빅데이터 기반 AI 응용 솔루션 개발자 전문과정

교과목명 : 통계 및 ml 기초

- 평가일 : 21.8.20
- 성명 : 김광훈 ....ㅋㅋㅋ
- 점수 :

In [68]:

```
import statsmodels.formula.api as smf
import statsmodels.api as sm
from scipy import stats
from scipy import stats
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('seaborn-dark')
%config InlineBackend.figure_format = 'retina'
pd.options.display.max_rows = 20
pd.options.display.max_columns = 20

plt.rcParams["figure.figsize"] = (14,4)
plt.rcParams['lines.linewidth'] = 2
plt.rcParams['lines.color'] = 'r'
plt.rcParams['axes.grid'] = True
```

Q1. 다음 괄호안에 들어갈 적합한 단어를 기술하세요.(20점)

확률 변수(random variable)은 확률 실험의 결과에 대한 숫자적 표현입니다. 가령 동전을 던진다고 했을 때 앞면을 1, 뒷면을 0이라고 표현한다고 하면 이는 실험 결과의 숫자적 표현이고 확률 변수입니다. 확률 변수는 그것이 취할 수 있는 값들이 한 개, 두개와 같이 셀 수 있으면 (1) 확률 변수(discrete random variable), 셀 수 없을 경우 (2) 확률 변수(continuous random variable)이라 부릅니다. 동전의 경우 나올 수 있는 경우의 수가 앞 면 또는 뒷 면으로 두 가지니까니 (1) 확률 변수입니다. 그러나 같은 반 아이들의 키 처럼 확률 변수가 가질 수 있는 가지 수가 무한한 경우 (2) 확률 변수에 속하게 됩니다.

확률변수가 취할 수 있는 값들에는 확률이 대응되어 있고, 이를 (3)라고 합니다. 즉 확률 변수들이 어떠한 형태로 놓여있을까, 어떻게 분포해 있을까를 나타내주는 함수입니다. (3) 역시 확률 변수가 이산형 확률 변수이냐, 연속형 확률 변수이냐에 따라서 (4), (5)로 나뉩니다.

(6)란 고정된 지역, 시간 또는 부피 등에서 관심 있는 사건의 관찰 수 또는 발생 횟수 X를 표현하는데 사용되는 분포입니다. 예를 들면 하루 동안 서버에 접속한 사용자 수, 어느 주말 일요일에 발생한 교통사고 사망자 수 등이 있습니다.

평균을 중심으로 좌우가 대칭인 종 모양을 그리는 (7)는 표준 편차가 높을 수록 그래프는 완만한 곡선 형태를 띄게 됩니다.

현실 세계의 많은 데이터들은 정규분포를 따르고 있습니다. 하지만 각 집단의 평균과 표준 편차가 모두 다르기 때문에 데이터들을 서로 비교하기가 어렵습니다. 예를 들어 A반의 수학 점수가 평균은 70점이고 표준 편자는 30점입니다. 반면 B반은 평균 65점에 표준 편차가 10점이라면 두 반 중 어느 반이 더 수학 점수가 높다고 할 수 있을까요? 이러한 비교를 위해서 전체 데이터를 평균으로 빼주고, 표준 편차로 나누어 주는 표준화(standardizing) 작업을 거치게 됩니다. 그 결과 두 집단의 수학 점수는 동일하게 (8)를 따르게 되며 평균이 0이고 분산이 1인 정규 분포가 됩니다. 수식으로는 Z = (X - mu) / sigma로 표현할 수 있습니다. 표준화를 거친 개별 데이터를 우리는 (9)라고 부릅니다.

(10)는 정규 분포인 모집단의 모평균을 표본 평균을 통해서 추측할 때 사용되는 분포입니다. 정규 분포와 유사한 형태를 갖지만 양 끝단에 데이터가 더 많이 분포하는 형태를 띕니다. 모집단에서 표본을 추출할 경우 표준 편차가 더 커질 것이라는 것을 예상할 수 있습니다. 이 때문에 곡선의 모양이 더 완만해 지는 것을 (10)로 설명한다고 이해하면 좋을 것 같습니다.

정규 분포의 경우 그래프의 형태를 표준 편차와 평균이 결정하였습니다. t-분포는 이 둘에 더해 수식 상에서 (n-1)에 해당하는 (11)가 그래프의 형태에 영향을 줍니다. (11)의 정의는 말 그래도 자유스러운 정도로 특정 분포에서 그래프의 모양을 결정하는 모수입니다. 대표적으로 t 분포와 카이제곱 분포가 ( )를 모수로 갖습니다.

(12)는 f 검정이나 분산 분석에서 많이 활용된다고 합니다. (12)는 두 모분산의 비율에 대한 추정을 할 때 사용됩니다.

(13)관찰된 빈도가 기대 빈도와 통계적으로 다른지를 판단하는 검증방법으로 사용한다.

모집단의 표본에서 얻은 통계량을 통해 모집단의 통계적 특성을 추측할 수 있습니다. 이러한 과정을 논리적으로 전개하기 위해서 필요한 것이 가설과 검정입니다. 가설(hypothesis)란 확률 분포에 대한 어떠한 주장이며 이를 증명하는 행위를 검정(testing)이라 합니다. 특히 확률 분포의 모수 값에 대한 가설을 검정하는 것을 (14)이라 부릅니다.

(15)은 처음부터 버릴 것을 예상하는 가설입니다. 기본적으로 참으로 추정되며 이를 거부하기 위해서는 증거가 반드시 필요합니다. 예를들어 형사가 용의자를 잡았을 경우에도 무죄 추정의 원칙에 따라서 '이 용의자는 무죄일 것이다' 라는 가설을 먼저 세우게 됩니다.(15)을 세울 때에는 특별한 증거가 없다면 참으로 여겨지는 가설을 (15)로 세우게 됩니다.

(16)란 귀무 가설이 맞는데도 이를 잘못 기각하여 발생하는 오류입니다. 용의자가 무죄가 맞지만 잘못하여 유죄 판결을 내리는 것과 같습니다. (17)란 대립 가설이 사실임에도 불구하고 귀무가설을 기각하지 못하는 오류를 말합니다. 용의자가 범인이 맞지만 무죄가 아니라는 것을 입증해내지 못하는 것을 말합니다.

(18)란 귀무 가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 관측될 확률입니다.

(19)은 귀무가설을 기각하는 기준이 되는 값이다.

통계모델에 사용하는 파라미터를 계수라고 하는 반면 머신러닝에서는 (20)라고 표현한다.

- 1. 이산 (확률 변수)
- 2. 연속 (확률 변수)
- 3. 확률 함수
- 4. 이산형 확률분포
- 5. 연속형 확률분포
- 6. 포아송 분포
- 7. 정규분포
- 8. 표준정규분포
- 9. z-score
- 10. t분포

- 11. 자유도 df
- 12. f분포
- 13. 카이제곱검정 (교차검정)
- 14. 모수검정
- 15. 귀무가설
- 16. 1종 오류
- 17. 2종 오류
- 18. p_value
- 19. 유의수준
- 20. 가중치

Q2. 다음의 확률밀도함수를 그래프로 표현하고 모집단 분포와 비교한 후 t분포의 특징을 기술하세요.(5)

- 표준정규분포
- 표준정규분포인 모집단에서 샘플 10개를 추출한 표본의 t분포
- 평균 0, 표준편차 2인 정규분포
- 평균 0, 표준편차 2인 정규분포인 모집단에서 샘플 5개를 추출한 표본의 t분포

In [2]:

```
mu = 0
std = 1
norm_ = stats.norm(mu,std)

sample_data = norm_.rvs(10)
smaple_mean = np.mean(sample_data)
sample_data = np.std(sample_data,ddof =1)
```

Q3. t검정과 관련 다음 사항에 대하여 기술하세요.(5)

- t검정을 사용하는 경우
- t검정에 사용하는 t-value 산출 방법
- t-value와 z score 간의 의미상 차이

<답1> 집단간 평균의 차이를 검정하는 방법. <답1> 모집단이 2개 이하일 경우에만 이에 해당된다.

<답2> t_value는 표본평균에서 모평균을 빼고 표준오차로 나눠 준 값. (표본평균 - 모평균) / 표준 오차

<답3> z_value는 모집단의 평균과 모집단을 모두 알지만 t_value값은 모집단의 평균은 알지만 분산을 모른다는 점에서 차이가 있다.

Q4. 남성 키의 평균 170 표준편차는 5인 모집단 생성(size=2000) 후 그 중 10개를 출력하세요.

In [3]:

```
tall = stats.norm(loc=170,scale=5)
data = tall.rvs(size=2000)
data[:10]
```

```
array([173.33847401, 163.09951989, 170.59177417, 167.3802138 ,
       168.0661604 , 176.14886994, 168.27453347, 162.51617548,
       173.51528427, 172.05744693])
```

Q5. Q4에서 생성된 모집단에서 20개의 샘플을 추출한 후 아래 사항을 수행하세요. 단 모집단의 분산은 알 수가 없다는 것을 가정한다.

- 추출한 샘플1, 샘플2로 부터 남성의 평균 키가 170이다라는 가설을 세우고 검정
- 샘플2를 기준으로 표준오차를 구하세요
- 샘플2를 기준으로 95% 신뢰수준으로 신뢰구간을 구하세요

In [9]:

```
#sample1 1 추출
np.random.seed(0)
sample1 = np.random.choice(data,10,replace=False)
sample1 = data[:20]
sample1
```

```
array([173.33847401, 163.09951989, 170.59177417, 167.3802138 ,
       168.0661604 , 176.14886994, 168.27453347, 162.51617548,
       173.51528427, 172.05744693, 161.01112076, 167.2384077 ,
       171.7775144 , 166.84132685, 169.09031015, 177.20111914,
       167.85130921, 169.57290859, 160.69590191, 176.1607646 ])
```

In [10]:

```
# sample 2 추출
np.random.seed(0)
sample2 = np.random.choice(data,20,replace=False)
sample2 = data[:20]
sample2
```

```
array([173.33847401, 163.09951989, 170.59177417, 167.3802138 ,
       168.0661604 , 176.14886994, 168.27453347, 162.51617548,
       173.51528427, 172.05744693, 161.01112076, 167.2384077 ,
       171.7775144 , 166.84132685, 169.09031015, 177.20111914,
       167.85130921, 169.57290859, 160.69590191, 176.1607646 ])
```

In [11]:

```
mu =  np.mean(sample2)
df = len(sample2)-1

sigma = np.std(sample2,ddof= 1)
se = sigma/np.sqrt(len(sample2))
se

t_value = (mu-179)/se
```

In [12]:

```
alpha = stats.t.cdf(t_value,df=df)
(1-alpha)*2

stats.ttest_1samp(sample2,170)

# 유의구간 0.25~ 귀무가설 기각
```

```
Ttest_1sampResult(statistic=-0.8062247994659046, pvalue=0.4300893546500032)
```

In [13]:

```
print(sample1.mean())
print(sample2.mean())
```

```
169.12145678287536
169.12145678287536
```

Q6. 어느 자동차 임대 회사에서 3가지 종류의 휘발유를 비교하는데 관심이 있다. 각 종류의 휘발유에 대하여 4번의 주행을 하여 반응변수로서 각 주행에 대한 리터당 주행거리를 구하였다.(qsample.csv) 다음 사항을 고려하여 휘발유간의 주행거리 차이가 유의미한지를 분산분석을 사용하여 검정하세요.

- 각 휘발유의 4개의 실험에서의 주행거리가 다름 = 실험오차(군내변동)
- 휘발유 종류에 따라 각 평균 주행거리가 서로 다름(군간변동)

In [25]:

```
import pandas as pd
df = pd.read_csv('Data/qsample.csv')
df
```

	product	distance
0	제품1	15.2
1	제품1	16.1
2	제품1	16.8
3	제품1	15.9
4	제품2	18.5
5	제품2	17.5
6	제품2	18.2
7	제품2	17.8
8	제품3	19.6
9	제품3	19.3
10	제품3	18.4
11	제품3	18.7

In [35]:

```
anova_model = smf.ols('distance~product',
                      data=df).fit()
display(sm.stats.anova_lm(anova_model,type=2))

anova_model.params

#따라서 군내변동 군간변동 둘다 의미가 있다.
```

	df	sum_sq	mean_sq	F	PR(>F)
product	2.0	18.666667	9.333333	30.215827	0.000102
Residual	9.0	2.780000	0.308889	NaN	NaN

```
Intercept         16.0
product[T.제품2]     2.0
product[T.제품3]     3.0
dtype: float64
```

Q7. 비만과 당뇨의 상관관계를 분석하기 위하여 조사한 결과로 qsample2 데이터를 작성하였다. 비만과 당뇨가 서로 독립적인지를 판단하기 위한 카이스퀘어 검정을 수행 하세요.

In [36]:

```
data = pd.read_csv("Data/qsample2.csv")
```

In [55]:

```
data
```

	당뇨	정상	index
0	10	10	비만체중
1	15	65	정상체중

In [62]:

```
# 분할표
cross = pd.pivot_table(data=data,values='정상',
                       aggfunc='sum',index='index',columns='click')
cross

stats.chi2_contingency(cross,correction = 100)
```

```
(0.0,
 1.0,
 0,
 array([[65.],
        [10.]]))
```

Q8.'7_1_beer.csv' 데이터 셋에서 온도의 변화에 따른 맥주 매상 예측 모델을 statsmodel 라이브러리를 활용하여 수행 후 summary() 결과를 해석하세요.

In [76]:

```
beer = pd.read_csv('Data/7_1_beer.csv')
beer.head()
```

	beer	temperature
0	45.3	20.5
1	59.3	25.0
2	40.4	10.0
3	38.0	26.9
4	37.0	15.8

In [77]:

```
lm_model = smf.ols(formula = "beer ~ temperature",
                   data=beer).fit()
```

In [78]:

```
lm_model.summary()
```

OLS Regression Results

Dep. Variable:	beer	**R-squared:**	0.504
Model:	OLS	**Adj. R-squared:**	0.486
Method:	Least Squares	**F-statistic:**	28.45
Date:	Fri, 20 Aug 2021	**Prob (F-statistic):**	1.11e-05
Time:	16:32:18	**Log-Likelihood:**	-102.45
No. Observations:	30	**AIC:**	208.9
Df Residuals:	28	**BIC:**	211.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	34.6102	3.235	10.699	0.000	27.984	41.237
temperature	0.7654	0.144	5.334	0.000	0.471	1.059

Omnibus:	0.587	**Durbin-Watson:**	1.960
Prob(Omnibus):	0.746	**Jarque-Bera (JB):**	0.290
Skew:	-0.240	**Prob(JB):**	0.865
Kurtosis:	2.951	**Cond. No.**	52.5

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

결정계수가 0.5, 수정결정계수가 0.4~5 사이로, 설명력은 많이 좋지 않다고 볼 수 있다. 또한

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Q9.iris 데이터셋으로 분류 분석을 수행하고 평가하세요.

In [83]:

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
```

In [84]:

```
iris = load_iris()
iris_label=iris.target
iris_data=iris.data
iris_df = pd.DataFrame(data=iris_data, columns=iris.feature_names)
iris_df['label'] = iris_label
iris_df.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

In [85]:

```
X_train,X_test,y_train,y_test = train_test_split(iris_data,
                                                 iris_label,
                                                 test_size=0.2,
                                                 random_state=10)
```

In [86]:

```
df_clf = DecisionTreeClassifier()
df_clf.fit(X_train,y_train)
```

```
DecisionTreeClassifier()
```

In [87]:

```
prediction = df_clf.predict(X_test)
```

In [90]:

```
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test,prediction))
```

```
0.9666666666666667
```

예측 정확도가 약 97%정도 되었다.

Q10. 'auto-mpg' 데이터셋에서 horsepower에 따른 mpg를 예측하는 선형회귀 모델을 생성한 후 정확도 평가를 수행하세요.(단, 종속변수의 정규성을 개선해야 함)

In [ ]:

```
df = pd.read_excel("./dataset/auto-mpg.xlsx")
ndf=df[['mpg','cylinders','horsepower','weight']]
ndf.head()
```