2. *기술통계*

2-1 표본에 따른 통계량

■ 대표값(대표지표 – 위치 지표)

- ✓ 평균 – 참고: ($\bar{X} \geq G \geq H$)
  - 산술평균: n 이 충분히 크고 분포가 대칭일 때
  - 기하평균: $\mathrm{G} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$
    - 대칭이 아닌 경우, 변수를 변화시켜 평균을 구하는 방법
    - 인구변동율, 물가변동율, 이자율 등 변화율, 비율 등에 사용된다.
  - 조화평균: $\mathrm{H} = \frac{n}{\sum(\frac{1}{x_i})}$
    - 시간적으로 계속 변화하는 변량, 속도 등에 사용되는 대표값
    - 단위당 평균 산출등에 이용되나, 거의 사용되지 않는다..
- ✓ 중앙값(ME; median)
  - 극단적인 이상치가 있는 경우, 평균보다 더 정확하다.
- ✓ 최빈값(MO; mode)
- ✓ 사분위수

■ 산포(다양성, 변이) 지표

- ✓ 종류
  - 절대적인 분포의 산포도: 범위, 평균편차, 사분편차, 표준편차
  - 상대적인 분포의 산포도: 변이계수, 사분위편차계수, 평균편차계수
- ✓ 절대적인 분포의 산포도
  - 범위: 최대값 – 최소값
  - 평균편차: 평균으로부터의 평균거리 = $\frac{\sum|x_i-\bar{x}|}{n}$
  - 분산(σ^2): $\mathrm{var(X)} = E[(X-\mu)^2] = \frac{\sum(X-\mu)^2}{N}$, 표준편차($\sigma$)
  - 사분위수 범위(IQR; interquartile range)
    - 3 사분위수 - 1 사분위수: 극단적인 값에 영향을 받지 않는다.
    - 대표지표가 중앙값인 경우, 표준편차 대신 사용한다.
    - 사분편차(Quartile Deviation): (3 사분위수 - 1 사분위수)/2
- ✓ 상대적인 분포의 산포도
  - 변동계수(변이계수;Coefficient of variation): $\mathrm{cv} = \frac{s}{\bar{x}} \times 100\%$
    - 평균값, 표준편차 등은 단위에 따라 값이 달라진다.
    - 변이는 단위가 없어서, 단위에 무관하다.
    - 추정통계학에서 표본의 크기를 설정하는데 많이 사용된다.
  - 사분위편차계수 = 사분위편차/중위수
  - 평균편차계수 = 평균편차/(중위수 또는 산술평균)

✓ 적률

- Pearson 이 도입한 것으로 분포에 대한 특성값을 나타내는 데 사용
- K 차 적률 $M_K = \frac{\sum f_i(x-a)^K}{n}$
  - f_i는 계급 도수이다. 계급이 없으면 1
  - N = $\sum f_i$
  - a 는 평균 등 임의의 값이 될 수 있다.
- 평균에 대한 1 차 적률은 0 이다.
- 평균에 대한 2 차 적률은 분산 = 원점에 대한 2 차적률 – 원점에 대한 1 차 적률의 제곱
- 평균에 대해 자료분포가 대칭인 경우 평균에 대한 홀수 적률은 0 이다.

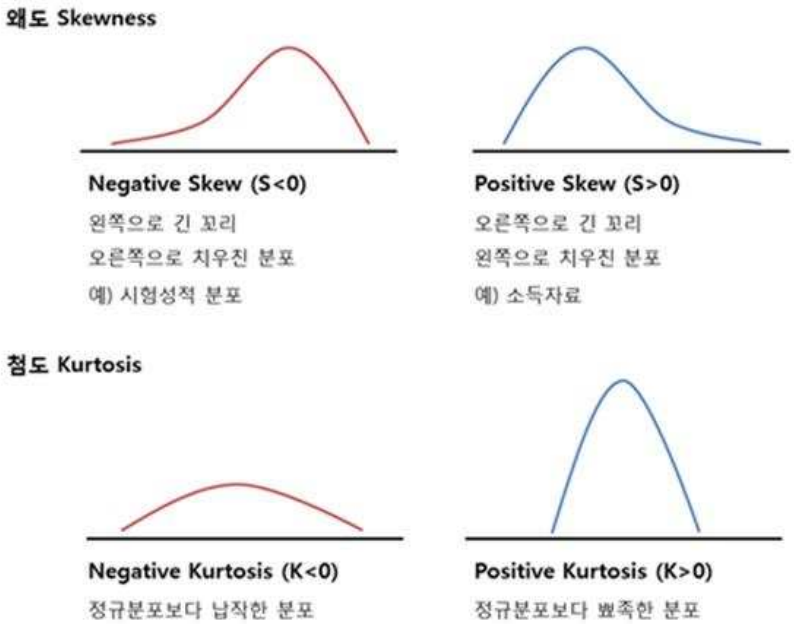
■ 분포의 모양

✓ 왜도(skewness) = $\frac{\sum(x-\bar{x})^3}{ns^3}$

- 분포가 기울어진 방향과 크기를 나타내는 지표
- 좌우대칭: 0 ➔ $\bar{X} = M_e = M_o$
- 왼쪽으로 치우치면(⟅): 양수 ➔ $\bar{X} > M_e > M_o$
- 오른쪽으로 치우치면(∫): 음수 ➔ $\bar{X} < M_e < M_o$
- 왜도는 3 차 적률을 s^3로 나눈 값이다.

✓ 첨도(kurtosis) = $\frac{\sum(x-\bar{x})^4}{ns^4} - 3$

- 분포의 모양이 얼마나 뾰족한가를 나타내는 지표
- 표준정규분포와 같으면: 0 또는 3 이라고도 한다.(정의에 따라)
- 표준정규분포보다 납작하면 음수
- 표준정규분포보다 뾰족하면 양수
- 첨도는 4 차 적률을 s^4로 나눈 값으로 정의하기도 한다.



■ 공분산과 상관계수

✓ 두 개의 확률변수 X, Y 에 대해

- 평균이 각각 μ_1, μ_2이고 표준편차가 σ_1, σ_2이라 하자.
- 만약 X 가 μ_1보다 클 때 Y 가 μ_2보다 커지고, X 가 μ_1보다 작을 때 Y 가 μ_2보다 작아지는 경향이 있으면
  - 표준화된 X 와 Y 의 곱 $\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)$은 양의 값을 가지게 된다.
  - 확률변수 X 의 증감에 따른 확률변수 Y 의 변화방향과 정도의 척도로 $\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)$을 사용한다.
- 상관계수 Corr(X, Y) = $\mathrm{E}\left[\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)\right] = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)\right]$
  - 일반적으로 -1 ≤ Corr(X, Y) ≤ 1 이다.
  - Y = aX + b 인 경우, 완전한 직선관계를 가질 때, a 가 양수면 상관계수는 1, a 가 음수이면 상관계수는 -1 의 값을 가진다.

✓ 공분산 = Cov(X, Y) = $\sigma_1\sigma_2$Corr(X, Y) = $\mathrm{E}[(X-\mu_1)(Y-\mu_2)] = \mathrm{E}(XY) - \mu_1\mu_2$

✓ 공분산과 상관계수의 성질

- Cov(aX+b, cY+d) = acCov(X, Y)
- Corr(aX+b, cY+d)
  - ac > 0, Corr(aX+b, cY+d) = Corr(X, Y)
  - ac < 0, Corr(aX+b, cY+d) = -Corr(X, Y)

✓ 두 확률변수의 합과 차에 대한 분산의 공식

- Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)
- Var(X – Y) = Var(X) + Var(Y) - 2Cov(X, Y)

✓ 두 확률변수가 독립인 경우

- E(XY) = E(X)E(Y) = $\mu_1\mu_2$
- Cov(X, Y) = $\mathrm{E}(XY) - \mu_1\mu_2 = 0$ ➔ Corr(X, Y) = 0
- Var(X ± Y) = Var(X) + Var(Y)

✓ 서로 독립인 두 확률변수 X, Y 에 대해

- 평균의 차의 분포를 생각해 보자.
- 즉, $\bar{X} - \bar{Y}$에 대해 $\frac{[(\bar{X}-\bar{Y})-(\mu_1-\mu_2)]}{\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}}$은 정규분포 N(0, 1)로 근사된다.

2-2 확률과 확률분포

■ 확률

✓ 개념

- 어떤 사건이 일어날 가능성의 정도를 나타내는 척도
- 0 ~ 1 사이의 실수값을 가진다.
- 표본공간: 한 실험에서 나타날 수 있는 모든 가능한 결과의 집합(S)

- 표본점: 표본공간 S 의 원소
- 사상 또는 사건: 표본공간의 부분집합

✓ 확률계산

✓ 확률변수

- 실험의 결과들이 수치이며, 그 값이 원소에 따라 변하면서 확률에 따르는 경우 확률변수라 부른다.
  - 이산확률변수: 특정 값만을 취하는 경우
  - 연속확률변수: 어떤 구간내의 임의의 값을 취하는 경우

✓ 기대값(E)

- 이산: $\sum x \cdot p(x)$
- 연속: $\int_a^b x \cdot p(x)dx$

■ 확률분포

✓ 개념

- 표본공간에서 나타나는 모든 값들과 그 값에 대응하는 확률을 동시에 표시한 것

✓ 확률변수의 기대값

- 실험을 지속적으로 반복했을 때, 평균적으로 기대할 수 있는 값으로 확률변수의 중심화 경향을 나타낸다.
- 기대값 $E(X) = \sum X \cdot P(X)$
- 기대값의 성질(a 는 상수, X/Y 는 확률변수)
  - E(a) = a
  - E(aX) = aE(X)
  - E(X+Y) = E(X) + E(Y)
  - E(X -Y) = E(X) – E(Y)
- 분산의 성질
  - V(a) = 0
  - V(aX) = a^2V(x)
  - V(X+Y) = V(X) + V(Y) + 2Cov(X, Y)
  - V(X-Y) = V(X) + V(Y) - 2Cov(X, Y)
  - V(X) = $\sum[X - E(X)]^2 \cdot P(X) = \sum X^2 \cdot P(X) - [E(X)]^2$

✓ 확률분포의 유형

- 이산확률분포: 이항분포, 프아송분포, 초기하분포, 기하분포, 다항분포
- 연속확률분포: (표준)정규분포, 지수분포, t-분포, F-분포, χ^2분포

2-3 이산확률분포

■ 이항분포(BINOMIAL)

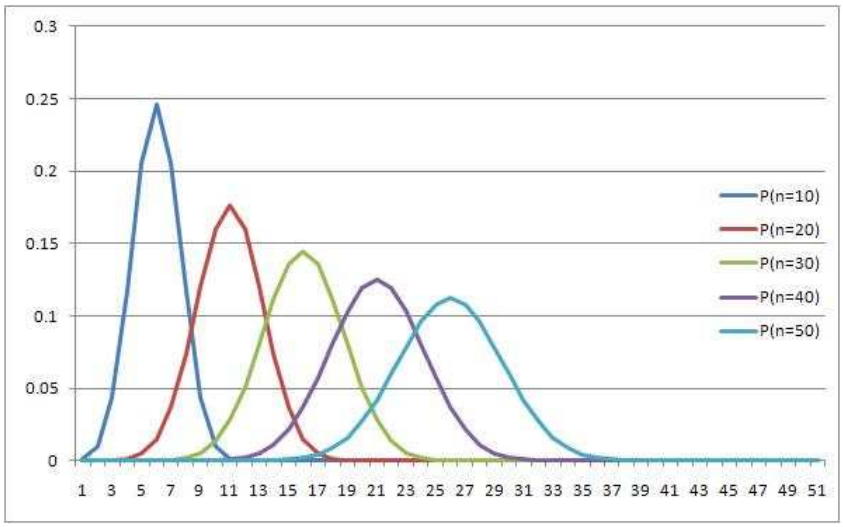
✓ 경우의 수가 단 두개(p, 1-p)

✓ n 번 시행시, A 가 x 번 발생할 확률 B(n, p) = $\binom{n}{x}p^x(1-p)^{n-x}$

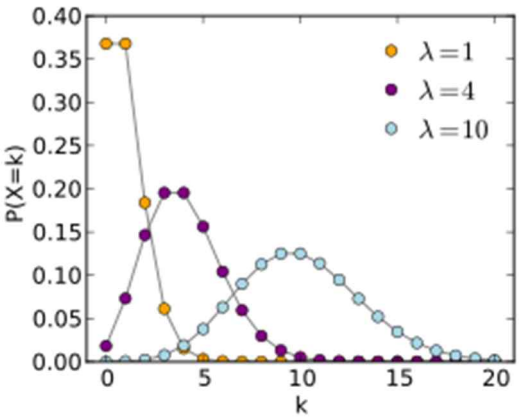
- 베르누이 시행을 n 번 반복한 것이라 할 수 있다.

✓ 특징

- 평균은 np, 분산은 npq (=np(1-p))
- p 가 1 또는 0 에 가깝지 않고 n 이 충분히 크면 정규분포에 가까워지고, 대칭에 접근한다.
- p 가 0.5 가 되면 좌우 대칭의 산모양이 된다.



■ 프아송분포(POISSON)



✓ 단위시간당, 단위공간당 사건발생 횟수에 적용되는 분포

- 주어진 시간, 정해진 공간에서 일어나는 성공의 횟수

✓ 분포함수 p(x) = $\frac{e^{-\lambda}\lambda^x}{x!}$

- x = 0, 1, 2, … n
- λ = 단위시간 또는 단위공간 내의 발생횟수의 평균(λ=np)

✓ 프아송 분포의 성립조건

- 독립성: 주어진 시공간에서 일어날 사건의 횟수는 다른 시공간에서 일어나는 사건의 횟수와 독립이다.
- 비례성: 어떤 시공간 내에서 사건이 한번 발생할 확률은 그 공간의 길이 또는 면적에 비례한다.
- 비집락성: 짧은 시공간 내에서 사건이 동시에 두 번 일어날 확률은 0 에 가까워 무시할 수 있다.

✓ 특징

- 확률변수의 **기대값과 분산은 모두 λ 이다.**
- 이항분포의 특수형태이다. (n 이 매우 크고, p 가 매우 작을 때)

✓ 응용분야

- 주어진 시간 동안에 도착한 고객의 수
- 주어진 생산시간 동안 발생하는 불량의 수
- 1 킬로미터 도로에 있는 흠집의 수
- 하룻동안 발생하는 출생자의 수
- 어떤 시간 동안 톨게이트를 통과하는 차량의 수
- 어떤 페이지 하나를 완성하는 데 발생하는 오타의 발생률
  - 400 글자당 오타가 2 개 발생한다고 하자
  - 한 페이지당 400 글자가 들어간다면, 임의의 페이지에 오타가 한 개 있을 확률은
    ① 단위(한페이지)당 사건(오타)가 발생하는 횟수는 2
    ② P(X=1) = $\frac{e^{-2}2^{1}}{1!}$ = 0.27

■ 초기하분포(HYPERGEOMETRIC)

✓ 개념

$$X \sim Hypergeometric(N,m,n)$$

$$P(X=x)=f(x;N,m,n)=\frac{{}_{m}C_{x}\,{}_{N-m}C_{n-x}}{{}_{N}C_{n}},\quad (x=max(0,n+m-N),\cdots,min(n,m))$$

x: number of successes in sample,
N: number of population$(=1,2,\cdots)$
m: number of successes in population $(=0,1,2,\cdots,N)$,
n: number of sample $(=1,2,\cdots,N)$

- 크기 N 인 유한모집단(성공 수 M, 실패수 N-M)에서 비복원으로 n 개의 표본을 취할 때 확률변수 X(표본내의 성공횟수)가 나타내는 분포
- 성공 확률이 매회 일정한(독립사건; 복원) 경우는 이항분포를 사용
  - 복원추출이거나 모집단의 수가 무한한 경우
- 일정하지 않은 경우(종속사건; 비복원)에는 초기하분포를 사용한다.
  - 비복원추출이며 모집단의 크기가 작은 경우

- 유한모집단의 크기 N 이 추출개수 n 보다 매우클 때 초기하분포는 이항분포로 접근한다.

✓ 특성

- 평균 = np (p = M/N)
- 분산 = $np(1-p)\cdot\frac{N-n}{N-1}$

✓ 사용 예

- 전체 52 장 카드(붉은 색 26, 검은 색 26)에서 10 장을 뽑을 때,
  - 붉은 색이 7 장일 확률
  - $\frac{\binom{26}{7}\cdot\binom{26}{3}}{\binom{52}{10}}$ = 0.10811

■ 기하분포(GEOMETRIC)

✓ 개요

- 단 한번의 성공을 위해 실패를 거듭해야 하는 경우
  - 각 시행은 독립시행으로 각 시행에서의 성공확률 p 는 항상 동일
  - 베르누이 시행을 처음으로 성공할 때까지 시행횟수 x 의 확률분포

✓ 확률분포함수

- $P(X=x) = p\cdot(1-p)^{x-1}$ (x = 1, 2, ...)
  - $P(X>r) = (1-p)^r$
  - $P(X\le r) = 1 - (1-p)^r$

✓ 특징

- 평균: E(X) = 1/p
- 분산: $V(X) = {}^{q}/_{p^2}$

✓ 사용 예

- 주사위를 던져 6 번만에 1 이 나올 확률: ${}^{1}/_{6}\cdot({}^{5}/_{6})^{5}$=0.067

■ 음이항분포(NEGATIVE BINOMIAL)

✓ 개념

- 성공확률이 p 인 베르누이 시행을 독립적으로 반복 실행할 때, k 번 성공할 때까지의 시행횟수 X 의 확률분포

2-4 연속확률분포

■ 정규분포

✓ 특징

- 평균과 표준편차에 의해 모양과 위치가 결정된다.
- 첨도는 0, 평균은 0, 표준편차가 1 인 분포가 표준정규분포이다.
- 산술평균 = 중위수 = 최빈수

- 개별치의 확률분포가 정규분포가 아니더라도 표본의 크기가 클수록 표본평균의 분포는 정규분포와 가까워진다.

✓ 정규분포곡선
- 왜도가 0 이고, 첨도는 (정의방법에 따라)3 이다.

✓ 표준편차 범위내 분포 확률
- 1σ 범위내 ➔ 0.683
- 2σ 범위내 ➔ 0.954
- 3σ 범위내 ➔ 0.997

✓ 정규분포 표본의 평균($\bar{X}$)이 갖는 분포 : $\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$ (중심극한정리)

■ 지수분포

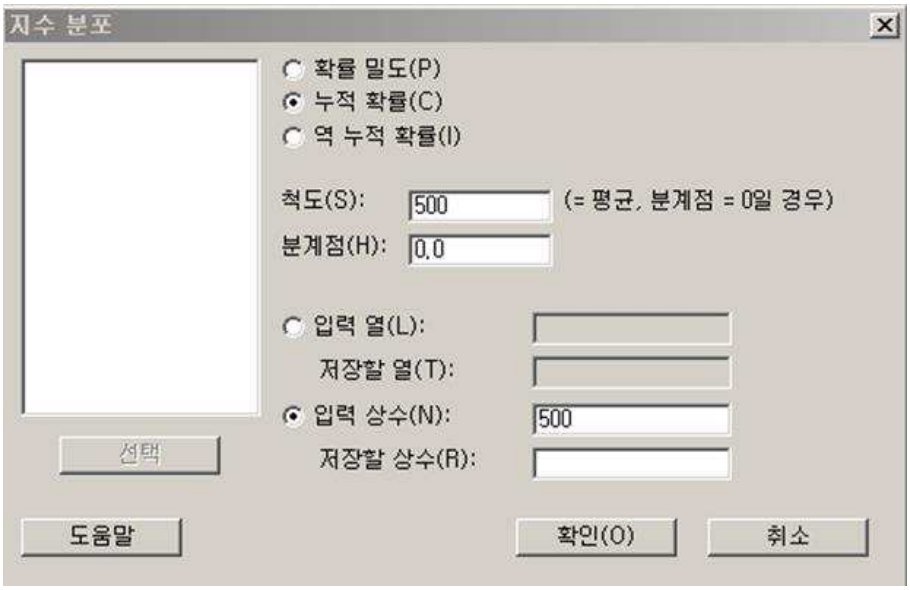
✓ 개념
- 프와송분포가 단위 시공간 내에 사건이 발생횟수에 대한 분포라면,
- 지수분포는 한 사건이 발생한 이후, 다음 사건이 발생할 때까지의 시간에 대한 확률분포이다.
  - 단위 시공간당 평균 발생건수 = λ 라면,

✓ 확률밀도함수
- $f(t, \lambda) = \lambda e^{-\lambda t}$ $(t \geq 0)$

✓ 특징
- 기대값 $E(t, \lambda) = 1/\lambda$
- $V(t, \lambda) = 1/\lambda^2$
- 누적분포함수 $F(t, \lambda) = 1 - e^{-\lambda t}$

✓ 예제 1
- 한 사무실에서 전화가 평균 10 분당 5 번 걸려온다.
- 이 사무실에서 전화가 걸려온 때부터 다음 전화가 걸려올 때까지 걸리는 시간을 분으로 측정하는 확률분포를 구하라.
  - 1 분당 0.5 회 전화가 온다. ➔ λ = 0.5
  - $f(t, 0.5) = 0.5e^{-0.5t}$
- 다음전화가 올 때까지 걸린 시간이 5 분 이내일 확률
  - $F(t, 0.5) = 1 - e^{-0.5\times5} = 0.918$
- 다음전화가 올 때까지 걸린 시간이 5 분 이상일 확률
  - 1 – 0.918 = 0.082

✓ 예제 2:
- 남송전자가 생성하는 형광등의 수명은 λ = 0.002(단위: 시간)의 지수분포를 따른다.
- 수명이 500 시간을 못 미칠 확률은? 0.6321

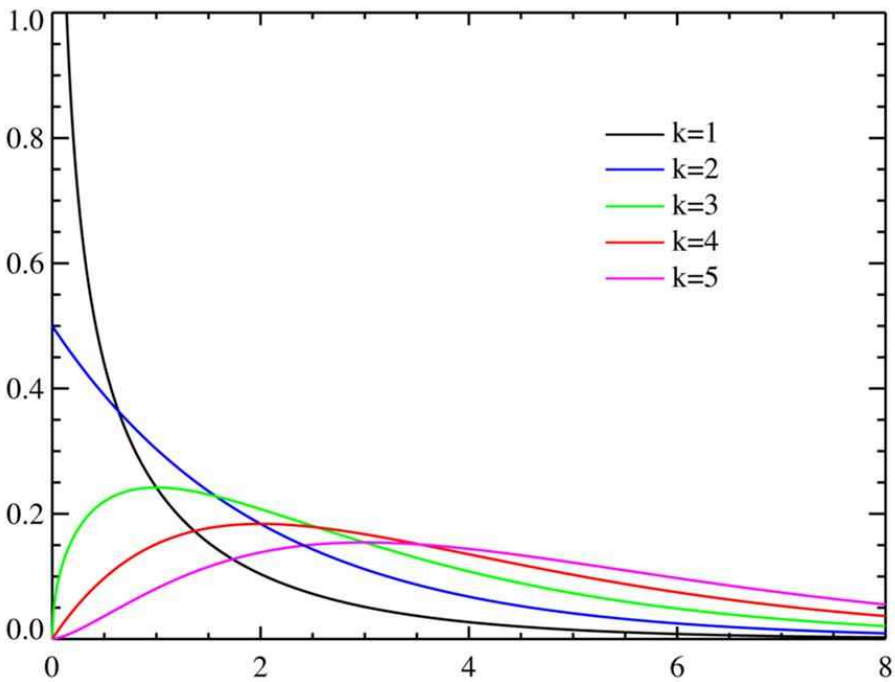


- 평균은 1/ λ = 500 시간이다.
- 척도는 500 으로 하고, 누적확률로 설정된 상태에서 입력상수는 500 으로 계산한다.

· 200 시간이 지난 후 500 시간 동안 고장이 없을 확률

- 뒤 설정에서 분계점을 200, 입력상수를 200+500=700 설정
- 역시 0.6321 이다. ➔ 출발시작부터 어떤 시간에 처음 발생할 확률이 같은 지수분포의 특성이 잘 나타난다.

■ 카이제곱분포



✓ 개요

· 정규분포를 따르는 모집단에서 추출한 표본분포 중에 카이제곱분포, t-분포, F-분포가 있다. (표본의 분포라 불림)

· k 개의 서로 독립적인 표준정규확률변수를 각각 제곱하여 합해서 얻어지는 분포로서 분산과 관련된 분포이다.

- 확률변수 Z_i가 N(0, 1)의 랜덤표본일 때, $\chi^2(k) = \sum_{i=1}^{k} Z_i^2$을 자유도가 k 인 카이제곱 분포라 한다.
- 정규분포의 표본 X_i에 대해서 : $\chi^2(k-1) = \frac{\sum_{i=1}^{k}(X_i-\bar{X})^2}{\sigma^2} = \frac{(k-1)s^2}{\sigma^2}$
  정의에 의해 $s^2 = \frac{\sum_{i=1}^{k}(X_i-\bar{X})^2}{k-1}$ 이며, μ 가 아닌 $\bar{X}$ 가 사용되었다.

· 위 그림과 같이 확률분포함수가 왼쪽으로 기울어진 이유는 Z_i가 0 을 중심으로 분포하기 때문이다.

· 이때, 카이제곱분포의 (1-α) 분위수를 $\chi_\alpha^2(k)$로 나타낸다.

- V ~ $\chi^2(k)$일 때, $P[V \geq \chi_\alpha^2(k)] = \alpha$이다.

✓ 확률분포함수 : gamma(지수/프아송 분포 매개변수의 prior) 분포의 특수예.

· 확률밀도함수: $f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} 1_{\{x\geq 0\}}$

· 누적분포함수: $f(x;k) = \frac{\gamma(k/2,\ x/2)}{\Gamma(k/2)} = P(k/2,\ x/2)$

✓ 특징

· 분포의 형태가 좌측으로 기울어진 분포이고, 자유도가 커질수록 정규분포에 접근한다.

· 여러 집단 사이의 독립성 검정과 적합도 검정을 하는데 사용된다.

· 표본의 산포로 모집단의 산포를 추정할 때 사용한다.

- $P(\chi^2(k) \geq \chi_\alpha{}^2(k)) = \alpha$
- $P(\chi^2(k) \leq \chi_\alpha{}^2(k)) = 1-\alpha$

· 기대값 = k ← 표준정규분포의 분산값이 1 이므로

· 분산 = 2k

✓ 카이제곱분포의 가법성

· 두 확률변수가 서로 독립이고, 각각 카이제곱분포를 따를 때, 이들의 합도 카이제곱분포를 따르는 성질

· $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$이고, 서로 독립이면, $V_1 + V_2 \sim \chi^2(k_1 + k_2)$

✓ **정규모집단에서의 표본 평균 갖는 분포**

· X1, …, Xn 이 정규분포 $N(\mu, \sigma^2)$에서의 랜덤표본이라 할 때,

· 카이제곱분포의 정의로부터 $\chi^2(n) = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$ 이다.

· 그런데, $\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{\sigma^2} + n\frac{(\bar{X}-\mu)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$

· $\chi^2(n) = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$이고, $\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)}$이므로 가법성에 의해

· $\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \chi^2(1)$ 이다. 즉, $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ 이다.

✓ **분산이 같고 서로 독립인 두 정규 모집단 X~ $N(\mu_1, \sigma^2)$, Y~ $N(\mu_2, \sigma^2)$**

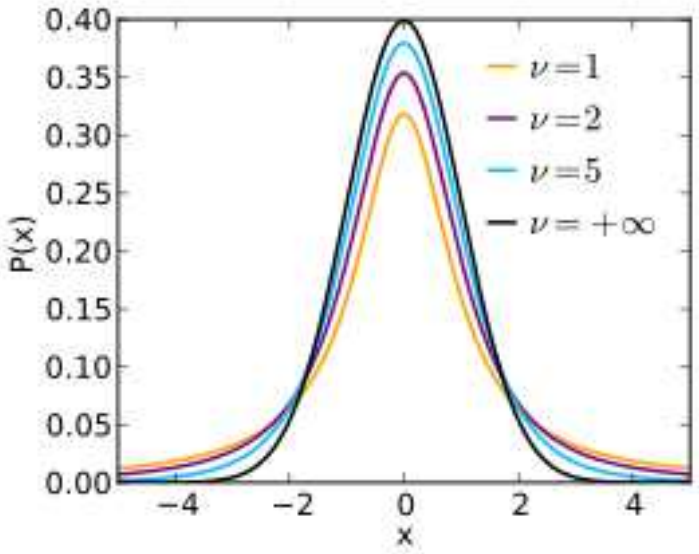
· X1, …, Xn_1과 Y1, …, Yn_2이 각각 정규 모집단의 랜덤 표본일 때,

· 카이제곱분포의 가법성에 의해 $\frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$

· **합동표본편차**(pooled sample variance) $\boxed{{S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{(n_1+n_2-2)}}$일 때,

- $\frac{(n_1+n_2-2){S_p}^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$이 성립한다.
  - 여기서 ${S_p}^2$는 공통분산 σ^2의 추정량으로서
  - 모분산이 σ^2인 두 랜덤표본을 이용하여 만든 것이다.

■ t-분포

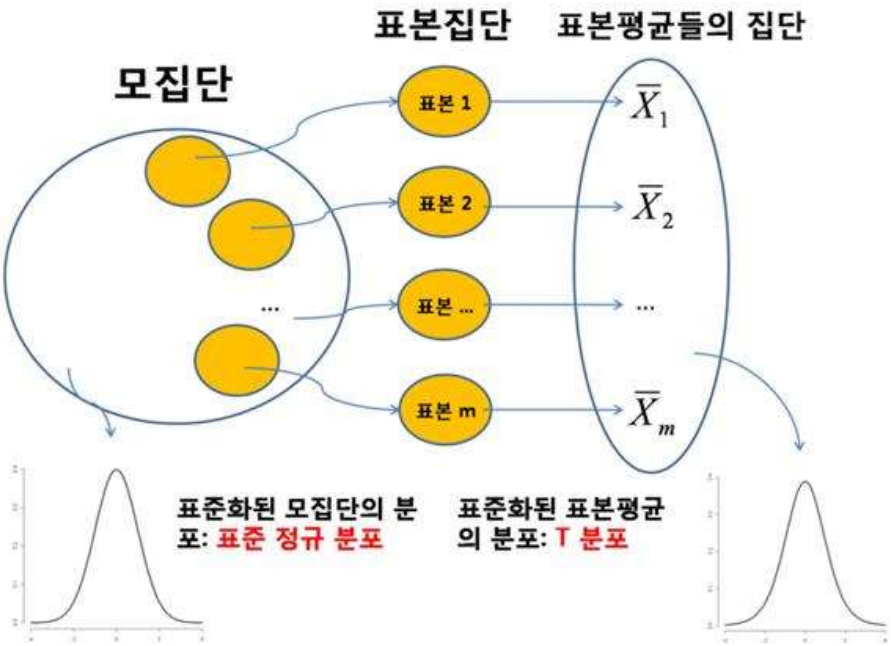


자유도에 따른 t-분포함수의 모양

✓ 정의: $t_k = \frac{Z}{\sqrt{V(k)/k}}$

✓ 개념

- **표본집단의 평균이 가지는 분포를 나타낸다.**
- 모집단에서 표본크기 n 인 표본을 m(여러) 개 추출했을 때, m 개의 표본평균의 평균이 그리는 분포함수가 t-분포함수이다.
- 확률변수$t_{n-1} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ → 자유도 n-1 의 t 값으로 변형하여 계산한다.
  - $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\sigma}{S}Z = \frac{Z}{\sqrt{V(n-1)/n-1}} = \mathrm{t(n-1)}$ (단, $\bar{X}$와 S^2은 서로 독립)
  - 정규모집단의 분산을 모를 때, 정규 모평균을 추정하기 위해 사용한다.



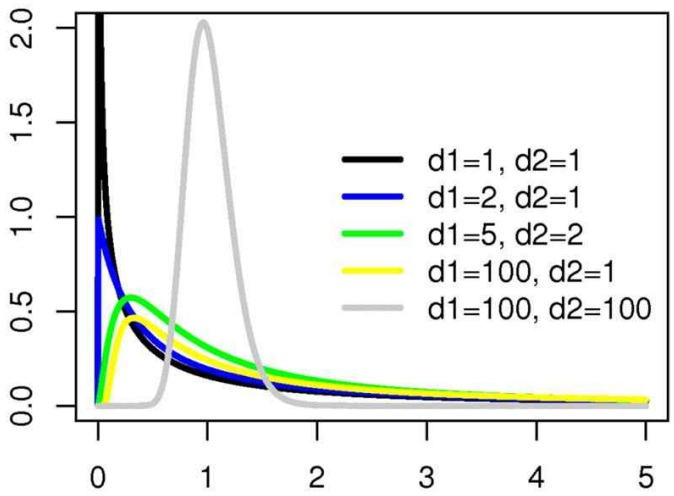
✓ 특징

- n 에 따라 그 모양이 변하며, 0 을 중심으로 좌우대칭이다.
  - 이때 자유도는 n-1 이다. ➔ 분산에서 (n-1)로 나누는 것과 같다.
  - n 이 충분히 크면 표준정규분포에 접근한다.
- 표본의 크기 n 이 30 보다 작을 때,
  - 모평균, 모평균의 차, 회귀계수의 추정이나 검정에 사용된다.
  - 모집단의 표준편차를 모르고, 모평균을 추정할 때 사용한다.
- t 분포의 분산값은 n/(n-2), n<2 인 경우는 ∞

✓ 분산이 동일한 두 정규모집단에서의 t 분포

- $\sigma_1 = \sigma_2 = \sigma$인 경우, $\frac{[(\bar{X}-\bar{Y})-(\mu_1-\mu_2)]}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim Z(0,1)$가 된다.
  - $\bar{W} = (\bar{X}-\bar{Y})$라 하면, $V(\bar{W}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$이므로...
- 이때, $\frac{[(\bar{X}-\bar{Y})-(\mu_1-\mu_2)]}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$는 ➔ $\frac{[(\bar{X}-\bar{Y})-(\mu_1-\mu_2)]}{\left(\frac{Sp}{\sigma}\right)\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{Z}{\sqrt{\frac{V(n_1+n_2-2)}{(n_1+n_2-2)}}} = t(n_1+n_2-2)$
  - 여기서 ${S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{(n_1+n_2-2)}$일 때,
  - $\frac{(n_1+n_2-2){S_p}^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$ 이다.

■ F-분포



✓ 개념

- 두 확률변수 V1, V2 가 각각 자유도 k1, k2 이고 서로 독립인 카이제곱분포를 따를 때, F(k1, k2) = $\frac{V_1/k_1}{V_2/k_2}$의 분포
- 따라서, F(k1, k2) = $\frac{{S_1}^2}{{S_2}^2}\frac{{\sigma_2}^2}{{\sigma_1}^2}$,
- 분산이 동일한 두 정규모집단에서의 F 분포
  - $\sigma_1 = \sigma_2 = \sigma$인 경우, $F(n_1-1, n_2-1) = \frac{{S_1}^2}{{S_2}^2}$

✓ 특징

- 두 정규모집단에서 확률로 추출한 표본으로부터 구한 두 표본분산과 두 모분산의 동일성 여부에 대한 추론에 사용한다.
  - X, Y 두집단의 분산이 동일하다는 가설 Ho: $\frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$
  - σ_1, σ_2는 모른다. s 를 이용, $\frac{{s_1}^2}{{s_2}^2} > F_\alpha(n_1 - 1, n_2 - 1)$이면 H_0 기각
- 주로, 분산분석과 실험계획법 등에 사용된다.
- 확률변수 F(k1, k2)에 대해, F(k1, k2) = 1/F(k2, k1) 이다.(정의에 의해)
  - $F_{1-\alpha}(k_1, k_2) = \frac{1}{F_\alpha(k_2, k_1)}$

✓ 사용예: P(F(15,9)≥?)=0.01 → 계산>확률분포>F-분포

- Fα(15,9)보다 클때의 확률이 0.01 이므로, 누적은 1-0.01=0.99 를 입력 상수값으로 설정해야한다.
- 확률로부터 Fα 를 구해야 하므로, "역 누적 확률"을 적용한다.
- 결과는 4.96 이다.

✓ t-분포와 F-분포의 관계

- $F(1,\ n_1 - 1) = \frac{{Z^2}/{1}}{{V(n_1-1)}/{n_1-1}} = T^2(n_1 - 1)$
- 즉, t 분포는 F 분포의 특수한 형태로 추출될 수 있다.

2-5 기출문제

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
21																				
41																				
61																				