Chapter 02. 데이터와 표본분포 - 오다원

2.6 정규분포 (Normal distribution)

- 오차 error : 데이터 포인트와 예측값 혹은 평균 사이의 차이
- 표준화(정규화)하다 : 평균을 빼고 표준편차로 나눈다.
- z -score : 개별 데이터 포인트를 정규화한 결과
- 표본정규분포 : 평균 = 0, 표준편차 = 1 인 정규분포
- **Quantile - 분위수**

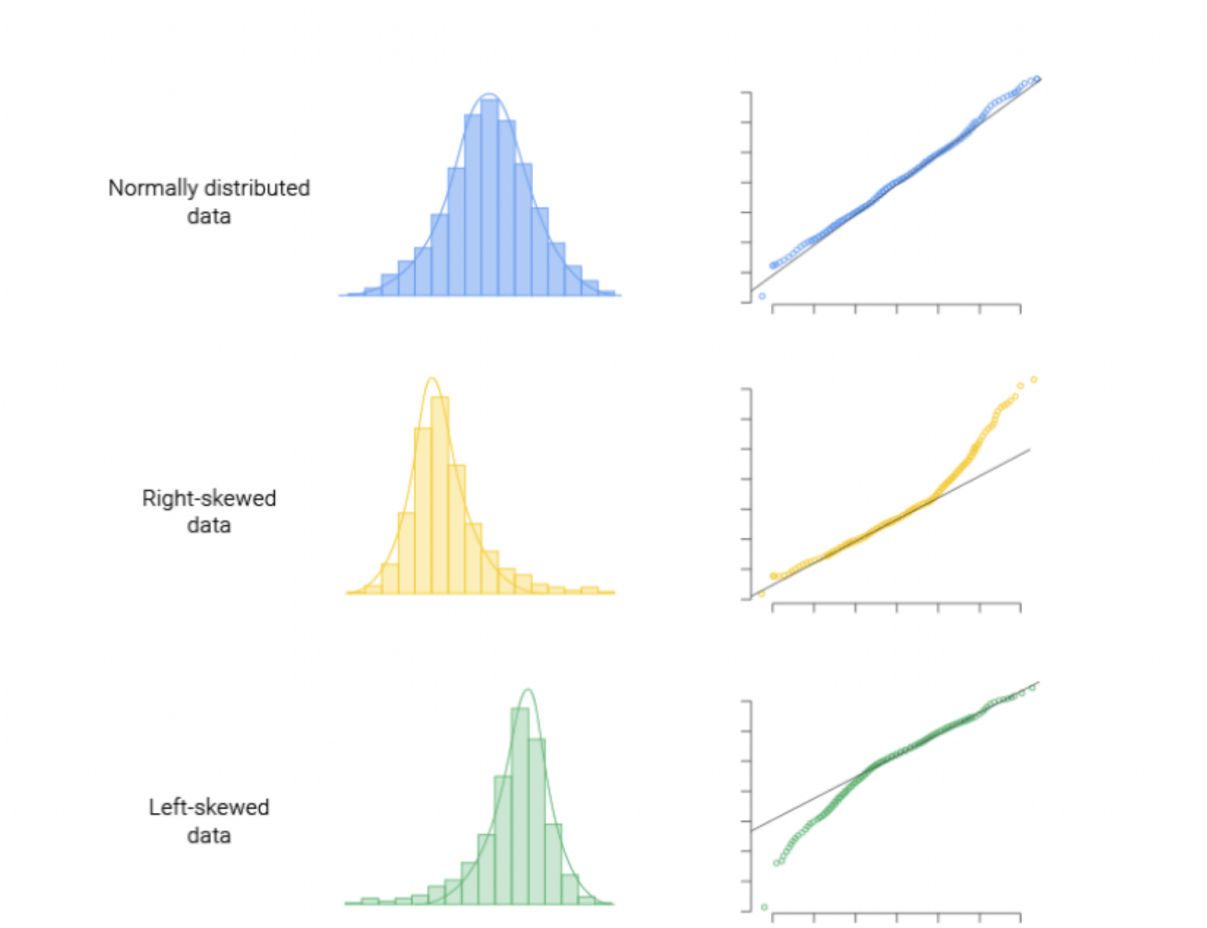
QQ그림(qq-plot) : 표본분포가 특정 분포에 얼마나 가까운지를 보여주는 그림

2.6.1 표준정규분포와 QQ 그림

- 표준정규분포 : x축 단위가 평균의 표준편차로 표현되는 정규분포

- QQ그림(Quantile-Quantile Plot) : **표본이 특정 분포에 얼마나 가까운지**를 시각적으로 판별하는데 사용
  - z점수를 오름차순으로 정렬하고 각 값의 z점수를 y축에 표시
  - x 축은 정규분포의 해당 분위 Quantile
  - y 축은 z score
  - 점들이 대략 대각선 위에 놓이면 표본분포가 정규분포에 가까운 것으로 간주

- 원시 데이터 자제는 대개 정규분포가 아니지만,
- 표본들의 평균과 합계,
  그리고 오차는 많은 경우 정규분포를 따름
- 데이터를 z점수로 변환하려면 데이터의 값에서 평균을 빼고 표준편차로 나눔



2.7 긴 꼬리 분포

- 꼬리 : 적은 수의 극단값이 주로 존재하는, 도수분포의 길고 좁은 부분
- 왜도 : 분포의 한쪽 꼬리가 반대로 다른 꼬리보다 긴 정도 skew

  평균과 중앙값이 같으면 왜도 0
  왜도가 양수일 때, 확률밀도함수의 오른쪽 부분에 긴 꼬리를 가지며 왼쪽에 데이터가 더 많이 분포
- 분포의 꼬리는 양 극한값에 해당
- 흑고니 이론 : 이례적인 사건이 정규분포로 예측되는 것보다 훨씬 더 자주 일어날 수 있다고 예측
  ex. 주식시장의 붕괴

- 주가 수익률 같은 **긴 꼬리 특성**을 가진 데이터는 QQ 그림에서 낮은 값들의 점은 대각선 보다 훨씬 낮고 높은 값은 선보다 훨씬 위에 위치
- 정규분포를 따르지 않으며 훨씬 더 많은 극단값을 관찰할 가능성이 있음을 의미
- 평균에서 표준편차 이내에 있는 데이터의 점들은 선에 더 가까이 위치

- 대부분 데이터는 정규분포를 따르지 않음
- **정규분a포를 따를 것이라는 가정은 자주 일어나지 않는 예외 경우에 관한 과소평가를 가져올 수 있음**

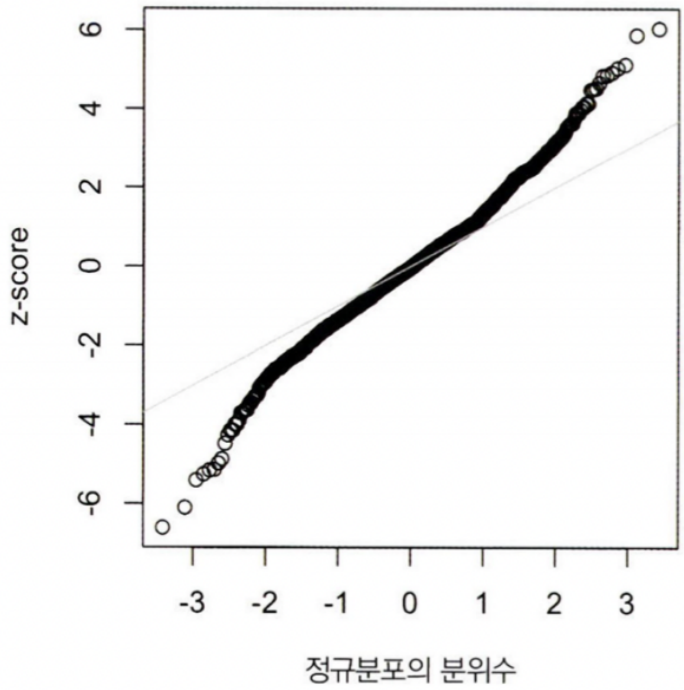
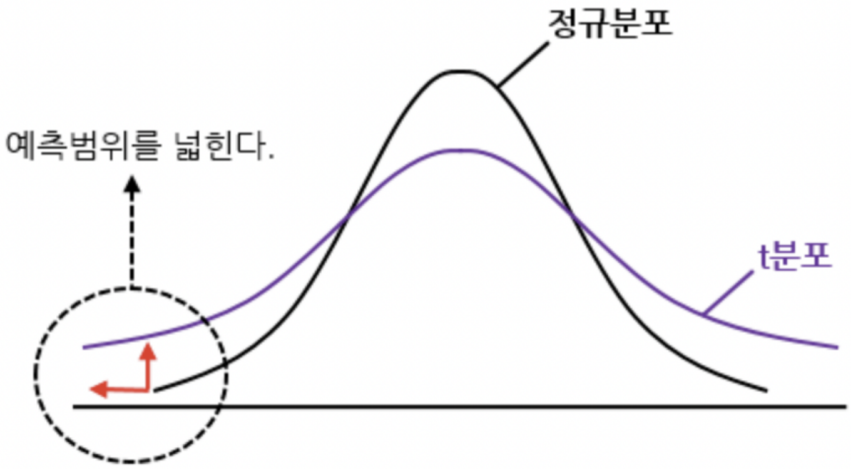


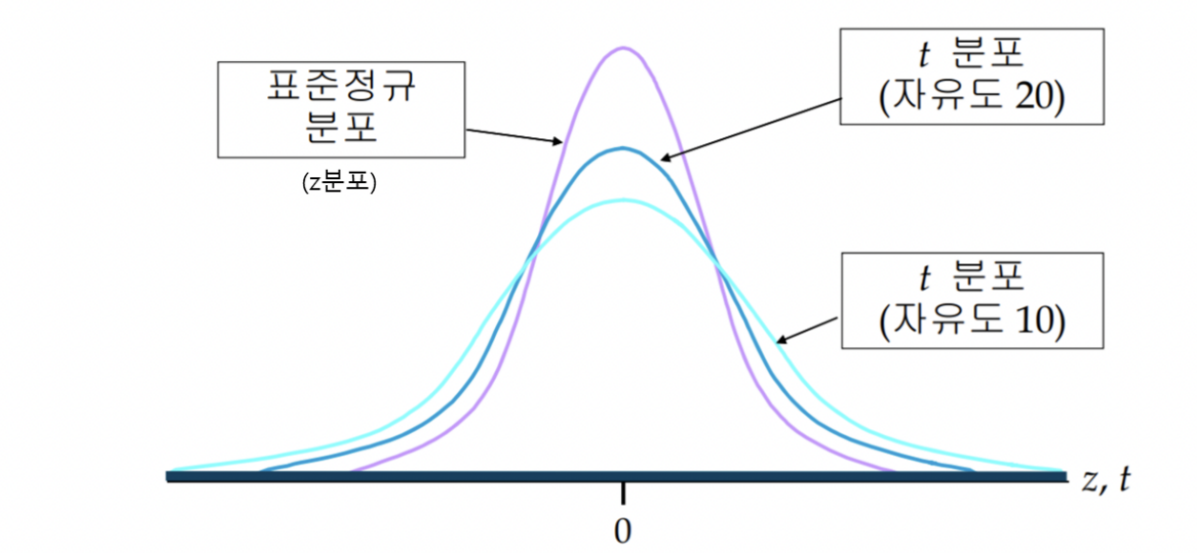
그림 2-12 넷플릭스(NFLX)의 일일 주식 수익률에 대한 QQ 그림

2.8 스튜던트의 t 분포

- n : 표본 크기
- 자유도 : 다른 표본크기, 통계량, 그룹 수에 따라 t분포를 조절하는 변수



- t 분포는 정규분포와 생김새가 비슷하지만 꼬리부분이 더 두껍고 긴 형태
- 표본평균의 분포는 일반적으로 t 분포와 같은 모양
- 표본크기에 따라 다른 계열의 t 분포가 있으며 표본이 클수록 더 정규분포를 닮은 t 분포



- t 분포는 표본평균, 두 표본평균 사이의 차이, 회귀 파라미터 등의 분포를 위한 기준으로 널리 사용됨

정규분포를 따르는 모집단에서 크기 n인 표본을 무작위 추출하였을 때 표본평균을 $\bar{x}$, 표본분산을 s^2(표본표준편차는 s)라고 하면, 통계량 t는 자유도 $n-1$인 t 분포를 따른다.

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=antifatekr&logNo=221061029807

- 표본평균이 $\bar{x}$, 크기 n 의 표본, s 가 표준편차일 때 표본평균 주위 90% 신뢰구간

- $t_n - 1(0.05)$: 자유도 n-1 를 갖는 t 분포의 양쪽 끝에서 5%를 잘라내는 t 통계량

$$\bar{x} \pm t_{n-1}(0.05) \cdot \frac{s}{\sqrt{n}}$$

2.9 이항분포

- 시행 : 독립된 결과를 가져오는 하나의 사건 (ex. 동전 던지기)

- 성공 : 시행에 대한 관심의 결과 (유의어 1, 즉 0에 대한 반대)
- 이항식 : 두 가지 결과를 갖는다 (예/아니오, 0/1, 이진)
- 이항시행 : 두 가지 결과를 가져오는 시행 (베르누이 시행)
- 이항분포 : n 번 시행에서 성공한 횟수에 대한 분포 (베르누이 분포)

- 이항분포 : 각 시행마다 그 성공 확률(p)이 정해져 있을 때, 주어진 시행 횟수(n) 중에서 성공한 횟수(x)의 도수 분포
- 이항분포의 평균 : n*p (성공확률이 p 일 때, n번의 시행에서 예상되는 성공 횟수)
- w분산 : n*p(1-p)

- 이항분포는 이항 확률을 구할 때 많은 계산이 필요하고,
  사실상 정규분포와 구별이 어려워 통계 절차에서는 평균과 분산으로 근사화한 정규분포를 사용
- 이항 결과는 무엇보다 중요한 결정 사항들을 나타내므로 모델을 만드는데 매우 중요 (생존/죽음 등)
- 이항시행은 두 가지 결과, 즉 하나는 확률이 p, 다른 하나는 확률이 1-p 인 실험
- n이 크고 p가 0 또는 1에 너무 가깝지 않은 경우, 이항 분포는 정규분포로 근사할 수 있음

FAQ

1. 이항분포에서 이항확률을 구할때 계산이 많이 필요하다는 의미

$$Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   - 여기에서 각각의 k와 p 등을 구하는데 많은 개념이 필요하고, 이것을 구한뒤 분포에 대한 계산을 하기 위한 다른 계산도 필요하다는 의미

2.10 카이제곱분포

카이제곱 검정/적합도 검정 (Chi-square Test)/(Goodness of Fit Test)

카이 제곱 검정은,

- **관찰된 빈도가 기대되는 빈도와 유의미하게 다른지를 검증하는 통계 검정 방법**이다.
- 주로 **범주형 자료**로 구성된 데이터 분석에 이용된다.
- 핵심은 두 범주형 변수가 서로 상관이 있는 지 혹은 독립 관계인지 이다.
- 참고로 범주형 자료는 categorical data 로, 월 소득 100만원 미만, 이상 등 구간에 대한 자료를 의미한다.

카이제곱 검정의 형태는 다음과 같다.

1. Goodness of fit test : 적합도 검정. (Pearson의 카이제곱 검정)

적합도 검정이란, 어떤 모집단의 표본이 그 모집단을 대표할 수 있는 지 검정하는 방법으로, 관찰 된 비율 값이 기대값과 같은지 여부를 검정하는 방법이다. 변수는 1개 이다.

Test of homogeneity : 동질성 검정.

동질성 검정이란, 두 집단의 분포가 동일한지 검정하는 방법이다.

1. Test for independence : 독립성 검정.

동립성 검정은 두 개 이상의 변수가 독립인지 검정하는 방법이다. 즉, 각 표본들이 관찰 값에 영향을 주는지 여부를 검정하는 방법이다.

적합도 검정

- 카이제곱 검정의 종류중 하나로

- 관측도수와 기대도수의 따른 비율차이가 통계적으로 유의한지 살펴보는 검정
- 관측도수 = 실제 표본조사 빈도
- 기대도수 = 알려진 모집단의 분포에 의해서 예상되는 빈도

그니까 **주어진 데이터의 분포가 예상되는 분포랑 비슷하냐 ?** ? 물어보면 대답할때 쓴다.

라고 하고 해보려고 했는데 ;;

범주형 변수가 진단 하나 뿐

그래서 범주형 변수를 하나 더 만들어 준다. 임의로 (뭔지 모르지만 일단 만들어!!)

In [6]:

```
def new(x):
  if x >20:
    return 1
  else :
    return 0
```

In [7]:

```
df['new'] = df['radius_mean'].apply(new)
```

In [8]:

```
df
```

Out[8]:

속성	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mea
0	842302	@2021년 9월 20일	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	@2021년 9월 20일	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	@2021년 9월 20일	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	@2021년 9월 20일	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	@2021년 9월 20일	20.29	14.34	135.10	1297.0	0.10030	0.13280
...	...		...	...	...	...	...	...
564	926424	@2021년 9월 20일	21.56	22.39	142.00	1479.0	0.11100	0.11590
565	926682	@2021년 9월 20일	20.13	28.25	131.20	1261.0	0.09780	0.10340
566	926954	@2021년 9월 20일	16.60	28.08	108.30	858.1	0.08455	0.10230
567	927241	@2021년 9월 20일	20.60	29.33	140.10	1265.0	0.11780	0.27700
568	92751		7.76	24.54	47.92	181.0	0.05263	0.04362

569 rows × 34 columns

In [9]:

```
# 크로스탭 쿼리를 통해 빈도교차표를 출력한다.  X = pd.crosstab(df.new, df.diagnosis, margins= True )
# Margins = 행과 열의 총합을 표시 할 것인가? X
```

Out[9]:

제목	diagnosis	B	M	All
제목 없음	@2021년 9월 22일	0	0	0
제목 없음	@2021년 9월 13일 오전 12:00	357	167	524
제목 없음	@2022년 1월 1일	0	45	45
제목 없음		357	212	569

범주형 변수가 아니더라도 범주형 변수로 바꿔서 이렇게 빈도를 나타내는 교차표를 만들었다.

ALL 은 총합

- 바꿔말하면 범주형 변수가 아니더라도 가능함
- 실제 표본조사 빈도(관측도수) 와 알려진 모집단의 분포에 의해서 예상되는 빈도(기대도수)가 통계적으로 유의한지 살펴보자 :)

In [10]:

```
# 진단과 radius_mean이 20이 넘는지 안넘는지의 열(new)을가지고 통계적으로 유의미한지 보았을때 관찰값 Ob = X.values[1,:2]
# 관측 비율 (여기서는 기존에 알려진 비율이 존재한다고 가정하자) Pr = np.array([0.7,0.3])
n = X.values[1,2]

E = n*Pr
stats.chisquare(Ob,E)
```

Out[10]:

```
Power_divergenceResult(statistic=105.0, pvalue=1.2213583807296 62e-24)
```

- p밸류가 0.05를 넘어감, 그니까 귀무가설H0을 그대로 살려두고 , H1인 대안가설을 죽임(기각)
- 관측치와 예상치(기대)는 다르다고 말할 충분한 근거가 없다.
- 관측도수와 기대도수의 차이는 통계적으로 의미 없다~
- 모집단의 분포와 표본의 분포는 차이가 없다. (차이가 있는지 보려고 했으니까)

사실 의미가 없는게

적합도 검정 조건

- 범주형 변수 값의 갯수를 알 때 (단순 랜덤 표본에 해당하는 값이어야 함)
- 범주형, 명목형, 연속형 데이터에는 적합하지 않음.
- 관측된 각 데이터 범주에서 최소 5개의 값이 기대될 정도의 사이즈.

가 되어야 한다.

물론 이게 아니어도 상관은 없다.

예를들어서 공을 뽑는다고 하자

1. 흰공 3 검은공 4 빨간공 2
2. 총 9개의 공을 뽑았다.
3. 90개의 공이 30개씩 담겨있었을 것이라고 예상할때
4. 뽑은 공의 색깔에 대한 분포와
5. 기대하는 분포 1/3 이 같은지 보는 것이다.
6. 딱봐도 다르네....
    - 이때 귀무가설 H0 은 변수X(공 색깔)의 관측 분포와 기대분포가 동일하다
    - 이때 대안가설 H1 은 변수X(공 색깔)의 관측 분포와 기대분포가 동일하지 않다.

   동일하지 않으니까 ㄱ,,,,귀무가설 기각

카이제곱 검정/독립성 검정 (Chi-square Test)/(Test of Independence)

독립성검정은

- 범주형인 두 변수가 서로 연관되어 있는지 여부를 검정
- 연속형 변수들 사이의 관계를 알아보는 상관분석이 있다면,
- 범주형 변수에는 독립성검정이 있습니다.

예를 들어봅시다.

성별과 흡연여부의 관계를 알고 싶어서 임의로 200명을 추출하여 성별 및 흡연여부를 조사하였습니다.

	흡연	비흡연	합계
남성	46	33	79
여성	25	96	121
합계	71	129	200

- 귀무가설 : 변수 X성별와 Y흡연유무는 서로 독립이다.

- 대립가설 : 변수 X성별와 Y흡연유무는 서로 독립이 아니다.

궁금하면 한번 해보시는걸로

In [11]:

```
# 새로운 열이 추가된 df를 그대로 써보자 df.head()
```

Out[11]:

Aa 제목	# 속성	# id	diagnosis	# radius_mean	# texture_mean	# perimeter_mean	# area_mean	# smoothness_mean	# compactness
제목 없음	0	842302	@2021년 9월 20일	17.99	10.38	122.8	1001	0.1184	0.2776
제목 없음	1	842517	@2021년 9월 20일	20.57	17.77	132.9	1326	0.08474	0.07864
제목 없음	2	84300903	@2021년 9월 20일	19.69	21.25	130	1203	0.1096	0.1599
제목 없음	3	84348301	@2021년 9월 20일	11.42	20.38	77.58	386.1	0.1425	0.2839
제목 없음	4	84358402	@2021년 9월 20일	20.29	14.34	135.1	1297	0.1003	0.1328

크로스탭도 그대로 쓰자 .... 범주형 변수가 없잖아 ? 데이터를 다른걸 쓰든지 해야지

대신 이번에는 margins= False

In [12]:

```
# 크로스탭 쿼리를 통해 빈도교차표를 출력한다.  X = pd.crosstab(df.new, df.diagnosis, margins=  False )
# Margins = 행과 열의 총합을 표시 할 것인가?# 이번에는 하지말자  X
```

Out[12]:

Aa 제목	diagnosis	# B	# M
제목 없음	@2021년 9월 22일	0	0
제목 없음	@2021년 9월 13일 오전 12:00	357	167
제목 없음	@2022년 1월 1일	0	45

In [13]:

```
stats.chi2_contingency(X)
```

Out[13]:

```
(79.39733176671615,
5.079466303860378e-19,
1,
array([[328.76625659, 195.23374341],
       [ 28.23374341,  16.76625659]]))
```

- 두 변수가 독립적인가 ? : 독립성 검정
- 위에 있는 부분이 카이제곱 검정 통계량
- 그 바로 밑에가 유의 확률 (10의 -19승을 곱해야 하니까 엄청나게 유의하다고 볼 수 있음)
- 1은 자유도
- 밑에는 각 셀에따른 기대빈도, 기대도수가 나타난다. (잘보면 표랑 비슷)

결과적으로 new 와 diagnosis 는 통계적으로 유의하게 연관성이 있다. = 독립적이지 않다.

카이제곱 검정/동질성 검정 (Chi-square Test)/(Test of Homogeneity)

동질성검정은 독립성검정처럼 변수가 2개입니다.

독립성검정이 두 변수의 관계를 알기 위해 하는 검정이지만,

동질성검정은 두 변수의 관계를 알기 위해 하는 검정은 아닙니다.

- 동질성검정은 한 변수의 요인들에 관심이 있습니다.
- 요인 보다는 그룹이라고 하는 것이 이해하기 쉽습니다.
- 각 그룹들이 동질한지 알고 싶은 것입니다. 여기서 동질하다는 것은 확률분포가 같다는 것입니다.
- 서울 부산, 인천 경기 와 같이 모집단이 다른곳에서 얻은 데이터가 통계적으로 (비율적으로) 유의미한지 보는 것입니다.
- 모집단 다른 두 표본 / 동질적이냐 이질적이냐 ?

In [14]:

```
# 크로스탭 쿼리를 통해 빈도교차표를 출력한다.  X = pd.crosstab(df.new, df.diagnosis, margins=  False )
# Margins = 행과 열의 총합을 표시 할 것인가?# 이번에는 하지말자  X
```

Out[14]:

Aa 제목	diagnosis	# B	# M
제목 없음	@2021년 9월 22일	0	0
제목 없음	@2021년 9월 13일 오전 12:00	357	167
제목 없음	@2022년 1월 1일	0	45

빈도교차표를 출력하고

In [15]:

```
stats.chi2_contingency(X)
```

Out[15]:

```
(79.39733176671615,
5.079466303860378e-19,
1,
array([[328.76625659, 195.23374341],
       [ 28.23374341,  16.76625659]]))
```

결과적으로는 귀무가설을 기각하니까 => 차이가 있다 동질적이지 않다.

근데 위에거랑 똑같네 ?

그럼 독립성 검정과 동질성 검정은 뭐가 다를까?

독립성 검정과 동질성 검정의 차이

동립성검정과 동질성검정은 변수에 종류에 의해 결정되지 않습니다.

우리의 '관점'과 '표본 추출 방법'에 의해 결정됩니다.

'성별'과 '흡연여부'를 두개의 변수로 취급할지 아니면 '흡연 여부' 만을 변수로 보고, 성별은 비교 대상이 되는 '그룹'으로 해석할지는 우리가 결정하는 것입니다.

1. 만약 '성별'과 '흡연여부'를 두개의 변수로 보고 독립성검정을 하겠다고 결정했다면, 표본추출은 전체집단 하나만 추출합니다.
2. 만약 '흡연 여부'만을 변수로 보고 남/녀 그룹의 동질성을 비교하고 싶다면, 남자 그룹과 여자그룹의 표본을 각각 추출합니다. 각 표본에서 흡연 여부를 조사합니다.

- 기댓값 : '데이터에 특이하거나 주목할 만한 것이 없다는 의미'로 대략 정의 가능
  - 변수 혹은 예측 가능한 패턴 사이에 상관관계가 없음
- 한 변수가 다른 변수와 독립적인지, 데이터 테이블의 각 셀에 있는 숫자에 의미가 있는지 검정

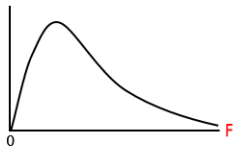
- 카이제곱 통계량 : 검정 결과가 독립성에 대한 귀무 기댓값에서 벗어난 정도를 측정하는 통계량
- 관측값과 기댓값의 차이를 기댓값의 제곱근으로 나눈 값을 다시 제곱하고 모든 범주에 대해 합산한 값
- 일반적으로 카이제곱 통계량은 관측 데이터가 특정 분포에 적합한 정도를 나타냄 (적합도)

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed frequency counts in each category
E_i = expected frequency counts in each category
k = number of categories

- 카이제곱분포 : 귀무 모델에서 반복적으로 재표본추출한 통계량 분포
  - 카이제곱 값이 낮다 : 기대 분포를 거의 따르고 있음
  - 카이제곱 값이 높다 : 기대한 것과 현저하게 다름

2.11 F 분포



〈 F분포 〉

- F 통계량 : 그룹 평균 간 차이가 정규 무작위 변동에서 예상할 수 있는 것보다 얼마나 큰지를 측정하는 것이며 각 그룹 내 변동성에 대한 그룹 평균 간 변동성의 비율을 의미 - 이러한 비교를 분산분석(ANOVA)이라함
- F 분포는 측정된 데이터와 관련한 실험 및 선형 모델에 사용
- F 통계량은 관심 요인으로 인한 변동성과 전체 변동성을 비교
- F 분포 : 정규분포를 이루는 모집단에서 독립적으로 추출한 표본들의 분산비율이 나타내는 연속 확률 분포
- **두 가지 이상 표본집단의 분산을 비교하거나 모집단의 분산 추정할 때 쓰임**
- 2개 이상의 표본평균들이 동일한 모평균을 가진 집단에서 추출되었는지
  아니면 서로 다른 모집단에서 추출된 것인지 판단하기 위하여 이용

2.12 푸아송 분포와 그 외 관련 분포들

- 람다 : 단위 시간이나 단위 면적당 사건이 발생하는 비율
- 푸아송 분포 : 표집된 단위 시간 혹은 단위 공간에서 발생한 사건의 도수분포
- 지수분포 : 한 사건에서 그 다음 사건까지의 시간이나 거리에 대한 도수분포
- 베이불 분포 : 사건 발생률이 시간에 따라 변화하는, 지수분포의 일반화된 버전

2.12.1 푸아송 분포

- 시간 단위 또는 공간 단위로 표본들을 수집할 때, 그 사건들의 분포를 알려줌
- 시간 별 혹은 공간별 사건 발생이 얼마나 다른지 알고 싶을 때

- 핵심 파라미터 : 람다(λ) - 어떤 일정 시간/ 공간의 구간 안에서 발생한 평균 사건 수
- 대기행렬 시뮬레이션을 위한 푸아송 분포를 따르는 난수 생성
    - 서비스센터에 1분당 2회 정도의 문의 전화가 올 때, 100분당 문의 전화 횟수

2.12.2 지수분포

- 람다(λ)를 이용하여 사건과 사건 간의 시간 분포 모델링
    - 톨게이트에 자동차가 도착하는 시간 사이
      고장이 발생하는 시간 모델링, 고객 상담에 소요되는 시간 모델링
- 지수 분포에서 난수를 생성하기 위해서는 n (난수 발생 개수), 비율(시간 주기당 사건 수)을 사용
  분당 평균적으로 0.2회 서비스 문의 전화가 걸려오는 경우, 100분 동안 서비스 센터 문의 전화 시뮬레이션
  주기별 평균 사건수 0.2인 지수분포에서 난수 100개 생성

```
stats.expon.rvs(scale=1/0.2, size=100)   # stats.expon.rvs(scale=5, size=100)
```

```
array([6.05107637e-01, 6.74366343e+00, 1.49636779e+01, 3.35726473e+00,
       1.31471735e+00, 1.39121200e+01, 3.42668301e+00, 1.36143831e+00,
       4.13493472e+00, 2.14941138e+00, 4.65395381e+00, 1.02662017e+00,
       3.96332671e+00, 2.48631257e+00, 4.32289169e+00, 5.99519586e+00,
       2.52139938e+00, 2.34281369e+00, 3.48832566e-02, 4.35052627e+00,
       1.50380571e+01, 9.78579781e+00, 3.56138987e+00, 4.28022221e+00,
       1.02264033e+01, 1.04791593e+01, 1.51364966e+01, 1.17230510e+01,
       7.83290928e+00, 1.28824938e+00, 5.01834883e-03, 6.97939163e+00,
       6.68651696e+00, 1.08169545e+00, 2.66978659e+00, 4.52892841e+00,
       9.58462139e-02, 1.55283838e+00, 3.65561344e-01, 2.33293373e+00,
       8.21293988e-01, 3.60623875e+00, 1.02662265e+01, 5.08266823e+00,
       1.43476638e+00, 5.76542848e-01, 7.52580288e-01, 1.90814150e-01,
       1.22610851e+00, 2.87763083e-01, 3.88136656e+00, 8.27241324e+00,
       1.54442687e+00, 1.39017985e+01, 5.44584499e+00, 5.68701274e+00,
       2.38576202e+01, 3.64795728e+00, 1.55663658e-01, 9.43105872e+00,
       1.77196481e+01, 2.48683256e+00, 1.85048424e-02, 5.84480295e+00,
       2.41604044e+00, 3.06088691e+00, 1.26668873e+01, 4.30152296e-03,
       1.53038423e+00, 7.52662594e+00, 3.06530745e+00, 6.61937201e+00,
       3.54541127e-01, 2.50537807e+00, 6.94588931e-01, 1.38003424e+00,
       7.09803676e+00, 3.60696102e+00, 1.98516369e+00, 6.06011146e+00,
       4.20538138e+00, 2.45324649e-01, 3.65918783e+00, 1.04595742e+00,
       1.75947493e+00, 5.85176161e+00, 9.38146145e-01, 4.12869076e+00,
       8.03939145e+00, 6.24748560e+00, 4.51514655e+00, 1.99043545e+00,
       5.46790883e+00, 2.53322624e-01, 4.18137967e+00, 8.36623623e-01,
       1.77472777e+00, 2.05246473e+01, 1.93046187e+01, 1.79934666e+01])
```

- 푸아송 분포와 지수분포에 대한 시뮬레이션 연구의 핵심은 **람다(λ)가 해당 기간동안 일정하게 유지된다는 가정**
    - 도로 교통상황의 경우 요일과 시간대에 따라서 다르게 나타나는데 이러한 경우 시간 주기 또는 공간을 일정 기간 충분히 동일하도록 영역을 잘 나누면 해당 기간의 분석 및 시뮬레이션이 가능

2.12.3 고장률 추정

- 정말 드물게 일어나는 사건의 경우 예측을 위한 데이터가 거의 없어서 사건 발생률 추정할 근거가 없음
  - 비행기 엔진 고장 등
- 20 시간 후 아무런 일도 일어나지 않았다면, 시간당 발생률이 1이 아니라는 것은 분명히 알 수 있음
- 시뮬레이션 또는 확률의 직접 계산을 통해 다른 가상 사건 발생률을 평가하고, 그 이하로 떨어지지 않을 임곗값 추산
- 데이터가 있으나 정확하고 신뢰할 만한 발생률을 추정하기에 불충분한 경우 적합도검정을 통해서 적용한 여러 발생률 중 어떤 것이 관찰된 데이터에 가장 적합한지 알 수 있음

2.12.4 베이불 분포

- 사건 발생률이 시간에 따라 지속적으로 변하는 경우 지수분포 또는 푸아송 분포는 유용하지 않음
  - 시간이 지날수록 고장 위험이 증가하는 기계 고장
- 지수 분포를 확장한 것으로, 형상 파라미터 β (베타) 로 지정된 대로 발생률이 달라질 수 있음
  - β > 1 이면 발생률은 시간이 지남에 따라 증가
    β < 1 이면 발생률은 시간이 지남에 따라 감소
- 베이불 분포는 사건 발생률 대신 고장 시간 분석에 사용되므로 두번째 파라미터는 구간당 사건 발생률보다 특성 수명으로 표현됨
  - η (에타, 척도변수)
- 형상 파라미터 β = 1.5(시간 흐름에 따라 발생률 증가), scale = 5000(특성 수명), size = 100 (수명)

```
stats.weibull_min.rvs(1.5, scale = 5000, size = 100)
```

```
array([ 4328.8902672 ,  7270.56731164, 10108.96621979,  2247.6555086 ,
        4491.67896451,  2196.77690747,  3026.79256127,  2131.23556399,
        1151.78232775,  4259.9912767 ,  3486.90674314,  9491.10384838,
         710.15946721,  1996.13044213,  7452.90433717,  5254.23464897,
        5917.11283145,   786.17246323,  2273.7653755 ,  6068.25769649,
        8484.21344843,  2907.52770128,  7305.74267807,  5306.84644739,
        3338.51258303, 10660.11805989,   840.89683678,   664.94368732,
        2193.8679954 ,  3973.68808212,  8004.78569336,  1502.74279776,
        2687.31003864,  8132.83603748,   587.57907422,   959.67978287,
       20028.5712402 ,  7811.91141773,  3002.8036799 ,  7028.87623996,
        4999.78182081,  3258.66131088,  3454.70645105,  6468.31243852,
        2732.26626535,  3569.88890395,  7799.07116382,  7145.45756525,
        3180.21301172,  4602.69798704,  2095.108545  ,  5373.14623875,
        6085.48003527,  2045.90679413,  6131.12469475,  5289.13776487,
        1574.88113031,  9660.56349871,  2682.94237077,  3911.0289521 ,
        3564.24004675,  1927.86532771,  7430.6235086 ,  2871.81039717,
        3386.17918643,  3670.61811146,  7405.65651513,  3394.81331199,
        8864.4795025 ,  5473.61465556,  1897.51240103,  1866.57241264,
        1283.53393172,  5892.21361987,  7408.67310898,  5754.5134088 ,
        4374.94716635, 14626.56645349,   631.7554765 ,  2105.78428221,
       12593.01810499,  4197.77977464,  8737.36714082,  3870.68877591,
         519.38259925,  4631.38830936,  3373.68154409,  5024.98710704,
        1050.78108407,   856.61729382,  1881.54410969,  2198.13664835,
        2677.94497894,   595.32061446,  1102.28385923,  4999.91479579,
        3940.40964081,  9830.08223475,  7726.6916701 ,  7916.29518671])
```