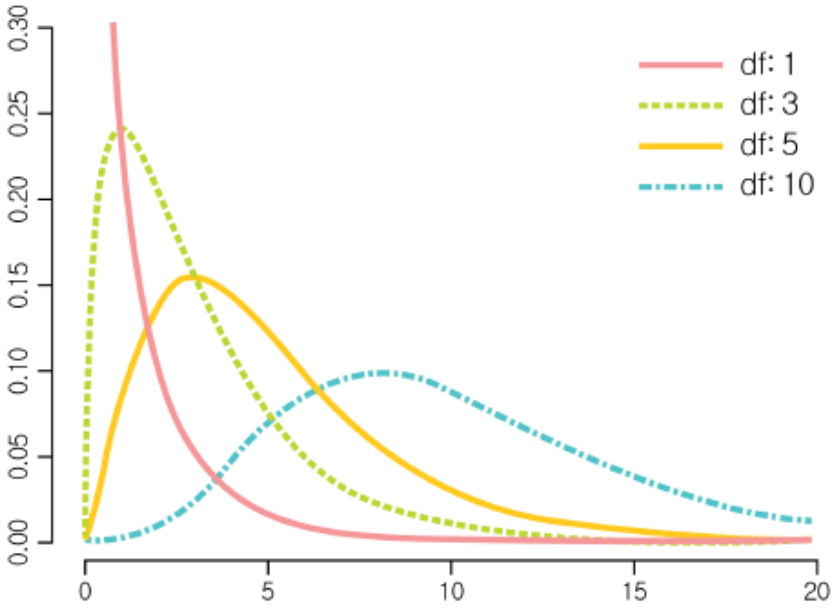
다양한 표본분포

χ^2-분포

- χ^2-분포
  - 표본분산과 관련이 있는 분포
  - χ^2-분포의 모수 : 자유도 k
    - 자유도가 작을 때 꼬리가 오른쪽으로 길게 늘어지는 형태이고,
      자유도가 증가할 수록 정규분포와 유사하게 평균을 중심으로 좌우대칭 형태를 갖습니다.



χ^2-분포

- χ^2-분포
  - 표준정규분포로부터 독립적으로 추출한 k개의 확률표본 $Z_1, Z_2, \ldots, Z_k$ 에 대해,
    - 각각 확률표본의 제곱은 각각 자유도가 1인 χ^2-분포를 따릅니다. $({Z_i}^2 \sim \chi^2(1))$
    - 확률표본들의 제곱의 합 ${Z_1}^2 + {Z_2}^2 + \ldots + {Z_k}^2 = \sum_{i=1}^{k} {Z_i}^2$은 자유도가 k인 χ^2-분포를 따릅니다.
    - 자유도가 k인 χ^2-분포의 기댓값과 분산은 다음과 같습니다. $(X \sim \chi^2(k))$
      - $E(X) = k, \quad Var(X) = 2k$

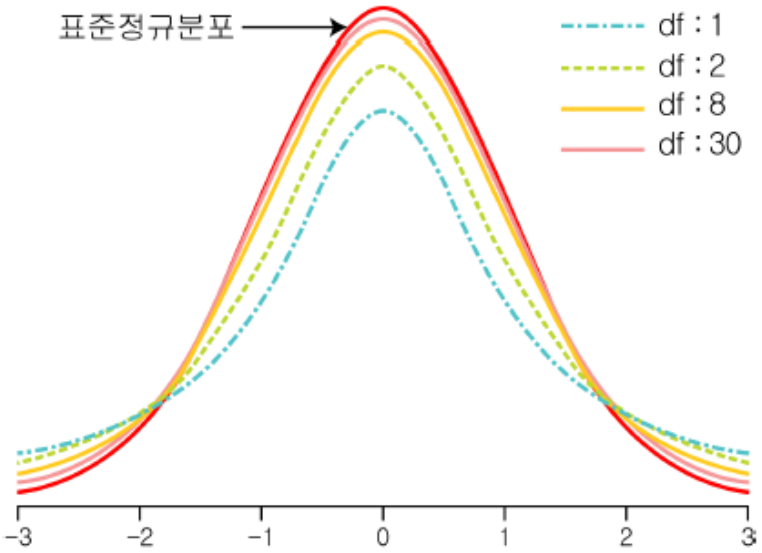
χ^2-분포

- 표본분산과 χ^2-분포
  - ① $X_1, X_2, \ldots, X_n$ 은 정규분포 $N(\mu, \sigma^2)$ 으로부터 추출한 n개의 확률표본입니다.
  - ② 표본평균을 $\bar{X}$, 표본분산을 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 이라 하고, 확률변수 $V = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$ 이라 할 때,
  - ③ ②에 의해, $V = \frac{(n-1)S^2}{\sigma^2}$ 이 됩니다. $((n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2)$
    - V는 자유도가 (n-1) 인 χ^2-분포를 따릅니다. $(V \sim \chi^2(n-1))$
  - ④ 표본분산 S^2의 기댓값
    - $V = \frac{(n-1)S^2}{\sigma^2}$ 으로부터 $S^2 = \frac{\sigma^2}{n-1}V$ 이고 기댓값은 다음과 같습니다.

      $E(S^2) = E\left(\frac{\sigma^2}{n-1}V\right) = \frac{\sigma^2}{n-1}E(V) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2, \quad V \sim \chi^2(n-1), E(V) = n-1$
    - 표본분산 S^2의 기댓값은 모분산 σ^2 입니다.

t-분포

- t-분포
  - 정규분포처럼 평균을 중심으로 좌우 대칭입니다.
  - t-분포의 모수는 자유도이며, 자유도에 따라 분포의 모양이 달라집니다.
    - t-분포는 정규분포와 비슷한 형태지만, 평균 주변에서 상대적으로 밀도가 낮고 양 끝으로 갈수록 꼬리 부분이 두툼한 형태를 갖습니다.
    - 또한 자유도가 증가할수록 표준정규분포를 닮아갑니다.



t-분포

- t-분포
  - 표본평균의 분포와 관련이 있습니다.
  - 두 개의 확률변수 Z와 V가 각각 표준정규분포와 자유도가 k인 χ^2-분포를 따르고($Z \sim N(0, 1^2)$, $V \sim \chi^2(k)$) 서로 독립인 경우 통계량 T를 다음과 같이 정의할 때,

$$T = \frac{Z}{\sqrt{V/k}}$$

  - 통계량 T는 자유도가 k인 t-분포를 따릅니다($T \sim t(k)$).
  - t-분포의 기댓값과 분산 : $X \sim t(k)$
    - $E(X) = 0\,, \quad Var(X) = \frac{k}{k-2}, \ (k > 2)$

t-분포

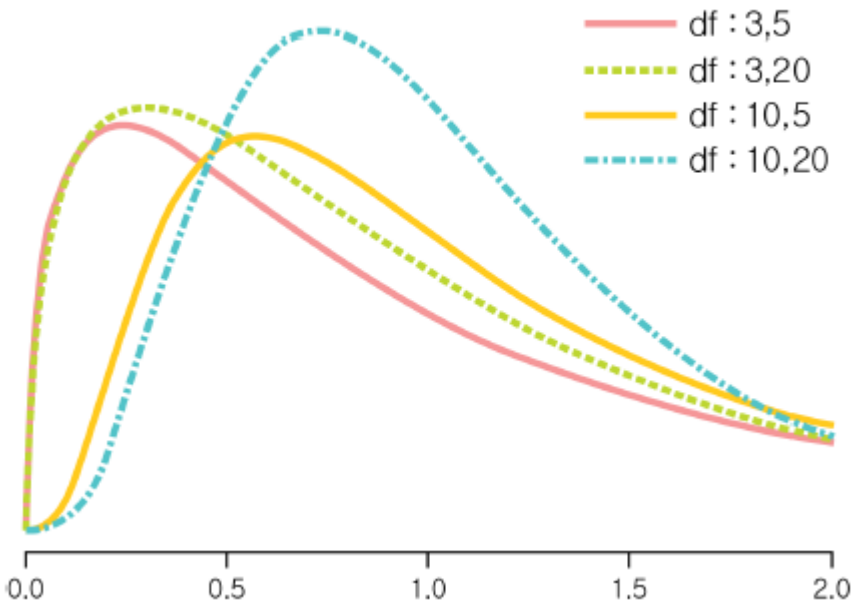
- **표본평균의 분포로써의 t-분포**
  - 정규분포로부터 추출된 n개의 확률표본 $X_1, X_2, \dots, X_n$ 의 평균들의 분포는 정규분포를 따름($N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$)을 앞에서 살펴봤습니다.
    - 표본평균들의 분포는 정규분포를 따르므로 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$
  - 정규분포로부터 추출된 n개의 확률표본 $X_1, X_2, \dots, X_n$의 표본평균을 $\bar{X}$, 표본분산을 S^2(표준편차 S) 이라 하면 다음의 통계량 T는 자유도가 (n-1)인 t분포를 따릅니다.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

    - 정규분포의 분산(표준편차)를 알지 못하는 경우 정규분포를 이용한 계산을 할 수 없어 표본들의 분산(표준편차)을 이용한 t-분포를 사용합니다.
    - 표본의 개수가 클수록 t-분포의 자유도가 증가하여 표준정규분포로 근사한 계산을 할 수 있습니다.

F-분포

- F-분포
  - 두 집단의 분산을 비교할 경우에 유용하게 사용하는 분포입니다.
  - 두 χ^2-분포를 이용하는 분포로 두 χ^2-분포의 모수들을 모수로 사용합니다.



F -분포

- $\boldsymbol{F}$-분포
  - 서로 독립인 두 개의 확률변수 V_1, V_2가 각각 자유도가 k_1, k_2 인 χ^2-분포를 따르고($V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$), 각각의 확률변수를 각각의 자유도로 나눈 통계량 F를 다음과 같이 정의할 때

$$F = \frac{V_1/k_1}{V_2/k_2}$$

  - 통계량 F는 자유도가 (k_1, k_2)인 F-분포를 따릅니다. $(F \sim F(k_1, k_2))$

F -분포

- **두 표본분산 분산비로써로의 F-분포**
  - F-분포는 독립인 두 χ^2-분포의 비율을 이용하는 것으로 두 모집단의 분산 비율을 알고자 할 때 사용할 수 있습니다.
    - 두 개의 정규분포($N(\mu_1, {\sigma_1}^2)$, $N(\mu_2, {\sigma_2}^2)$ 에서 확률표본 $X_1, X_2, \dots, X_n$ 과 $Y_1, Y_2, \dots, Y_m$을 서로 독립으로 추출했을 때, 각 확률표본의 통계량 $V_1 = \frac{(n-1){S_1}^2}{{\sigma_1}^2}$ 은 자유도가 (n-1)인 χ^2-분포를, $V_2 = \frac{(m-1){S_2}^2}{{\sigma_2}^2}$은 자유도가 (m-1)인 χ^2-분포를 따릅니다(${S_1}^2$, ${S_2}^2$은 각 확률표본의 표본분산).
    - 이 때 V_1, V_2가 서로 독립이고 각각을 자유도 (n-1)과 (m-1)로 나눈 값의 비율인 다음의 통계량 F는 자유도가 (n-1, m-1) 인 F분포를 따릅니다. ($F \sim F(n-1, m-1)$)

$$F = \frac{{V_1}/{(n-1)}}{{V_2}/{(m-1)}} = \frac{{\frac{(n-1){S_1}^2}{{\sigma_1}^2}}/{n-1}}{{\frac{(m-1){S_2}^2}{{\sigma_2}^2}}/{m-1}} = \frac{{{S_1}^2}/{{\sigma_1}^2}}{{{S_2}^2}/{{\sigma_1}^2}}$$

F-분포

- $\boldsymbol{F}$-분포
  - 통계량 $F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_1}^2}$에서, $P(a < F < b) = P\left(a < \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_1}^2} < b\right) = P\left(a\frac{{S_2}^2}{{S_1}^2} < \frac{{\sigma_2}^2}{{\sigma_1}^2} < b\frac{{S_2}^2}{{S_1}^2}\right)$
  - 자유도가 k인 t-분포를 따르는 통계량 $T = \frac{Z}{\sqrt{V/k}}$에 대해 다음이 성립하여 $T^2 \sim F(1,k)$ 입니다.

$$T^2 = \frac{Z^2}{V/k} = \frac{Z^2/1}{V/k} = \frac{V_1/1}{V/k}, \qquad Z^2 = V_1 \sim \chi^2(1)$$

  - 확률변수 X가 자유도가 (n, m)인 F-분포를 따를 때 기댓값과 분산
    - $E(X) = \frac{m}{m-2}, \qquad m \geq 3$
    - $Var(X) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}, \qquad m \geq 5$

R에서 χ^2-분포, t-분포, F-분포

함수	시작문자	함수명	함수 형태
확률함수 $P(X=x)$	d	chisq t f	dchisq(x, df) dt(x, df) df(x, df1, df2)
분포함수 $P(X \leq x)$	p	chisq t f	pchisq(x, df) pt(x, df) pf(x, df1, df2)
분위수함수 $P(X \leq x) = q$	q	chisq t f	qchisq(q, df) qt(q, df) qf(q, df1, df2)
난수생성함수	r	chisq t f	rchisq(n, df) **rt(n, df)** rf(n, df1, df2)

R에서 χ^2-분포, t-분포, F-분포

- 자유도가 3인 χ^2-분포 ($X \sim \chi^2(3)$)
  - $P(X \leq 3)$: 기댓값 이하일 확률

```
> pchisq(3, df=3)
[1] 0.6084
```

  - $P(X \leq x) = 0.95$: 어떤 값 이하의 확률이 0.95인지
    - $1 - P(X > x) = 0.95$

```
> qchisq(0.95, df=3)
[1] 7.815
```

  - 자유도가 3인 χ^2-분포로 부터 10개의 난수 추출(표본 추출)

```
> rchisq(10, df=3)
 [1] 1.5351 0.6144 0.4465 2.0851 7.4897
 [6] 1.3151 1.8537 2.4792 6.6432 1.0466
```

R에서 χ^2-분포, t-분포, F-분포

- 자유도가 5인 **t**-분포 ($\boldsymbol{X \sim t(5)}$)
  - $P(X \leq 0)$: 기댓값 이하일 확률 (t-분포는 좌우대칭입니다.)

```
> pt(0, df=5)
[1] 0.5
```

  - $P(-2.571 \leq X \leq 0)$

```
> pt(0, df=5) - pt(-2.571, df=5)
[1] 0.475
```

  - $P(X \leq x) = 0.975$: 어떤 값 이하의 확률이 0.975인지

```
> qt(0.975, df=5)
[1] 2.571
```

  - 자유도가 5인 t-분포로 부터 10개의 난수 추출(표본 추출)

```
> rt(10, df=5)
 [1] -0.4034 -1.7081  0.7053 -0.7799 -0.1970
 [6]  0.2080  1.3272 -0.5517  0.9401  0.4124
```

R에서 χ^2-분포, t-분포, F-분포

- 자유도가 (3, 5)인 $\mathbf{F}$-분포 ($\boldsymbol{F \sim F(3, 5)}$)
  - $P(X \leq 5.409)$

```
> pf(5.409, df1=3, df2=5)
[1] 0.95
```

  - $P(X \leq x) = 0.95$: 어떤 값 이하의 확률이 0.95인지
    - $1 - P(X > x) = 0.95$

```
> qf(0.95, df1=3, df2=5)
[1] 5.409
```

  - 자유도가 (3, 5)인 $\mathbf{F}$-분포로 부터 10개의 난수 추출(표본 추출)

```
> rf(10, df1=3, df2=5)
 [1] 1.4447 0.4273 2.7416 0.6569 0.4024
 [6] 1.0139 0.7866 0.7381 0.2702 0.7502
```