5.2 분산분석

- 분산분석의 이론과 파이썬을 이용한 구현 방법
- 일원배치 분산분석, 정규선형모델에서 분산분석의 위상

5.2.1 이 절의 예제

종속변수로는 매상, 독립변수로는 날씨만을 사용합니다.

- 날씨 - 흐림, 비, 맑음 (3가지 수준)

이는 '일원배치 분산분석'이라고 불리는 방법

날씨에 따라 매상이 변화한다고 할 수 있는지 지금부터 검정으로 조사해서 알아보자

5.2.2 분산분석이 필요한 시기

분산분석은 평균값의 차이를 검정하는 방법

이는 t검정을 이용하는 것이 간단하지만 사용할 수 없을 때가 있다.

- 분산분석을 사용해야 할 때는 3개 이상의 수준 간의 평균 값에 차이가 있는지 검정할 때 이다.
    - ex) 날씨의 3가지 경우에 맥주 매상이 유의미하게 차이가 나는지
    - 이처럼 3개 이상의 수준을 가진 테이더가 대상이다. 2개 수준은 t검정
- 분산분석은 모집단이 **정규분포**를 따르는 데이터에 대해서만 적용 가능하다.
- 수준 사이의 분산값이 다르지 않다는 조건도 충족해야한다.

5.2.3 검정의 다중성

검정의 다중성 - 검정을 반복함으로써 유의미한 결과를 얻기 쉬워지는 문제

<유의수준을 0.05라고 정하고 검정 진행>

제 1종 오류를 저지를 확률 : 5% / 검정을 연속 2회 진행

- 이 때 어느 한쪽의 검정에 대해서라도 귀무가설을 기각할 수 있다면 대립가설을 채택한다는 규칙으로 검정을 시행

제 1종 오류를 저지를 확률 : 1 - (0.95*0.95) = 0.0975, **거의 10%**

- 검정을 반복하면 귀무가설이 기각되기 쉬워지고 제 1종 오류를 저지를 확률이 높아진다.

예를 들어 맑음, 비, 흐림의 세 가지 수준으로 매상이 달라지는지 검정할 때, t검정을 실시하면 **다중성 문제** 발생

분산분석은 개별 카테고리가 아니라 날씨에 따른 분석으로 한 번에 검정할 수 있음

5.2.4 분산분석의 직감적 사고방식 : F비

- 귀무가설 : 수준 간의 평균값에 차이가 없다.
- 대립가설 : 수준 간의 평균값에 차이가 있다.

(수준 : 날씨, 물고기 종류 등 카테고리형 변수)

분산분석에서는 데이터의 변동을 **오차**와 **효과**로 분리한다.

그리고 ***F*비**라 부르는 통계량을 계산

$$F\text{비} = \frac{\text{효과의 분산 크기}}{\text{오차의 분산 크기}}$$

- 효과 : 날씨에 따른 매상의 변동
- 오차 : 날씨라는 변수를 이용해서 설명할 수 없는 맥주 매상의 변동

영향의 크기는 분산을 이용해서 정량화 한다. 오차 영향의 크기도 잔차의 분산을 계산함으로써 구한다.

분산의 비율을 취한 통계량으로 검정을 시행하기 때문에 **분산분석**, **ANOVA**라고 불린다.

F비의에 대해 알고 있으면 분산분석을 이해하는데 도움이 된다.

5.2.5 유의미한 차이가 있을 때와 없을 때의 바이올린플롯

분산분석 검정 방법을 파악하기 위해

의미한 차이가 있을 때/아닐 때의 데이터 특징을 확인

그림 5-6 유의미한 차이가 있을 것 같은 바이올린플롯의 예

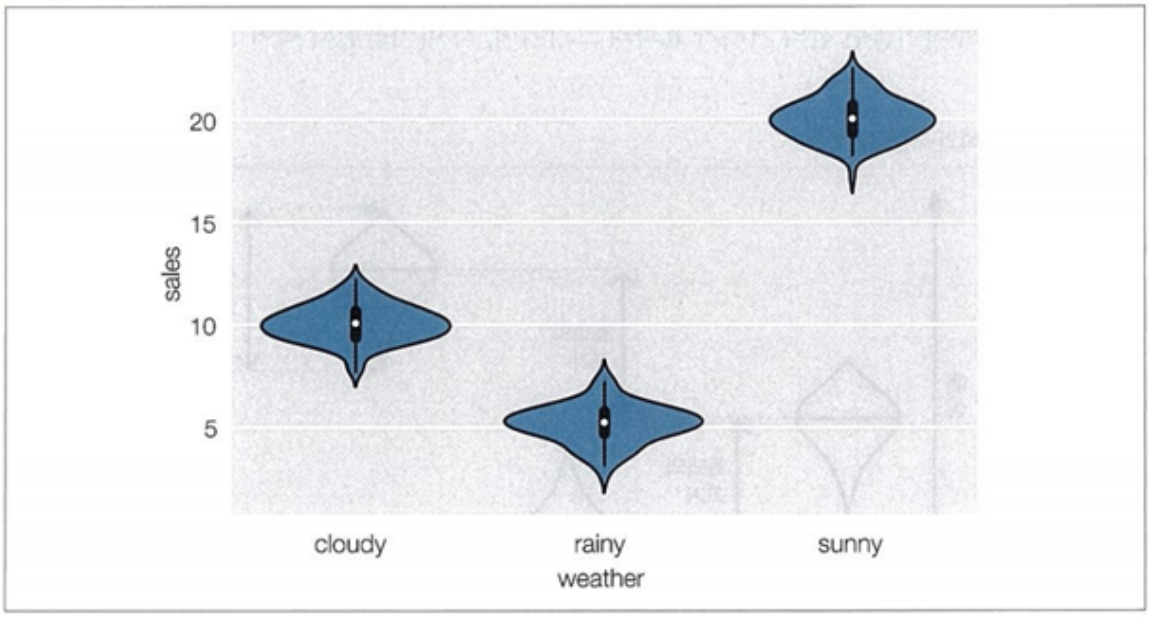
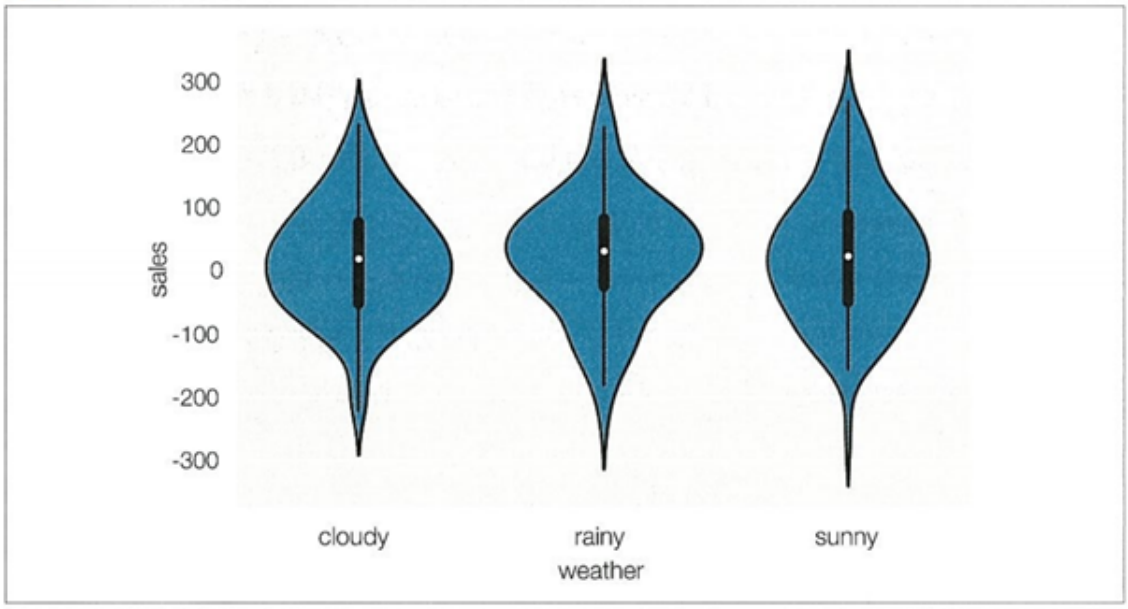


그림 5-7 유의미한 차이가 없을 것 같은 바이올린플롯의 예

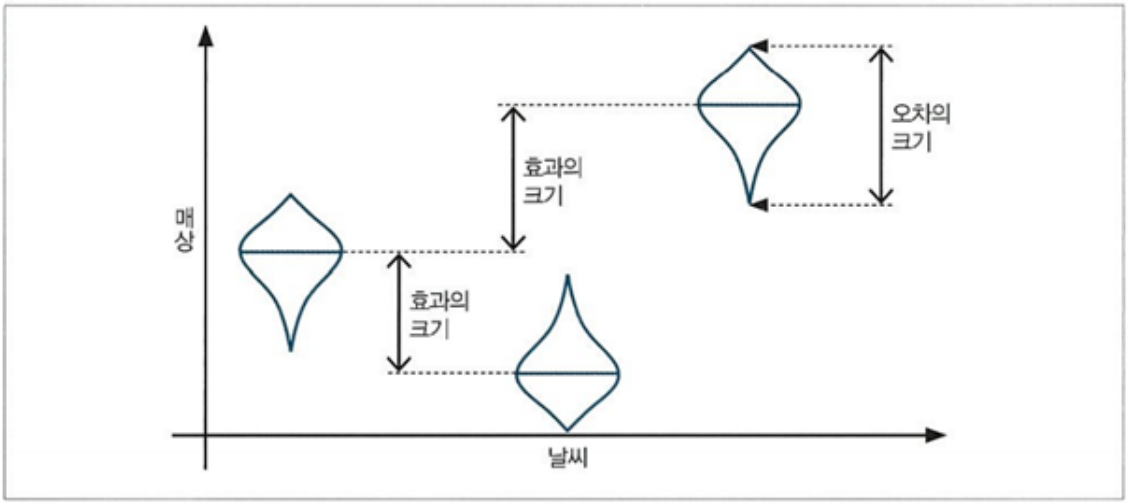


5.2.6 분산분석의 직감적 사고방식 : 오차 및 효과의 분리

오차의 크기와 효과의 크기에 대해 그림을 이용한 해석 시도

- 바이올린 간의 거리 : 효과의 크기
- 바이올린 폭 : 오차의 크기

그림 5-8 분산분석의 직감적인 해석



- 바이올린이 서로 떨어져 있다는 것 : 날시에 따라 매상이 크게 변한다는 것
- 바이올린 간의 거리 = 날씨 효과의 크기
- 같은 날씨라고 해도 매상은 일정하지 않으므로 날씨로 설명할 수 없는 차이를 크기 - 오차의 크기로 표현
-

5.2.7 군간변동과 군내변동

군간변동 - 바이올린 간의 거리, 즉 효과의 크기

군내변동 - 바이올린의 폭, 즉 오차의 크기

분산분석에서는

데이터의 분산을 이 2개로 나눈 뒤, 그 비율을 취한 것을 통계량으로 사용하여 검정을 시행

5.2.8 분석 준비

```
# 수치 계산에 사용하는 라이브러리
import numpy as np
import pandas as pd
import scipy as sp
from scipy import stats

# 그래프를 그리기 위한 라이브러리
from matplotlib import pyplot as plt
import seaborn as sns
sns.set()

# 통계모델 추정에 사용하는 라이브러리
import statsmodels. formula.api as smf
import statsmodels.api as sm
# 표시 자릿수 지정
%precision 3

#그래프를 주피터 노트북에 그리기 위한 설정
%matplotlib inline
```

5.2.9 데이터 작성과 표시

```
# 샘플 데이터
weather = [
    "cloudy", "cloudy",
    "rainy", "rainy",
    "sunny","sunny"
]
beer = [6,8,2,4,10,12]

# 데이터프레임으로 모으기
weather_beer = pd.DataFrame({"beer" : beer, "weather" : weather})
print(weather_beer)
```

5.2.10 분산분석 (1) : 군간 제곱과 군내 제곱 계산

계산 결과를 보기 쉽게 하기 위해 작은 데이터를 준비해서 대상으로 지정

```
# 날씨에 의한 영향
effect = [7,7,3,3,11,11]

# 군간 제곱합
mu_effect = sp.mean(effect)
squares_model = sp.sum((effect - mu_effect)**2)
squares_model
```

샘플사이즈가 작기 땜누에 바이올린플롯이 아니라 상자그림으로 제작

```
sns.boxplot(x = "weather",y = "beer", data = weather_beer,color= 'gray')
```

그래프

날씨별 매상의 평균치 계산

비오는 날 - 매상이 적음, 맑은 날 - 많음. 흐린 날 - 그 중간

```
print(weather_beer.groupby("weather").mean())
```

5.2.10 분산분석(1) : 군간 제곱과 군내 제곱 계산

일원배치 분산분석을 구현해보자

먼저 **효과의 크기, 즉 군간변동 계산**

날씨마다 매상의 평균값은 계산되어 있음.

EX) 흐린 날의 매상 평균값은 7. 여기서 날씨가 흐리게 되면 매상은 7만원이 될 것으로 기대할 수 있다.

각각의 날이 2일씩 있으므로 날씨의 영향만을 생각했을 때, 매상은

```
effect = [7,7,3,3,11,11]
```

이것의 흩어진 정도를 구함으로써 군간변동을 구할 수 있다.

군간변동의 분자에 해당하는 군간 편차제곱합을 계산

```
mu_effect = sp.mean(effect)
squares_model = sp.sum((effect - mu_effect)**2)
squares_model
```

오차는 원래 데이터에서 효과를 빼는 것으로 계산

```
resid = weather_beer.beer - effect
resid
```

군내 편차제곱합

*오차의 평균값은 0*이라는 점에 주의

```
squares_resid = sp.sum(resid**2)
squares_resid
```

5.2.11 분산분석(2) : 군간 분산과 군내 분산 계산

표본분산을 구할 때와 다르게, 불편분산을 계산하기 위해서는 샘플사이즈에서 1을 빼서 나눴다.

이와 마찬가지로 분산분석에서도 군간, 군내 분산을 계산할 때, **샘플사이즈를 자유도로 나누어야 한다.**

- 군간 변동의 자유도 (df_model) : 수준 -1
- 군내 변동의 자유도 (de_resid): 샘플사이즈 - 수준

In
```
# 군간 평균제곱(분산)
variance_model = squares_model / df_model
variance_model
```

Out
```
32.0
```

In
```
# 군내 평균제곱(분산)
variance_resid = squares_resid / df_resid
variance_resid
```

Out
```
2.0
```

5.2.12 분산분석(3) : p값 계산

F비와 p값 계산

F비는 군간 분산과 군내 분산의 비로 계산할 수 있다.

```
In  f_ratio = variance_model / variance_resid
    f_ratio
```

```
Out 16.0
```

p값은 F분포의 누적분포함수에서 계산할 수 있다.

파라미터로는 F비와 2개의 자유도를 넘긴다.

```
In  1 - sp.stats.f.cdf(x = f_ratio, dfn = df_model, dfd = df_resid)
```

```
Out 0.02509457330439091
```

p값이 0.05 이하이므로 날씨에 의해 매상이 유의미하게 변화한다고 판단 가능

일원배치 분산분석의 계산

- 데이터를 효과의 크기, 오차의 크기로 분리
- 각각의 크기를 분산으로 정량화
- 효과의 크기 : 군간변동, 오차의 크기 : 군내변동
- 군간 분산과 군내 분산의 비율, 즉 F비를 통계량으로 사용
- 모집단이 등분산 정규분포를 따를 때, F비는 F분포를 따른다는 것이 밝혀져 있으므로 F분포의 누적분포함수에서 p값을 계산하고 0.05이하인지 판정

5.2.13 독립변수가 카테고리형인 일반선형모델

날씨에서 매상을 예측하는 일반선형모델은 다음과 같다.

$$\text{맥주 매상} \sim \mathcal{N}(\beta_0 + \beta_1 \times \text{비} + \beta_2 \times \text{맑음}, \sigma^2)$$

- 비 : 비가 오면 1, 아니면 0
- 맑음도 마찬가지
- 각각의 영향을 나타내는 파라미더와 곱해져 있음
- 흐림의 경우에는 어떻게 되게요?!?

5.2.14 더미변수

카테고리형 변수를 모델에 넣을 때 사용하는 것

- 비일 때1, 그 외에는 0 - **날씨라는 카테고리형 변수를 그대로 모델에 넣는 것이 어렵기 때문에 더미변수 사용**

다만 statsmodels를 사용해서 모델링을 하는 경우네느 더미변수의 존재를 의할 일이 그다지 없을 것임

5.2.15 를 이용한 분산분석

방금 일원배치 분산분석을 실시한 데이터를 일반선형 모델의 구조로 모델링해보자

```
anova_model = smf.ols("beer ~ weather", data = weather_beer).fit()
```

- 한 번 모델링 해두면 간단히 분산분석 실행 가능
- sm.stats.anova_lm 함수를 사용
- 파라미터로 typ = 2를 넘김

```
print(sm.stats.anova_lm(anova_model, typ = 2))
```

5.2.16 분산분석표

sm.stats.anova_lm 함수의 결과로 출력된 표의 형식

군간과 군내의 편차 제곱합 - sum_sq, 자유도 df, F비, p값이 정리되어 있다.

표를 통해 샘플사이즈나 수준의 개수를 알 수 있으니까 보는 법을 알아두면 좋다.

5.2.17 모델의 계수 해석

```
anova_model.params
```

모델의 식과 어떻게 대응되는지 확인

$$\text{맥주 매상} \sim \mathcal{N}(\beta_0 + \beta_1 \times \text{비} + \beta_2 \times \text{맑음}, \sigma^2)$$

5.2.18 모델을 사용해서 오차와 효과 분리하기

추정된 모델의 계수를 이용해서 훈련 데이터에 적용한 결과

```
fitted = anova_model.fittedvalues
fitted
```

이 적용 결과는 각 수준의 평균과 일치하다.

- 독립변수를 카테고리형 변수로 한 일반선형모델의 추측지(예측치)는 각 수준의 평균값과 일치한다는 것이다.

잔차 - 적용한 결괏값과 실제 데이터의 차이

```
anova_model.resid
```

5.2.19 회귀모델의 분산분석

5.1의 모형 다시 계산

```
beer = pd.read_csv("5-1-1-beer.csv")
lm_model = smf.ols(formula = "beer ~ temperature", data = beer).fit()
```

독립변수가 카테고리형 변수라고 해도 F비 계산 가능

자유도 정의

독립변수가 연속형인 데이터의 경우 용어가 달라진다.

- 군간변동의 자유도 - **모델의 자유도** : 추정된 파라미터 수 -1
    - 단순회귀모델의 계수는 절편과 기울기 2개이므로 이것은 1이 된다.
- 군내변동의 자유도 - **잔차의 자유도** : 샘플사이즈 - 추정된 파라미터 수

```
df_lm_model = 1 # 모델의 자유도
df_lm_reside = 28 # 잔차의 자유도 (30-2)
```

F비 계산

In

```
# 모델을 적용한 값
lm_effect = lm_model.fittedvalues
# 잔차
lm_resid = lm_model.resid
# 기온의 효과의 크기
mu = sp.mean(lm_effect)
squares_lm_model = sp.sum((lm_effect - mu) ** 2)
variance_lm_model = squares_lm_model / df_lm_model
# 잔차의 크기
squares_lm_resid = sp.sum((lm_resid) ** 2)
variance_lm_resid = squares_lm_resid / df_lm_resid
# F비
f_value_lm = variance_lm_model / variance_lm_resid
f_value_lm
```

Out

```
28.44698368850461
```

분산분석표 출력

```
In    print(sm.stats.anova_lm(lm_model, typ = 2))
```

```
Out
                  sum_sq    df          F    PR(>F)
temperature  1651.532489   1.0  28.446984  0.000011
Residual     1625.582178  28.0        NaN       NaN
```

요약하면

```
In    lm_model.summary()
```

Dep. Variable:	beer	R-squared:	0.504
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	28.45
Date:	Mon, 15 Jul 2019	Prob (F-statistic):	1.11e-05
Time:	01:42:52	Log-Likelihood:	-102.45
No. Observations:	30	AIC:	208.9

- F - statistic : F비
- Prob(F - statistic) : 분산분석의 p값

독립변수가 1개인 경우에는 계수의 t검정 결과와 분산분석의 결과가 일치하게 된다.