8.5 스튜던트 t 분포, 카이제곱분포, F분포

- 정규분포에서 파생된 분포들 → 스튜던트 t, 카이제곱, F
- 위 파생분포들 통계량 분포 라고도 부르며, 나중에 공부할 가설검정에 쓰인다.

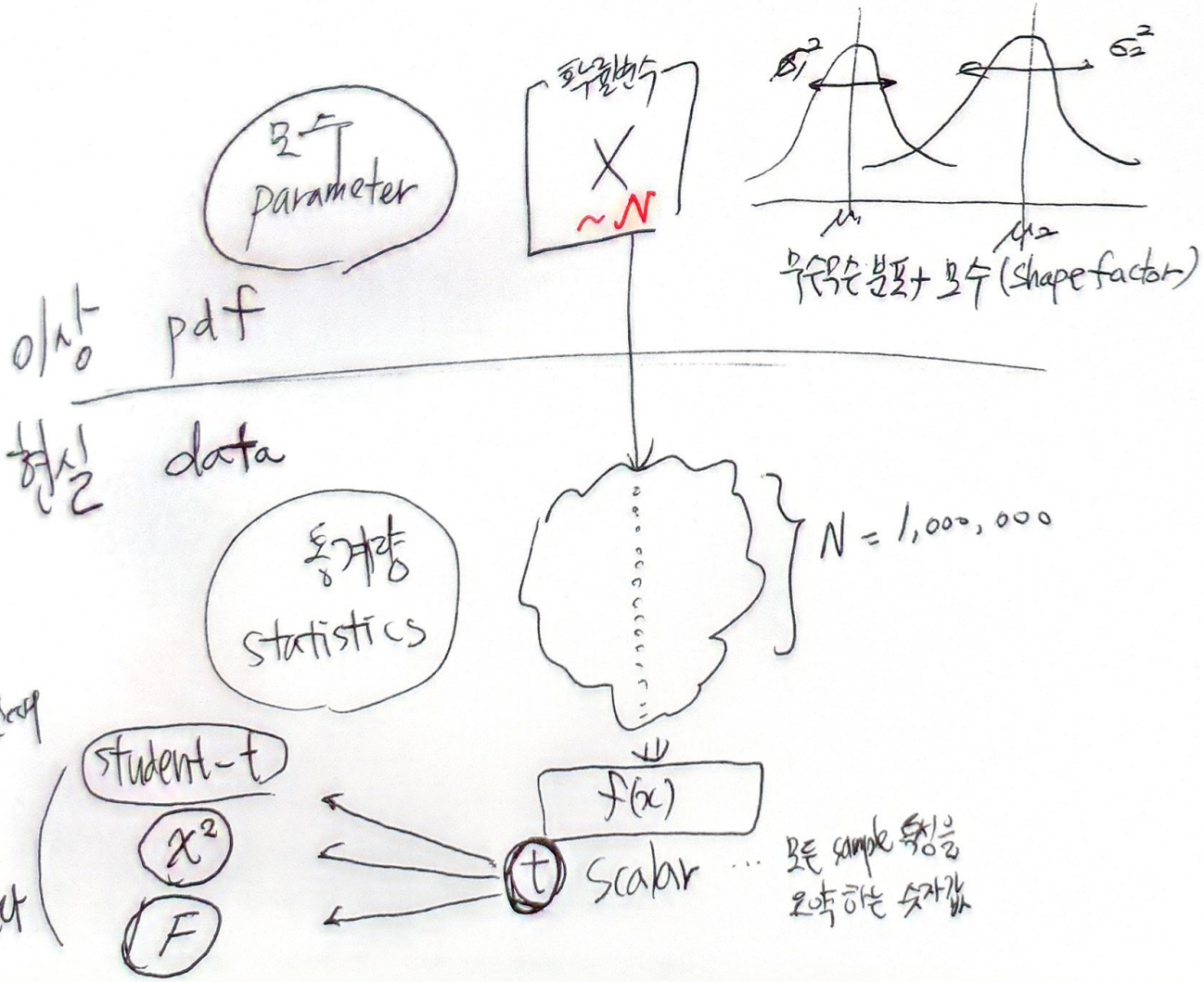
스튜던트 t 분포

현실 데이터 中.. (정규분포와 상당히 유사하지만, 양 끝단의 비중이 정규분포에 비해 더 큰 데이터)

= 정규분포보다 극단적 현상이 더 자주 발생했다는 의미

= fat tail

ex) 주식 수익률은 보통 정규분포를 따른다고 가정
But 정규분포에서는 거의 발생할수없는 극단적 사건 발생
= black swan



스튜던트 t분포 (or t분포) 확률밀도함수

$$t(x;\mu,\lambda,\nu) = \frac{\sqrt{\lambda}\,\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\lambda\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

* λ = 정규분포의 정밀도 $(\sigma^2)^{-1}$ = β에 대응하는 개념
* $\Gamma(x)$ = gamma function

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,du$$

정규분포와 달리, 정수값을 가지는 자유도 (degree of freedom)라는 모수(parameter) ν를 추가적으로 가진다. 스튜던트 t 분포에서는 모수 ν로 2 이상의 자연수를 사용한다.

모수 $\nu = 1$ ⟶ 코시분포 (Cauchy distribution)

코시분포에서 양수인 부분만 사용 ⟶ 하프코시분포 (Half-Cauchy distribution)

t 통계량

(previous) 정규분포의 표본을 표준편차로 나누어 정규화한 t통계량은 항상 정규분포가 된다.

↳ BUT t통계량을 구하려면 확률분포의 정확한 표준편차를 알아야 한다.

--- 현실적으로 불가능!

--- 표본에서 측정한 표본표준편차 (sample standard deviation)으로 정규화를 수행해야 함

t 통계량 = N개 표본 $x_1,\dots,x_N$에서 계산한 표본평균을 표본표준편차로 정규화한 값

t-통계량은 자유도가 N-1인 스튜던트 t 분포를 이룬다.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} \sim t(x; 0, 1, N-1)$$

이 식에서 $\bar{x}$, s는 각각 표본평균, 표본표준편차다.

$$\bar{x} = \frac{x_1 + \cdots + x_N}{N}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

이 정리는 추후 정규분포의 기댓값에 관한 구간 추정에서 사용된다.

→ N이 커질수록 t의 분포는 정규분포에 다가간다.

카이제곱분포

정규분포를 따르는 확률변수 X의 N개의 표본 $x_1, \cdots, x_N$의 합 (또는 차)은 표본분산을 적용하여 스튜던트 t 분포를 따른다는 것을 배웠다.

그런데 이 N개의 표본들을 단순히 더하는 것이 아니라 제곱하여 더하면 양수값만을 갖는 분포가된다. 이 분포를 카이제곱 (chi-squared) 분포라고 하며, $x^2(x; \nu)$로 표현한다. 카이제곱분포도 스튜던트 t 분포처럼 자유도 ν를 갖는다.

$$x_i \sim N(\alpha)$$

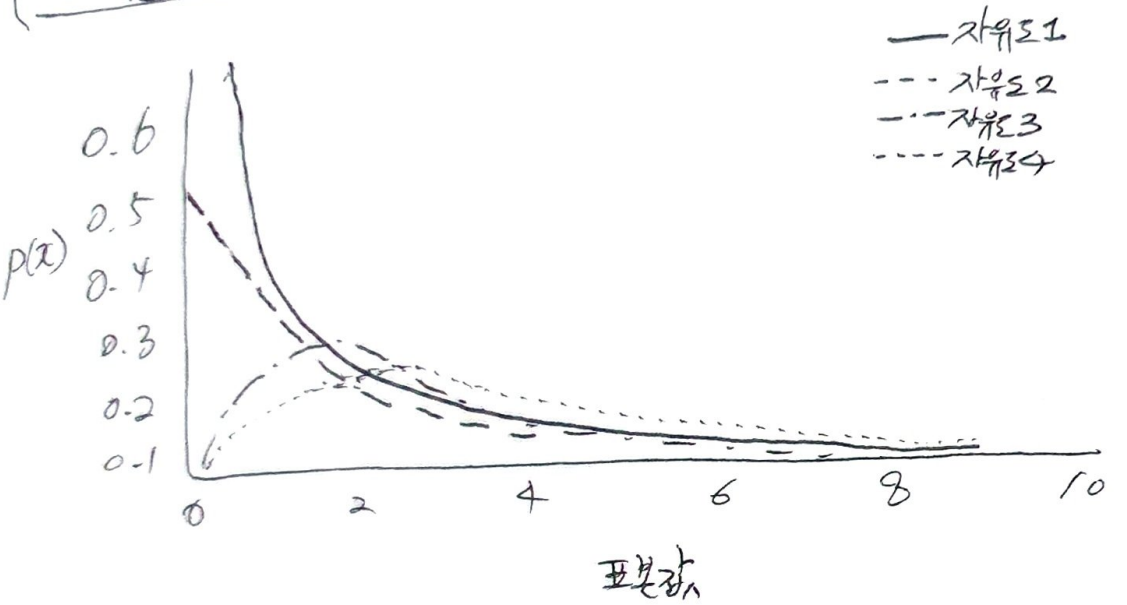
$$\downarrow$$

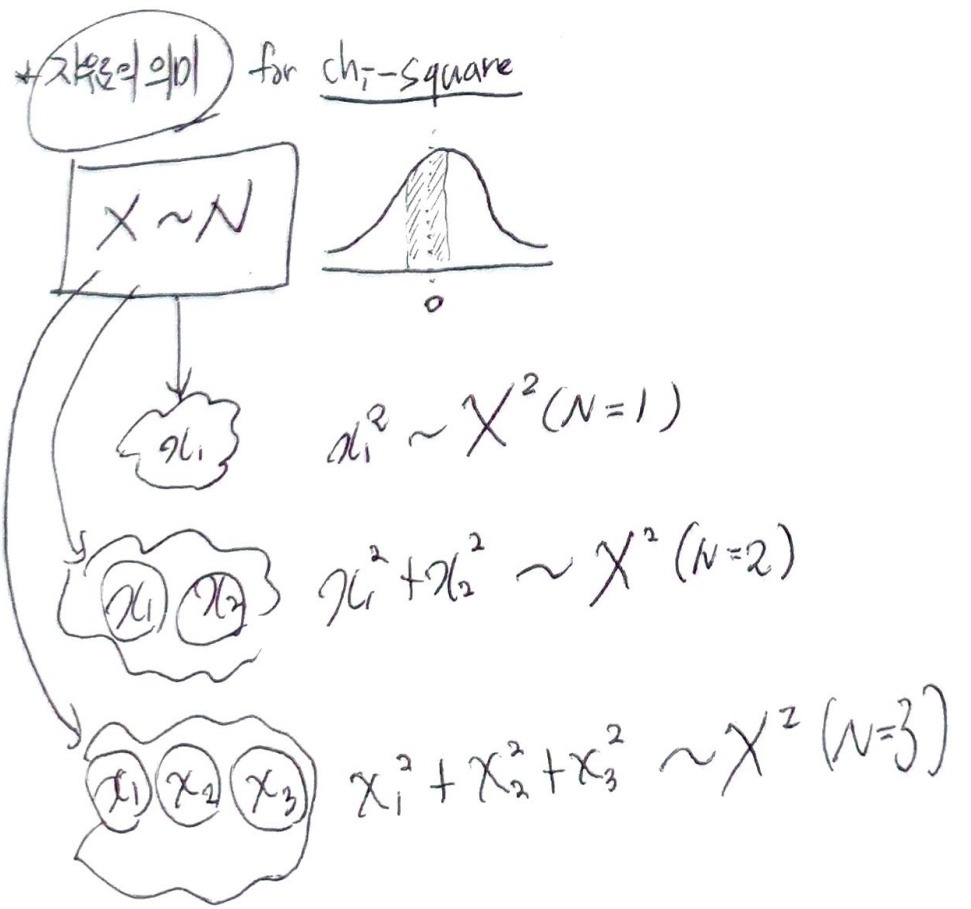
$$\sum_{i=1}^{N} x_i^2 \sim x^2(x; \underset{\parallel}{\nu}) \quad N$$

카이제곱분포의 확률밀도함수는 다음과 같다.

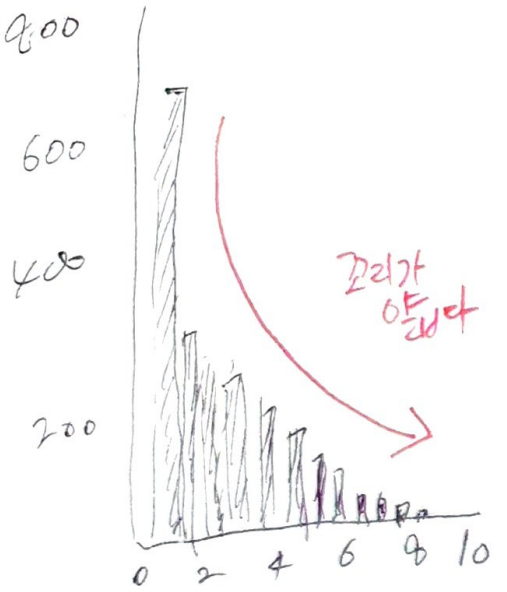
$$\chi^2(x;\nu) = \frac{x^{(\nu/2-1)} e^{-x/2}}{2^{\nu/2}\,\Gamma\left(\frac{\nu}{2}\right)}$$

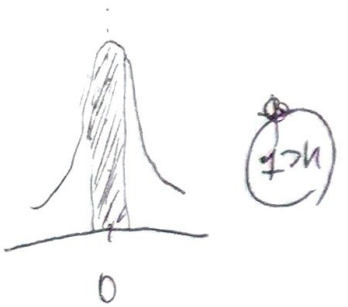
scipy.stats.chi2()



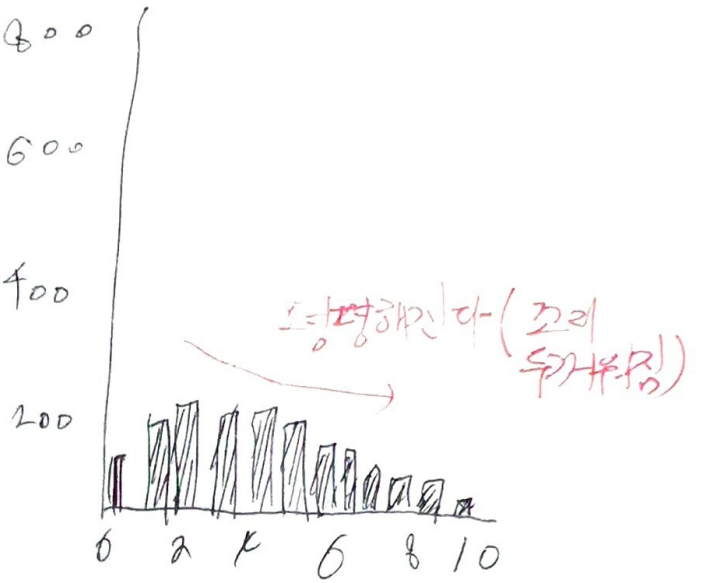


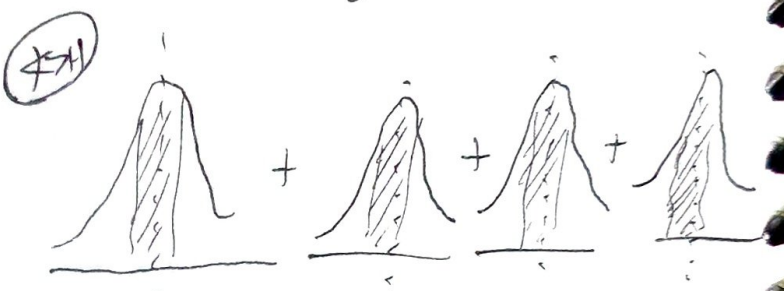
제곱합의 분포 (N=1)



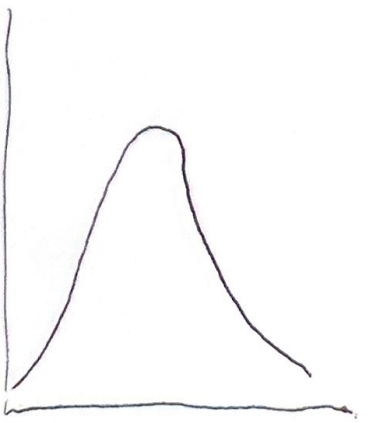


제곱합의 분포 (N=4)





N = 6 ···· (정규분포의 정점 개수 하나) ··· N = 30



F 분포

(스튜던트 t, 카이제곱분포) —— 정규분포를 따르는 확률변수 X로부터 나온 N개의 표본에서 만들 수 있다.

- 카이제곱분포를 따르는 독립적인 두 개의 확률변수 $\chi_1^2(x;N_1)$과 $\chi_2^2(x;N_2)$의 확률변수 표본을 각각 x_1, x_2라고 할 때, 이들 각각 N_1, N_2로 나눈 뒤 비율을 구하면 $F(x;N_1,N_2)$ 분포가 된다. N_1, N_2는 F분포의 자유도 모수라고 한다.

$$x_1 \sim \chi^2(N_1),\ x_2 \sim \chi^2(N_2) \rightarrow \frac{\frac{x_1}{N_1}}{\frac{x_2}{N_2}} \sim F(x;N_1,N_2)$$

F 분포의 확률밀도함수는 다음과 같다.

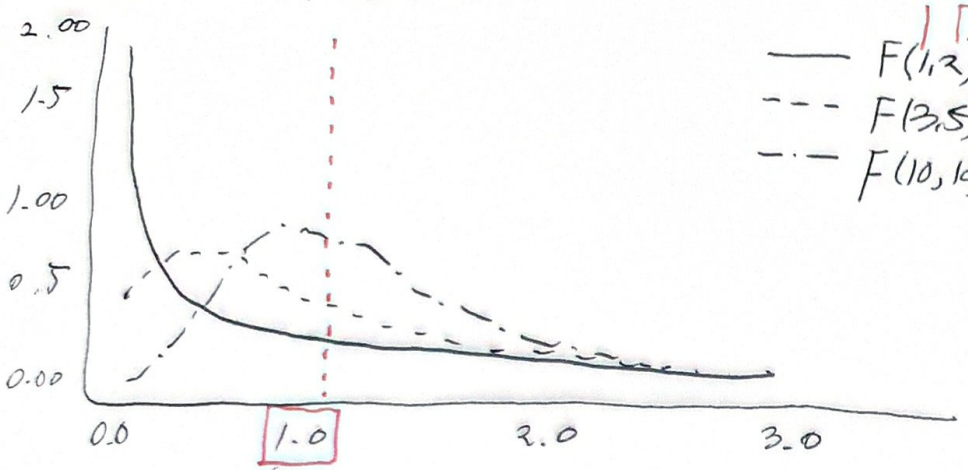
$$f(x;N_1,N_2) = \frac{\sqrt{\frac{(N_1x)^{N_1}\,N_2^{N_2}}{(N_1x+N_2)^{N_1+N_2}}}}{x\,B\left(\frac{N_1}{2},\frac{N_2}{2}\right)}$$

$B(x)$ = 베타, 특수함수

* 스튜던트 t 분포의 표본값을 제곱한 값은 F분포를 따른다.

$$t(N)^2 = F(1,N)$$

<자유도에 따른 F분포의 모양>



↳ 동일한 분포(카이제곱)
에서 2개의 확률변수표본을 가져와 나눈다면,
직관적으로 생각했을 때 1이 많이 나오지 않을까?
→ 자유도가 높아질수록 최빈구간이 1에 다가설 확률↑

활용

스튜던트 t분포, 카이제곱분포, F분포는 모두 정규분포의 통계량분포(statistics distribution)의 일종이다. 선형회귀분석에서 이 통계량 분포들은 각각 다음 값에 대한 확률모형으로 사용된다.

- 스튜던트 t분포 : 추정된 가중치에 대한 확률분포
- 카이제곱분포 : 2차 제곱합에 대한 확률분포
- F분포 : 비교 대상이 되는 선형모형의 2차제곱합에 대한 비율의 확률분포