3회차 파이통!
기초통계학 스터디 정리자료

카테고리

전체보기 (20)

통계학(statistics)
- 기초통계학
- 실생활 통계
- 파이썬 통계

생명과학(biology)

Python
- 기본개념

영어(english)

끄적끄적

활동정보

블로그 이웃 2 명
글 보내기 0 회
글 스크랩 0 회

사용중인 아이템 보기

DOMINO
(min0893)

통계학, 코딩, 생명과학, 영어
삶의 기록

프로필 쪽지

+이웃추가

RSS 2.0 | RSS 1.0 | ATOM 0.3

파이썬 통계 6개의 글

목록열기

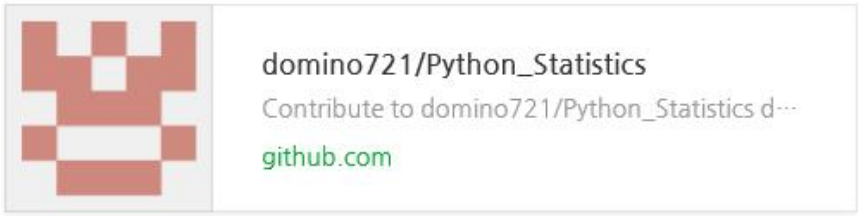
파이썬 통계

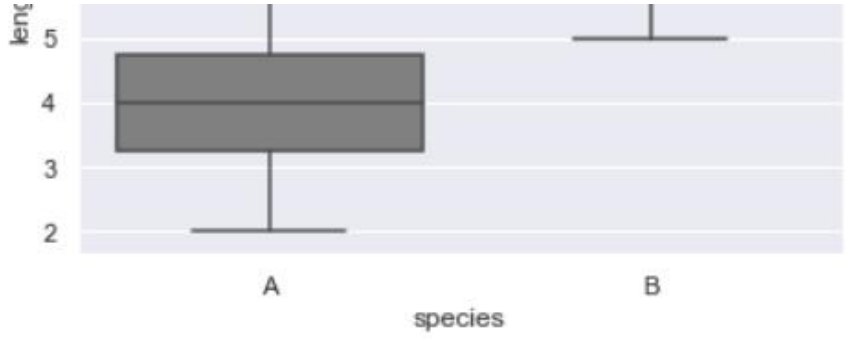
파이썬(Python)을 이용한 데이터 분석 총정리

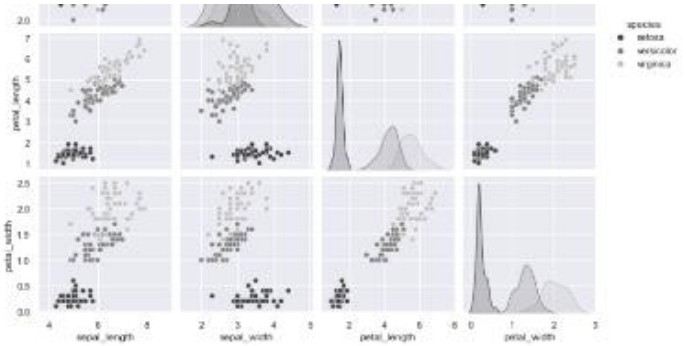
DOMINO · 2021. 4. 4. 20:14

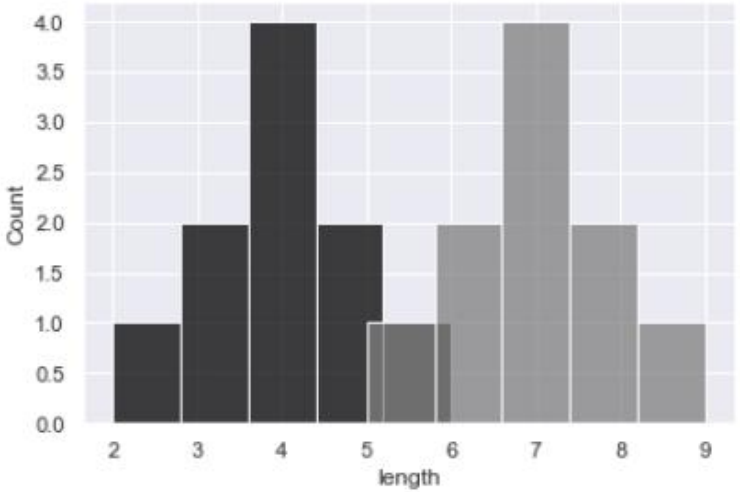
URL 복사 +이웃추가

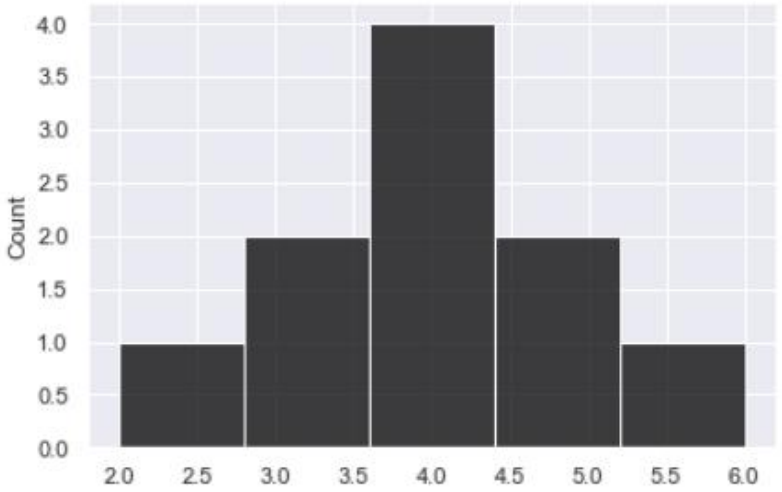
건국대학교 교내 프로그램인 Learning & Sharing을 통해 스터디를 진행하고 있다.
이번주에는 파이썬을 이용해 데이터를 분석하는 방법에 대해서 공부를 해보았다.
gitHub를 통해 정리를 해두었다. 자세한 내용은 GitHub를 통해 참고하면 될 듯하다.
(https://github.com/domino721/Python_Statistics)











맨 위부터 순서대로 1변량 데이터 꺾은선 그래프, 막대그래프, 바이올린플롯, 산포도, 상자수염그림, 페어플롯, 2변량데이터 히스토그램, 1변량 데이터 히스토그램 이다.

https://github.com/domino721/Python_Statistics/blob/main/Ch.3/3-3_matplotlib%EA%B3%BC%20seaborn%EC%9D%84%20%EC%9D%B4%EC%9A%A9%ED%95%9C%20%EB%8D%B0%EC%9D%B4%ED%84%B0%20%EC%8B%9C%EA%B0%81%ED%99%94-Copy1.ipynb

domino721/Python_Statistics

Contribute to domino721/Python_Statistics d···

github.com

기술통계 및 추측통계 카드뉴스

파이썬 이용 데이터 분석

2021. 04. 01

Learning and Sharing

기술통계

1. 데이터 집계

수치계산에 사용하는 라이브러리 import

In [9]:
```
import numpy as np
import pandas as pd
import scipy as sp
from scipy import stats
```

1변량 데이터

1가지 종류의 데이터
ex. 물고기의 몸길이

다변량 데이터

여러 개의 변수를 조합한 데이터
ex. 물고기 종류별 몸길이

기술통계

2. 데이터 통계량 구하기

1변량 데이터

- scipy 함수 이용하여 통계량 구하기
- 데이터 비교를 쉽게 하기 위해 **표준화** 사용하기

다변량 데이터

- 깔끔한 데이터 : 행 하나에 1개의 결과가 나타나도록 정리
- groupby 함수 사용하여 그룹별 통계량 구하기
- 교차분석표
- 공분산은 2개의 데이터의 상관관계를 확인하는 통계량
- 분산-공분산 행렬

$$\mathrm{Cov}(x,y)=\begin{bmatrix}\sigma_x^2 & \mathrm{Cov}(x,y)\\ \mathrm{Cov}(x,y) & \sigma_y^2\end{bmatrix}$$

- 피어슨 상관계수는 공분산의 절댓값 크기를 1로 제한, 표준화하는 것

기술통계

3. 데이터 시각화 (그래프 그리기)

그래프를 그리기 위한 라이브러리 import

In [ ]:
```
from matplotlib import pyplot as plt
import seaborn as sns
sns.set()
```

1변량 데이터

히스토그램 (sns.histplot)

: 측청치의 도수 표현

다변량 데이터

상자수염그림 (sns.boxplot)

: 다양한 통계량을 나타냄

바이올린플롯 (sns.violinplot)

: 커널밀도추정 ->

데이터가 집중된 부분 (도수) 파악

페어플롯 (sns.pairplot)

: 카테고리별 색을 나누어 그래프 표현

추측통계

모집단에서 표본 추출 시뮬레이션

Why ?

"Life is too short, You need Python"

파이썬 시뮬레이션 -> n번 샘플링 반복 -> n개의 실현값

What ?

모집단에서의 표본추출 = 정규분포를 따르는 난수 생성

stats.norm.rvs -> 정규난수 생성 시뮬레이션

How ?

```
In [ ]: # 난수씨드 설정
        np.random.seed(1)
        # (평균4, 표준편차0.8)인 모집단에서 사이즈10인 표본 10000개 추출
        for i in range(0, 10000):
            stats.norm.rvs(loc=4, scale=0.8, size=10)
```

** 난수씨드 지정하면 매번 같은 데이터가 랜덤하게 선택됨.

추측통계

표본의 성질

큰 수의 법칙

표본의 크기 커지면 표본평균이 모평균에 가까워짐

중심극한정리

표본의 개수 충분하면 표본 통계량은 정규분포를 따름

>>> 표본평균의 표준편차 < 모집단의 표준편차
: 극단적 데이터 값들이 배제되기 때문

>>> 표본분산의 평균값과 모분산의 차이 존재
-> 불편분산 사용하여 편향 제거

표본의 분포- t 분포

What ?

모집단분포가 정규분포일 때 t값의 표본분포

t 값 = (표본평균 - 모평균) / 표준오차

** t 분포 형태는 샘플 사이즈에 따른 자유도에 영향 받음.

Why ?

모분산을 모르는 상황에서, 표본평균의 분포에 대해 설명

How ?

In [ ]:

```
# 난수씨드 설정
np.random.seed(1)
# t값을 저장할 변수 설정
t_value_array = np.zeros(10000)
# 정규분포 클래스의 인스턴스
norm_dist = stats.norm(loc=4, scale=0.8)
# 시뮬레이션 실행
for i in range(0,10000):
    sample = norm_dist.rsv(size=10)
    sample_mean = sp.mean(sample)
    sample_std = sp.std(sample, ddof=1)
    sample_se = sample_std/sp.sqrt(len(sample))
    t_value_array[i] = (sample_mean-4)/sample_se
```

정도를 가능하고,
추정결과가 조사에 의미가 있는지 검정

추정

점추정

모수를 어느 1개의 값으로 추정

구간추정

추정값이 '폭' 을 가지는 추정

신뢰도와 이를 만족하는 신뢰구간

검정

가설검정 절차

1 **가설 수립** : 귀무가설 vs 대립가설
2 **유의수준 결정** : 귀무가설 기각 기준
3 **기각역 설정** : 양측검정 / 단측검정
4 **통계량 계산** : 검정통계량
5 **의사결정** : 임계치와 비교

공모전 주제 브레인스토밍

노동자의 안전과 건강.

1. 낮밤 근무에 따른 질병 발병율./
   ↳ ex) 경찰 담직. 간호직, 군인. 소방관.

2. 탄력근무 vs 정시출퇴근. 근무. — ②의 효율성(건강) / ①에 따른 건강.
   산재유형분석.

3. 대/중 소기업간 안전/질병률.
   + 외기업

4. 청년 아르바이트 건강실태 (사고)
   ↳ 배달업종관련.

5. 쿠팡 과로사 → 일용직 노동자. 부상부위/사망원
   작업별 부상부위.

6. 직군별 코로나 발병률.

7. 주재원 관련. (외국/내국) 데이터 수집
   코로나 이후 아시아인 인권

8. 실버 일자리
- 창출
- 문제분석,

9. 대기업 권고사직.
- 정신건강
- 후의 직종선택 ⇒ 심리적 문제.
- 재취의 연령대.

육아출산,

산재_ 사각지대,

우울, 정신건강_

직장내 괴롭힘.

취업시 수도권 집중이유 → 편의시설.
급여,
워라밸_