

A/B test

- 버튼의 디자인만 다른 두 가지 버전의 웹사이트를 방문자에게 무작위로 보여준 후 디자인의 효용성(클릭 비율)을 측정하는 문제를 생각해보자.



[웹사이트 디자인에서 A/B test의 예]
(Wikipedia)

기타 출처 #1

A/B test

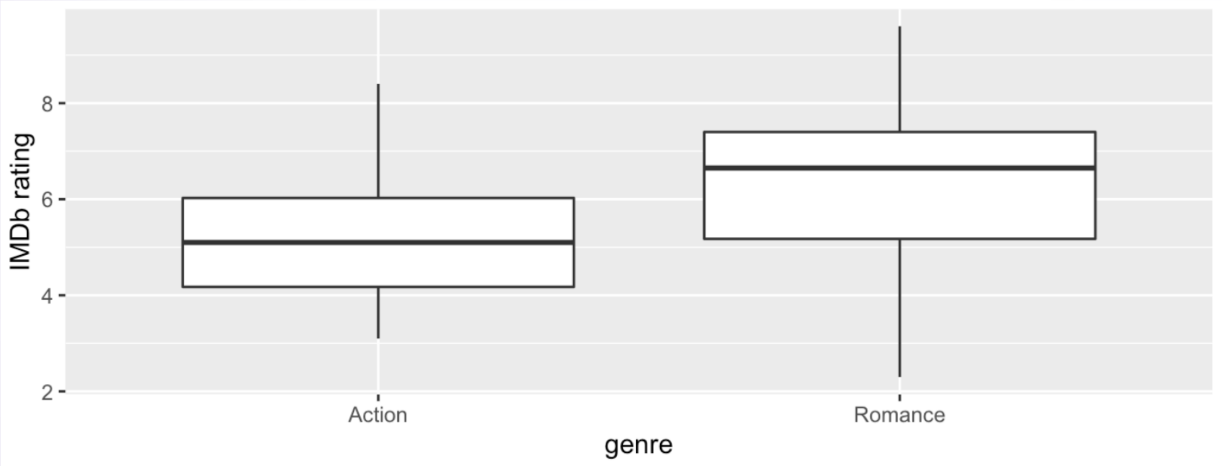
- 웹페이지나 앱에서 서로 다른 2개의 UX의 선호도를 평가하기 위해 사용하는 방법을 의미한다.
- 일반적으로 임의로 사용자(방문자)를 각각의 UX에 할당한 후 선호도를 조사한다.
- 통계학에서 2 표본 검정과 같은 문제로 생각할 수 있다!
- 앞의 예제에서는 두 집단의 비율의 차이가 있는지 여부를 검정하는 문제로 생각하면 된다.

액션 영화와 로맨스 영화중 어느 장르가 더 인기가 있을까?

- IMDb는 영화, TV 드라마에 관한 정보(출연 배우, 줄거리, 평점)를 제공하는 온라인 데이터베이스이다.
- IMDb를 이용하여 일반적으로 액션 영화와 로맨스 영화 어느 장르의 평점이 높은지 알아보자.
- 자료는 R package ggplot2movies에 있는 58,788개의 IMDb 평점을 이용하자.
- 우리는 이중 임의로 68편(액션 영화 32편, 로맨스 영화 36편)을 골라서 어느 장르가 더 평점이 높은지 알아보고자 한다.

로맨스 vs 액션

- Boxplot으로 두 그룹의 평점을 비교해 보자.



로맨스 vs 액션

- 가설검정의 절차를 이용해서 분석해보자. 먼저 귀무가설과 대립가설은 다음과 같다.
  - 귀무가설: IMDb에서 로맨스와 액션의 평균 평점이 같다.
  - 대립가설: IMDb에서 로맨스와 액션의 평균 평점이 다르다.
- 검정통계량은 일반적으로 **(추정치- 귀무가설하에서 추정하고자 하는 모수의 값)/표준오차**로 주어진다.
- 이 경우 추정하고자 하는 모수는 전체 데이터베이스에서 두 그룹의 평균 평점의 차이이며 추정치는 표본 평균의 차이 $(= \overline{x}_R - \overline{x}_A)$를 사용할 수 있다.

로맨스 vs 액션

- 귀무가설하에서는 두 그룹의 평균 rating의 차이는 0이다.
- 표준오차의 경우 다음 두 가지 경우를 고려할 수 있다.
  - 두 집단의 분산이 같은 경우:
  - 두 집단의 분산이 다른 경우
- P값의 계산
  - 순열검정을 이용하는 방법
  - t-분포를 이용하는 방법

표본오차의 추정

- 두 집단의 분산이 다른 경우 다음 사실을 이용한다
- $Var(\overline{x}_R - \overline{x}_A) = Var(\overline{x}_R) + Var(\overline{x}_A) = s_R^2/n_R + s_A^2/n_A$ 여기서 n_R, n_A는 로맨스와 액션 그룹의 표본크기, s_R^2, s_A^2는 로맨스와 액션 그룹의 표본분산을 나타난다.
- 따라서 $\overline{x}_R - \overline{x}_A$의 표본오차는 $\sqrt{s_R^2/n_R + s_A^2/n_A}$

포본오차의 추정

- **두 집단의 분산이 같은 경우** 두 그룹의 변동을 모두 고려해서 하나의 추정치를 구한다.

$$s_p^2 = \frac{\sum_{i=1}^{n_R}(x_{Ri} - \bar{x}_R)^2 + \sum_{i=1}^{n_A}(x_{Ai} - \bar{x}_A)^2}{n_R + n_A - 2} = \frac{(n_R - 1)s_R^2 + (n_A - 1)s_A^2}{n_R + n_A - 2}$$

- 따라서 $\bar{x}_R - \bar{x}_A$의 표본오차는 $\sqrt{s_p^2/n_R + s_p^2/n_A} = s_p\sqrt{1/n_R + 1/n_A}$
- 일반적으로 두 집단의 분산이 다르다고 보는 것이 타당하고 선행연구 결과나 분산이 같다는 가정이 합리화될 수 있는 경우에 사용하는 것이 타당하다. 즉 데이터를 관측한 후에 어떤 방법을 사용할지 결정하는 것이 아니라 미리 분석방법을 결정해야 한다.

검정통계량의 분포

- **분산의 추정방법과 관계없이 모집단이 정규분포를 따른다면 귀무가설하에서 검정통계량은 t -분포를 따른다.**
- **표본오차의 추정방법이 달라지는 경우 t-분포의 자유도가 달라진다.**
  - 두 그룹의 분산이 같은 경우: $n_R + n_A - 2$
  - 두 그룹의 분산이 다른 경우: 근사식을 사용하여 자유도가 정수가 아닌 경우가 나올 수 있다. R과 같은 프로그램을 사용하여 구할 수 있고 일반적으로 두 그룹의 표본크기 중 최솟값보다 크다고 생각할 수 있다.

P값 계산

- **2 표본검정에 p값은 다음과 같은 방법으로 귀할 수 있다.**
  - 모집단이 정규분포를 따른다는 가정하에 귀무가설하에서 검정통계량이 t-분포를 따른다는 사실 이용
  - 모집단의 분포에 대한 가정없이 순열검정을 이용하여 귀무가설하에서 검정통계량의 표본분포를 근사

다시 로맨스 vs 액션

- 먼저 검정통계량을 계산해 보자. 분산이 같다고 가정할 이유가 전혀 없기 때문에 분산이 다른 경우의 검정통계량을 사용하자.

$$\frac{\bar{x}_R - \bar{x}_A}{\sqrt{s_R^2/n_R + s_A^2/n_A}} = \frac{6.32 - 5.28}{\sqrt{1.61^2/36 + 1.36^2/32}} = 2.906$$

- 이론적으로 계산한 t-분포의 자유도는 65.85이며 p값은 0.002이다.
- 순열검정을 통해서도 비슷한 p값을 얻을 수 있다.
- 결론은 귀무가설을 기각한다. 즉 IMDb 데이터 베이스에서 로맨스와 액션의 평균 평점은 다르다고 말할 수 있다.

오늘의 강의 요점

- A/B test 란?
- 두 표본의 비교
  - 두 그룹의 분산이 다른 경우
  - 두 그룹의 분산이 같은 경우

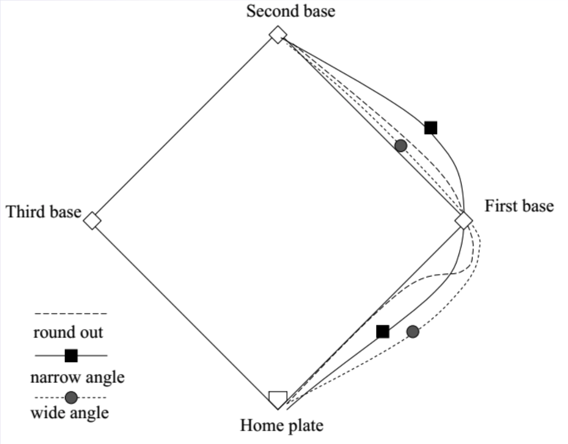
출처

#1 https://en.wikipedia.org/wiki/A/B_testing

알고 보면 쉬운 두 집단의 비교

2. 쌍체비교

1루는 어떻게 돌아야 하나?



[1루를 좁은 각도와 넓은 각도로 도는 경우]
(Design and Analysis of Experiments with R. p.138)

- 안타를 친 후 2루에 도달하기 위해서 1루를 큰 각도로 도는 것이 빠를까? 아니면 작은 각도 도는 것이 빠를까?
- 이 문제를 가설검정으로 고려한다면 귀무가설과 대립가설은 무엇인가?

기타 출처 #1

1루는 어떻게 돌아야 하나?

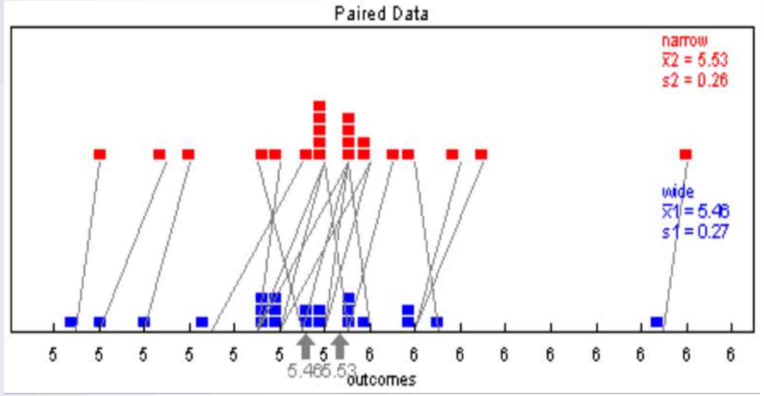
- 1970년 Woodward은 본인의 석사 논문에서 이 문제의 해답을 얻기 위해 다음과 같은 실험을 하였다.
- 먼저 주자가 홈플레이트에서 35피트를 지나는 점을 출발점으로 하고 2루 베이스에서 15피트 떨어진 점을 통과할 때까지의 시간을 측정하기로 하고 22명의 피실험자를 선발하였다.
- 실험을 공정하게 하기 위해서 각 주자들이 좁은 각도와 넓은 각도로 각각 뛰게 하고 뛰는 순서는 임의로 정하였다.
- 두 번의 달리기 사이에는 휴식 시간이 주어졌다.

1루는 어떻게 돌아야 하나?

- 22명의 기록은 다음과 같다.

피실험자	1	2	3	4	5	6	7	8	9	10	…
넓은 각도	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
좁은 각도	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
차이	-0.05	-0.05	0.10	0.10	0.15	-0.05	0.05	0.15	0.15	0.10	…

1루는 어떻게 돌아야 하나?



[점그림을 이용한 각도별 주루기록]
(www.isi-stats.com)

- 위의 그림은 실제 관측치를 dotplot을 이용하여 표시하였다.
- 두 그룹의 요약치는 다음과 같다.
- 좁은 각도: 평균 5.53, 표준편차 0.26
- 넓은 각도: 평균 5.46 표준편차 0.27

기타 출처 #2

1루는 어떻게 돌아야 하나?

- 앞에서 배운 2표본 t-검정을 이용해서 분석해보자. 같은 사람들이 2번 뛰었기 때문에 각 그룹의 분산을 같다고 가정하더라도 무방해 보인다. 이 경우 검정통계량은

$$\frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{5.53 - 5.45}{0.265\sqrt{1/22 + 1/22}} = 0.9338$$

- 귀무가설하에서 검정통계량의 분포는 자유도가 42 (=22+22-2)인 t-분포를 따른다.

$$\Pr(T_{42} > 0.9338) + \Pr(T_{42} < -0.9338) = 2 \times \Pr(T_{42} > 0.9338) = 0.006$$

1루는 어떻게 돌아야 하나?

- 하지만 2 표본 검정에서는 두 개의 표본이 서로 독립이라는 가정이 있는데 이 예제에서는 한 사람이 두 번 달리기를 했기 때문에 이렇게 쌍으로 주어진 데이터는 두 그룹이 독립이 아니다.

- 이 경우 개별 달리기 기록의 차이(= 넓은 각도기록 - 좁은 각도 기록)를 구한 후 차이들의 평균(= 검정통계량)이 아주 극단적인 값을 가지는 경우 귀무가설을 기각한다고 생각하면 자연스러운 가설검정 절차가 될 것이다.

피실험자	1	2	3	4	5	6	7	8	9	10	…
넓은 각도	5.50	5.70	5.60	5.50	5.85	5.55	5.40	5.50	5.15	5.80	…
좁은 각도	5.55	5.75	5.50	5.40	5.70	5.60	5.35	5.35	5.00	5.70	…
차이	-0.05	-0.05	0.10	0.10	0.15	-0.05	0.05	0.15	0.15	0.10	…

1루는 어떻게 돌아야 하나?

- 차이의 평균은 0.075, 표준편차는 0.088이다. 이 경우 검정 통계량의 값은 다음과 같다.

$$\mathrm{T} = \frac{0.075 - 0}{0.088/\sqrt{22}} = 3.998$$

- 귀무가설하에서 검정통계량은 자유도가 21인 t-분포를 따른다. 따라서 p값은

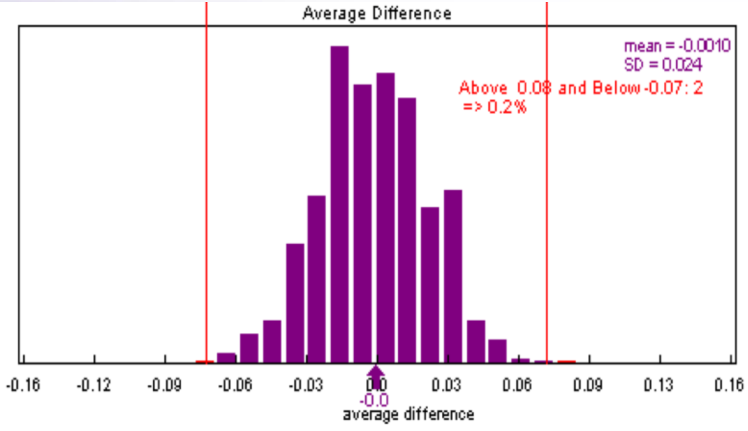
$$\Pr(\mathrm{T}_{21} > 3.998) + \Pr(\mathrm{T}_{21} < -3.998) = 2 \times \Pr(\mathrm{T}_{21} > 3.998) = 0.006$$

- 즉 1루를 크게 도는 것이 2루에 도달하는 시간을 단축할 수 있다!

쌍체비교에서 순열검정

- 만약 이 자료를 이용해서 순열검정(permutation test)를 한다면 어떤 방법으로 할 수 있을까?
- 두 그룹의 비교에서는 개별 데이터의 그룹 멤버쉽을 재배분한 후 검정통계량의 값을 계산하고 이러한 과정을 반복해서 검정 통계량의 분포를 제시하였다.
- 베이스러닝 예제의 경우 귀무가설은 1루를 도는 각도의 크기에 관계없이 2루에 도달하는 시간은 똑같다.
- 그렇다면 주루기록의 차이를 "넓은 각도의 기록 – 좁은 각도의 기록"로 사용하거나 "좁은 각도의 기록 – 넓은 각도의 기록"으로 사용하던지 관계없이 똑같은 결론에 도달해야 한다.

쌍체비교에서 순열검정

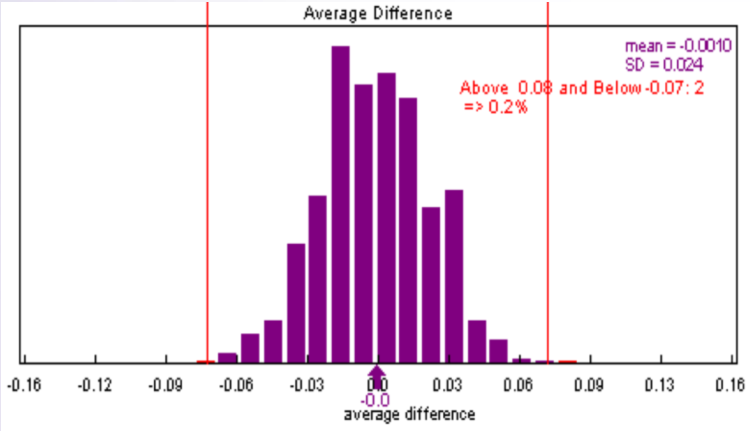


[순열을 이용한 주루기록차이의 표본분포]
(www.isi-stats.com)

- **즉 기록의 차이의 부호를 그대로 두거나 바꾸는 것을 고려할 수 있다. 따라서 전체 가능한 순열의 숫자는 2^{22}이다.**

기타 출처 #3

쌍체비교에서 순열검정

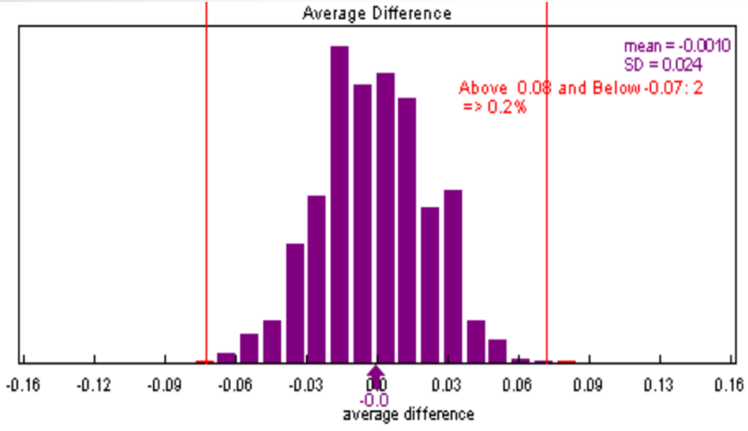


[순열을 이용한 주루기록차이의 표본분포]
(www.isi-stats.com)

- 전체 가능한 순열의 개수가 너무 많기 때문에 임의로 순열을 1,000개 생성하여 차이의 표본분포를 구한 결과가 위의 그림이다.

기타 출처 #3

쌍체비교에서 순열검정



[순열을 이용한 주루기록차이의 표본분포]
(www.isi-stats.com)

- 여기서 실제 관측된 경우인 0.075보다 같거나 더 큰 경우를 0.001이고 양측검정을 고려하기 때문에 p값은 0.002이다.

기타 출처 #3

오늘의 강의 요점

- 쌍체비교와 이표본 검정의 차이점
- 쌍체비교에서의 순열검정

출처

#1 J. Lawson, (2014), Design and Analysis of Experiments with R, Champman and Hall

#2~3 http://www.rossmanchance.com/ISIapplets.html

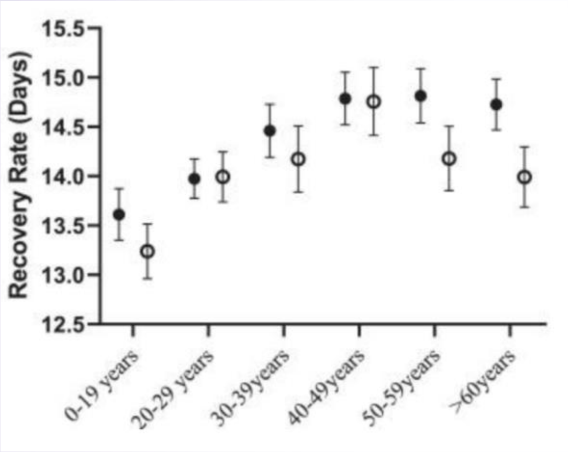
알고 보면 쉬운 두 집단의 비교

3. 연령대별로 코로나19 회복 기간이 차이가 있나요? - 분산분석

코로나19에서 회복하는 데 걸리는 시간은 연령별로 차이가 있는가?

- 코로나 19의 사망률이 일반적으로 남성이 더 높고 연령과도 비례한다고 알려져 있다.
- 그렇다면 코로나 19에서 회복되는 기간은 연령별로 차이가 날까?
- 이스라엘의 텔아비브 대학의 연구자들이 코로나 환자 5,769 명의 자료를 이용하여 위의 질문에 답변을 하였다.
- 이 자료는 20세부터 59세 사이의 3,370명의 남자와 2,399 명의 여자로 이루어져 있다.

3개 이상의 그룹에서 평균의 비교



[코로나 회복 기간에 대한 성/연령 별 상자 그림]
(ScienceDirect)

- 왼쪽 그림은 성/연령별 코로나 19 회복 기간에 대한 상자 그림이다.
- 전 연령층에 걸쳐 남성이 여성에 비해 회복 기간은 상대적으로 짧은 것으로 보이고 연령대별로도 회복 기간의 차이는 있어 보인다.
- 이 차이가 실제 통계적으로 혹은 의학적으로 유의미한 차이인가?

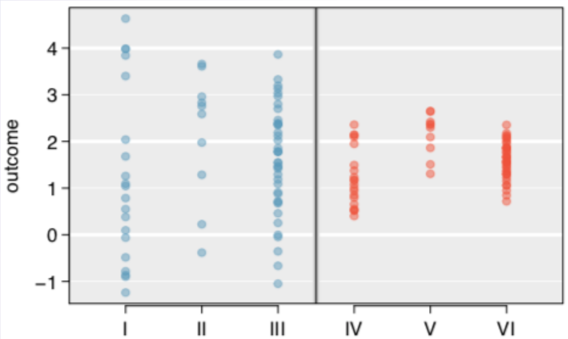
기타 출처 #1

3개 이상의 그룹에서 평균의 비교

- 이 질문에 답변하기 위해 먼저 우리는 남성 3,370명이 연령대별로 코로나 회복 기간에 차이가 있는지를 알아보자.

연령	0-19	20-29	30-39	40-49	50-59	60이상
평균 (표준오차) 회복 기간	13.61 (0.26)	13.97 (0.20)	14.46 (0.27)	14.79 (0.27)	14.81 (0.28)	14.73 (0.24)
표본수	510	859	502	460	457	582

3개 이상의 그룹에서 평균의 비교

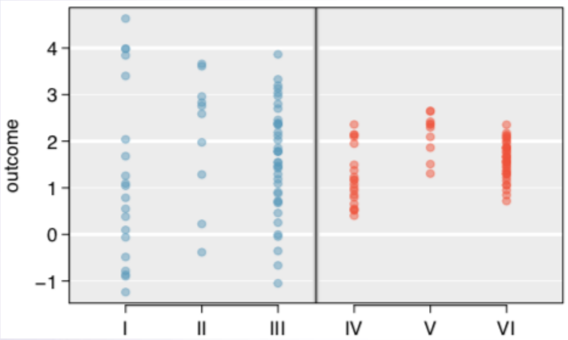


[3개의 그룹에서 그룹간의 평균을 비교 시 그룹내 변동이 미치는 영향]
(OpenIntro Statistics, p286)

- 만약 각 연령대별 표준오차(또는 표준편차)가 굉장히 큰 경우를 가정해보자. 이 경우 각 연령대별 회복 기간 평균의 차이가 어느정도 있더라도 같은 연령대 안에서 표준편차가 크다면 그 차이는 커 보이지 않을 수 있다.

기타 출처 #2

3개 이상의 그룹에서 평균의 비교



[3개의 그룹에서 그룹간의 평균을 비교 시 그룹내 변동이 미치는 영향]
(OpenIntro Statistics, p286)

- 결론적으로 우리는 연령대별 평균 간의 변동(그룹간 변동)과 연령대 안의 변동(그룹안의 변동)을 비교해서 통계적으로 그룹간 평균들이 같은지 여부를 확인하게 된다.

- 물론 연령대 안의 변동들이 모두 다르다고 한다면 위의 방법은 유효하지 않기 때문에 그룹 안의 변동은 거의 동일하다고 가정한다.

기타 출처 #2

분산분석의 절차

- 이렇게 3개 이상의 그룹의 평균을 비교하는 분석방법을 분산분석(ANalysis Of VAriance), 줄여서 ANOVA라고 표현한다.
- 분산분석에서 귀무가설과 대립가설은 다음과 같다.
  - 귀무가설: 각 그룹의 평균은 동일하다.(연령대별로 코로나 (평균)회복 기간은 같다.)
  - 대립가설: 각 그룹별 간 평균은 같지 않다.(연령대별로 코로나 (평균) 회복 기간은 차이가 있다.)
- 검정통계량은 그룹 간의 평균의 변동(연령대별 평균회복 기간의 변동)과 그룹 내의 변동(같은 연령대에서 회복 기간의 변동)의 비교로 표현할 수 있다.

데이터의 구조

- 다음과 같은 관측치를 가지고 있다고 가정하자. 예를 들면 앞의 예제에서 $\mathrm{K} = 6,\ \mathrm{n}_1 = 510$ 이다.

Group	관측치	평균	분산
Group 1	$x_{11}, \dots, x_{1n_1}$	$\bar{x}_1 = \sum_{i=1}^{n_1} x_{1i}/n_1$	$s_1^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2/(n_1 - 1)$
Group K	$x_{K1}, \dots, x_{Kn_1}$	$\bar{x}_K = \sum_{i=1}^{n_K} x_{1i}/n_K$	$s_1^K = \sum_{i=1}^{n_K} (x_{1i} - \bar{x}_1)^2/(n_K - 1)$

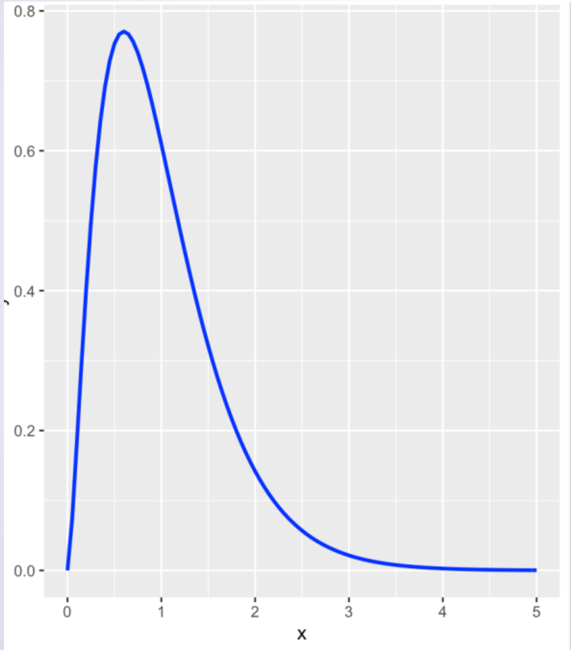
그룹 간 평균의 변동과 그룹 내의 변동

- 먼저 그룹 내의 변동은$MSE = \sum_{i=1}^{K}\sum_{j=1}^{n_k}(x_{ij} - \overline{x}_i)^2/(n-k)$ 로 표시할 수 있다. 여기서 $n = \sum_{i=1}^{K} n_k$로 전체 데이터의 숫자를 의미한다.

- 그룹 간 평균의 변동은 $\sum_{i=1}^{K}(\overline{x}_i - \overline{x})^2/(k-1)$으로 생각할 수 있다. 여기서 여기서 $\overline{x} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} x_{ij}/n$ 즉 전체 평균이다. 하지만 위의 공식은 그룹별 자료의 개수가 차이가 많이 나더라도 각 그룹 평균과 전체 평균과의 차이를 똑같이 반영한다는 단점이 있다. 그래서 그룹 간 변동은 다음과 같이 표현한다.

$$MSG = \sum_{i=1}^{K} n_i(\overline{x}_i - \overline{x})^2/(K-1)$$

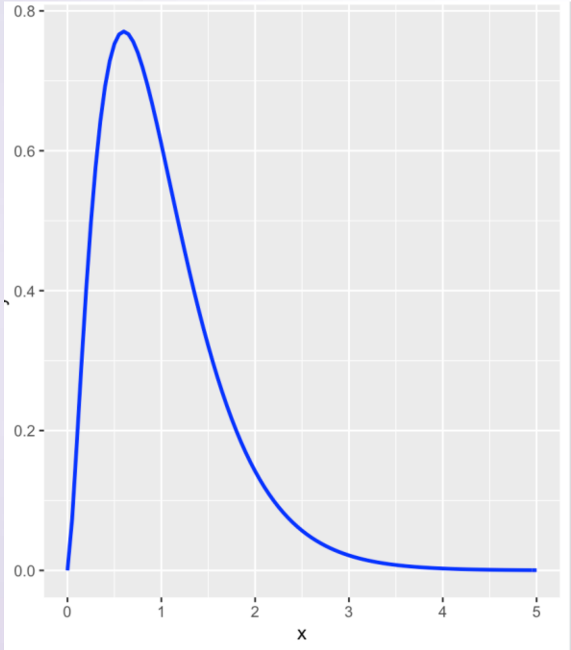
검정통계량과 F-분포



[F(5,3364)의 분포]

- 검정통계량은 따라서 F= (그룹 간의 평균의 변동)/(그룹내의 변동) 으로 정의된다.
- 이제 검정통계량의 표본분포를 알아야 관측된 검정통계량의 값이 얼마나 큰지 결정할 수 있다. 다행히 귀무가설하에서 검정통계량은 F분포를 따른다는 것이 알려져 있다.

검정통계량과 F-분포



[F(5,3364)의 분포]

- F분포는 두 개의 모수가 있는데 분모의 자유도 (=n-K)와 분자의 자유도(=K-1)가 그 모수에 해당한다. 코로나 예제의 경우 귀무가설하에서 검정통계량의 분포는 F(5,3364)를 따른다.

코로나19에서 회복하는 데 걸리는 시간은 연령별로 차이가 있는가?

- 코로나19 회복 기간의 예제에서 검정통계량 F =134.85/34.41 =3.92이다.
- P값은 $Pr(F_{5,3364} > 3.92) = 0.0015$ 이다.
- 즉 연령대별로 평균 회복 기간의 차이는 있다고 할 수 있다.
- 분산분석의 가정
  - 그룹간 등분산성
  - 자료는 정규분포를 따른다.(아주 극심하게 이상해 보이지 않으면 괜찮다.)

오늘의 강의 요점

- 3개 이상의 그룹의 평균의 비교
- 분산분석의 주요 가정: 그룹별 변동은 (거의)같다.

출처

#1 Voinsky I, Baristaite G, Gurwitz D. Effects of age and sex on recovery from COVID-19: Analysis of 5769 Israeli patients. J Infect. 2020;81(2):e102-e103. doi:10.1016/j.jinf.2020.05.026

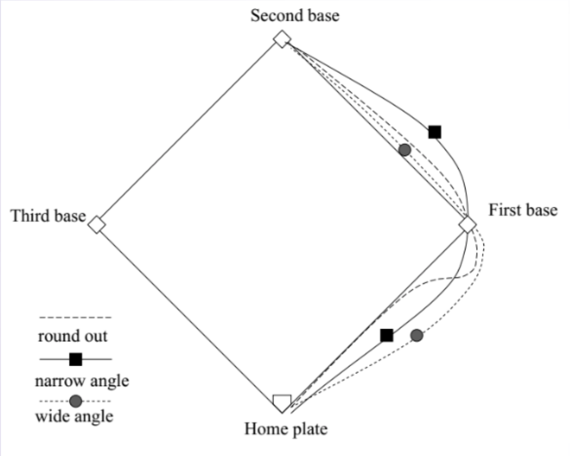
https://www.sciencedirect.com/science/article/abs/pii/S0163445320303030

#2 Diez, D.M., Barr, C.D. and çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

알고 보면 쉬운 두 집단의 비교

Lab 12. 사례연구

쌍체비교



[1루를 좁은 각도와 넓은 각도로 도는 경우]
(Design and Analysis of Experiments with R, p.138)

- 1루를 크게 도는 것과 작게 도는 것 중 어느 방법이 2루에 빨리 도달하는가?
- Applet을 이용해보자
- http://www.rossmanchance.com/ISIapplets.html

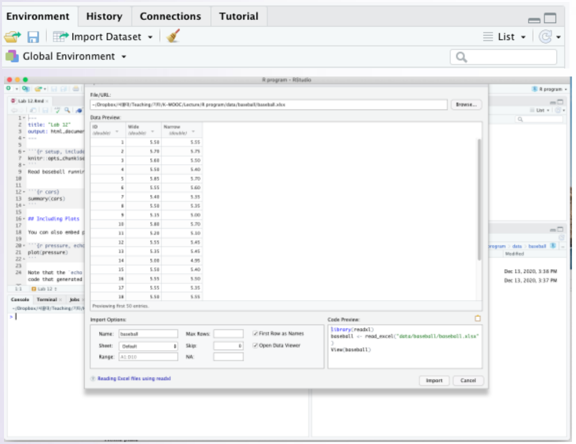
쌍체비교

ID	Wide	Narrow
1	5.5	5.55
2	5.7	5.75
3	5.6	5.5
4	5.5	5.4
5	5.85	5.7
6	5.55	5.6
7	5.4	5.35
8	5.5	5.35
9	5.15	5
10	5.8	5.7
11	5.2	5.1
12	5.55	5.45
13	5.35	5.45
14	5	4.95
15	5.5	5.4
16	5.55	5.5
17	5.55	5.35
18	5.5	5.55
19	5.45	5.25
20	5.6	5.4
21	5.65	5.55
22	6.3	6.25

- 이제 R을 이용하여 직접 이 자료를 분석해보자
- 먼저 엑셀을 이용하여 왼쪽 그림과 같이 자료를 입력하고 본인의 working directory 아래에 data라는 subdirectory를 만들어서 baseball.xlsx라는 이름으로 저장하자.
- 아래 명령어를 직접 입력하여 자료를 읽는다

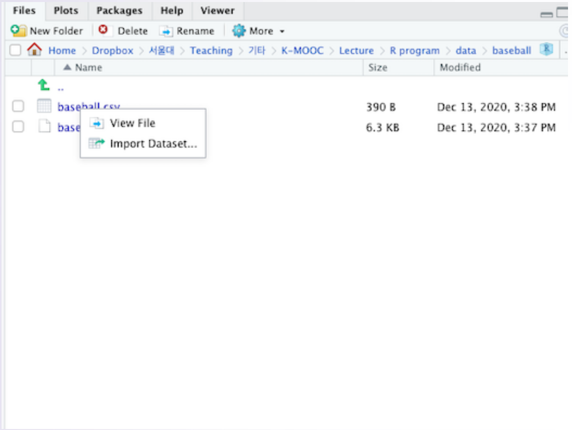
```r
library(readxl)
baseball <- read_excel("data/baseball/baseball.xlsx")
View(baseball)
```

쌍체비교



- 또 다른 방법은 오른쪽 상단에 있는 메뉴를 이용할 수 있다.
- Import Dataset을 클릭하면 아래 화면이 등장하고 하단의 Import 버튼을 누른 후 엑셀 파일 읽기를 선택하면 데이터를 읽을 수 있다.

쌍체비교



- 또 다른 방법은 오른쪽 하단에서 현재 directory를 subdir-ectory인 data로 바꾼 후 입력하고자 하는 파일을 클릭하면 왼쪽 그림과 같이 두가지 메뉴가 보인다.
- 이 메뉴에서 Import Data-set을 선택하면 역시 자료를 읽을 수 있다. 이 경우에는 .csv파일도 읽을 수 있다.

쌍체비교

Read data from an excel file.

```
library(readxl)
baseball <- read_excel("data/baseball/baseball.xlsx")
attach(baseball)
```

Run paired t-test

```
t.test(Wide, Narrow, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  Wide and Narrow
## t = 3.9837, df = 21, p-value = 0.0006754
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03584814 0.11415186
## sample estimates:
## mean of the differences
##                   0.075
```

- Paired-t 검정을 사용하기 위해서는 자료가 정규분포를 따르는지 여부를 확인해야 한다.
- 정규성 가정이 의심스러울 경우 순열검정을 사용하면 된다.

비교 평균의 집단 2

- IMDb 데이터베이스에서 로맨스와 액션 장르 평균 평점비교를 비교하기 위해 액션 영화 32편과 로맨스 영화 36편을 임의로 뽑았다.

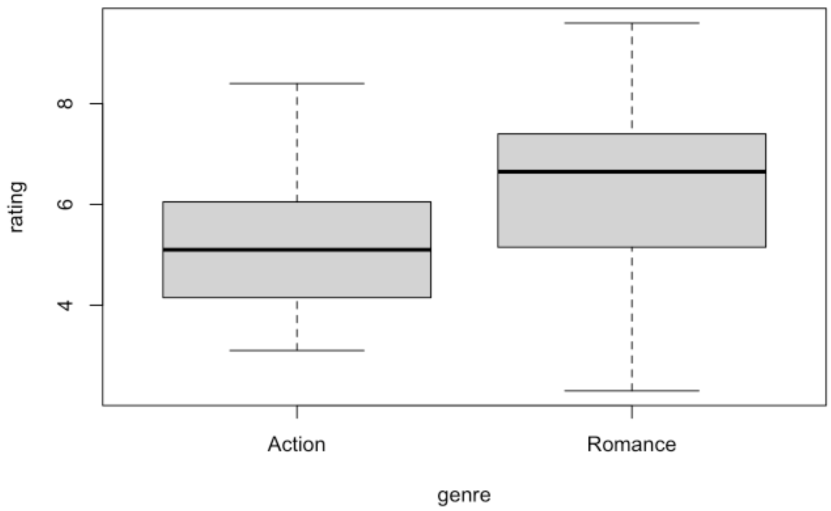
```r
library(moderndive)
movies_sample
```

```
## # A tibble: 68 x 4
##    title                      year rating genre
##    <chr>                     <int>  <dbl> <chr>
##  1 Underworld                 1985    3.1 Action
##  2 Love Affair                1932    6.3 Romance
##  3 Junglee                    1961    6.8 Romance
##  4 Eversmile, New Jersey      1989    5   Romance
##  5 Search and Destroy         1979    4   Action
##  6 Secreto de Romelia, El     1988    4.9 Romance
##  7 Amants du Pont-Neuf, Les   1991    7.4 Romance
##  8 Illicit Dreams             1995    3.5 Action
##  9 Kabhi Kabhie               1976    7.7 Romance
## 10 Electric Horseman, The     1979    5.8 Romance
## # … with 58 more rows
```

```r
attach(movies_sample)
```

이표본 t-검정

```
boxplot(rating~genre)
```



이표본 t-검정

```
t.test(rating~genre)
```

```
##
##  Welch Two Sample t-test
##
## data:  rating by genre
## t = -2.9059, df = 65.85, p-value = 0.004983
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.766773 -0.327671
## sample estimates:
##  mean in group Action mean in group Romance
##              5.275000              6.322222
```

```
t.test(rating~genre, var.equal=TRUE)
```

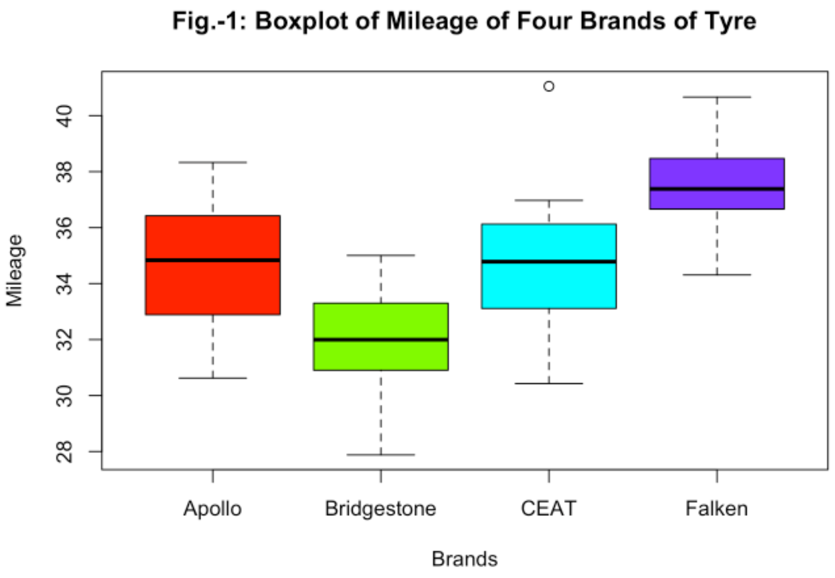
```
##
##  Two Sample t-test
##
## data:  rating by genre
## t = -2.8772, df = 66, p-value = 0.005399
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7739131 -0.3205314
## sample estimates:
##  mean in group Action mean in group Romance
##              5.275000              6.322222
```

3집단 이상 집단에서 평균의 비교

- 다음 예제는 r-blogger.com 에서 인용하였다 (https://www.r-bloggers.com/2017/08/one-way-anova-in-r/)
- 4개 브랜드의 자동차 타이어의 수명을 비교하고자 한다. 자료는 tyre.csv파일로 제공되며 브랜드별로 15개 타이어의 수명이 사용 마일리지로 기록되어 있다.
- 우리는 브랜드별로 수명의 차이가 있는지 궁금하다.

브랜드 별 타이어 수명

```
boxplot(Mileage~Brands, main="Fig.-1: Boxplot of Mileage of Four Brands of Tyre", col= rainbow(4))
```



브랜드 별 타이어 수명

- Boxplot으로 살펴본 결과 등분산성 가정은 크게 문제가 없어 보인다.
- ANOVA에서 F-검정결과는 귀무가설을 기각한다. 즉 브랜드 별로 수명의 차이가 있다.

```
model1<- aov(Mileage~Brands)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Brands       3  256.3   85.43   17.94 2.78e-08 ***
## Residuals   56  266.6    4.76                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 가정에 관한 점검은 회귀분석과 마찬가지로 잔차분석을 통해서 확인할 수 있다.

오늘의 강의 요점

- 쌍체비교에서의 평균이 0인지 여부에 관한 검정: Paired t-검정
  - 2표본에서 평균의 비교: 2표본 t-검정
- 3개이상의 표본에서 평균의 비교: 분산분석
- 가정에 대한 체크는 반드시 해야 한다(정규성, 등분산성)

출처

#1 J. Lawson, (2014), Design and Analysis of Experiments with R, Champman and Hall