표본 크기와 표준오차

- 무작위로 n개의 압정을 표본으로 뽑아서 그 중 무결함 압정 x개를 골라낸 경우, 표본에서 성공확률은 p에 가까운 값이므로 그것을 $\hat{p}$ 라 하고 'p-HAT' 라고 읽는다.

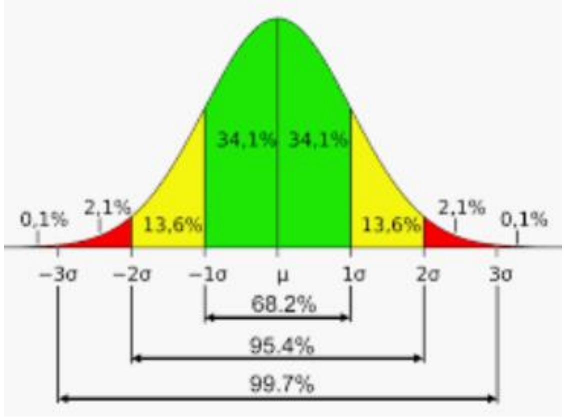
- $\hat{p}$은 그 표본에서 성공 횟수 x 를 표본 크기 n으로 나눈 것이다.

$$\hat{p} = \frac{x}{n}$$

- $\hat{p}$의 측정값들은 p를 중심으로 분포할 것이고,
- 표준편차 또는 분산은 표본 크기에 비례할 것이다.

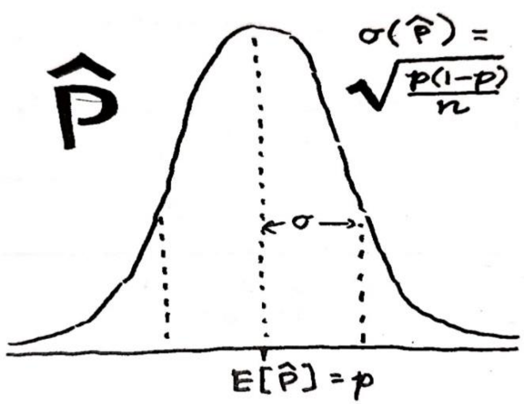
$$\frac{1}{\sqrt{n}}$$

- $\hat{p}$는 정규분포에 가깝기 때문에,
- 기본법칙에 따라 측정값의 68%가 정확한 값 p로부터 표준편차의 범위 내에 위치한다는 결론을 내릴 수 있다.



표본 크기와 표준오차

- 일반적으로 통계는 대부분 4단계로 진행한다
  - 1단계, 미지의 매개변수로 모집단을 정의한다.
  - 2단계, 이론적인 표본분포와 표준편차에 대한 추정치를 구한다.



- $\hat{P}$에 대한 평균, 분산, 표준편차

$\hat{P}$의 평균 $E[\hat{P}] = p$

$\hat{P}$의 분산 $\sigma^2(\hat{P}) = \frac{p(1-p)}{n}$

$\hat{P}$의 표준편차 $\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

n이 크면 $\hat{P}$는 근사적으로 정규분포를 따른다.

  - 3단계, 무작위로 표본을 추출해서 추정치를 찾는다.
  - 4단계, 그 결과와 표본오차를 확인한다.

표본 크기와 표준오차

- 예를 들어, p가 0.85이고 n=1, 4, 16, 25, 100, 10000개의 압정을 표본으로 뽑았다면, 각각의 $\sigma(\hat{p})$은 다음과 같다.

n	1	4	16	25	100	10,000
$\sqrt{n}$	1	2	4	5	10	100
$\sigma(\hat{p})$	.3570	.1785	.0890	.0710	.0357	.0036

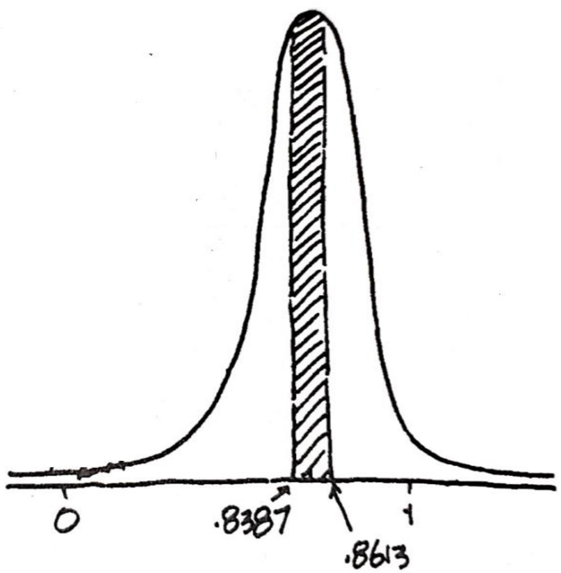
$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

n이 크지면 표준오차는 작아진다.

```
sqrt(0.85*(1-0.85))/sqrt(1)
#[1] 0.3570714
sqrt(0.85*(1-0.85))/sqrt(4)
#[1] 0.1785357
sqrt(0.85*(1-0.85))/sqrt(16)
#[1] 0.08926786
sqrt(0.85*(1-0.85))/sqrt(25)
#[1] 0.07141428
sqrt(0.85*(1-0.85))/sqrt(100)
#[1] 0.03570714
sqrt(0.85*(1-0.85))/sqrt(10000)
#[1] 0.003570714
```

표본 크기와 표준오차

- 예를 들어, p가 0.85이고 n=1000개의 압정을 표본으로 뽑았다면, 아마 우량품은 x=832개쯤 뽑혔을 것이고 $\hat{p}$=0.832가 되고, 표준편차 $\sigma(\hat{p})$ = ?



$$\sigma(\hat{\mathrm{P}}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{(0.85)(0.15)}}{\sqrt{1000}} = 0.0113$$

$$(0.85 - 0.0113) \le \hat{p} \le (0.85 + 0.0113)$$

$$0.8387 \le \hat{p} \le 0.8613$$

측정값의 약 68%가 0.8387 $\le \hat{p} \le$ 0.8613 구간 내에 있다고 예상할 수 있다.

신뢰구간

- 추정치 $\hat{P}$의 표준오차(standard error, SE)는 다음과 같다.

$$\sigma(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad \Rightarrow \quad SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

- 추정치 $\hat{p}$의 분포가 평균 p, 표준편차 σ인 정규분포 인 경우

평균p를 모르고 폭도 $\sigma(p)$의 배수이므로 $\hat{P}$의 표준오차를 쓴다.

$$0.95 = \Pr(\hat{p} - 1.96\sigma(p) \le p \le \hat{p} + 1.96\sigma(p))$$

$$0.95 = \Pr(\hat{p} - 1.96SE(\hat{p}) \le p \le \hat{p} + 1.96SE(\hat{p}))$$

- 이 식은 모집단의 정확한 비율 p가 아래의 확률구간에 있을 확률로 나타낸다.

$$\left(\hat{p} - 1.96SE(\hat{p}), \hat{p} + 1.96SE(\hat{p})\right)$$

- 표본을 여러 번 걸쳐 추출하면 그 중 95%는 이 구간 내에 p가 포함 될 것이다.

신뢰구간

[문제] 여론조사에서 1,000명의 유권자로 구성된 단 하나의 무작위 표본만 취해서

추정치 $\hat{p}$=0.550을 찾아내 평균p를 추정하기

$$\mathbf{SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = \frac{\sqrt{(0.55)(0.45)}}{\sqrt{1000}} = 0.0157}$$

- **p가 아래의 범위 내에 있다고 95% 확신한다.**

$$\mathbf{\hat{p} \pm 1.96SE(\hat{p}) = 0.550 \pm (1.96)(0.0157) = 0.550 \pm 0.031}$$

$$\mathbf{0.519 \le p \le 0.581}$$

- **다시 말하면, p=55%이고 오차 한계는 3%가 된다(조사는 보통 95% 신뢰수준을 택함).**

확률분포(이산, 연속확률분포)

- 확률분포는 크게 이산확률분포(Discrete probability distribution)과 연속확률분포(Continuous probability distribution)으로 나눌 수 있다.

구 분	확률 분포
이산확률분포 (Discrete probability distribution)	이항분포(Binomial distribution), 초기하분포(Hypergeometric distribution), 포아송분포(Poisson distribution)
연속확률분포 (Continuous probability distribution)	정규분포(Normal distribution), t-분포(t-distribution), F분포(F-distribution), 균등분포(Uniform distribution), 카이제곱분포(Chisq-distribution), 감마분포(Gamma distribution)

중심극한정리

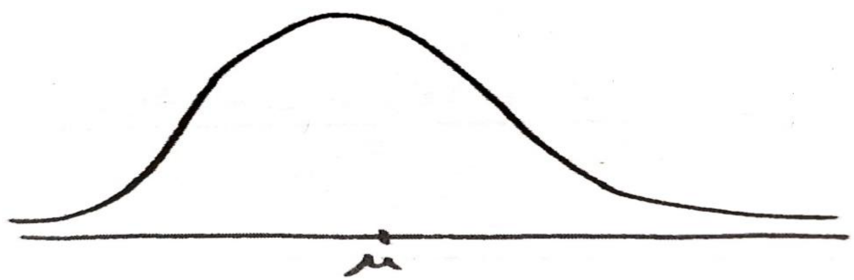
- 평균 μ, 표준편차 σ 인 모집단에서 크기 n인 표본들을 무작위로 추출하면 n이 커질수록 표본 평균의 추정량 $\bar{X}$는 평균 μ, 표준편차 $\frac{\sigma}{\sqrt{n}}$인 정규분포에 가까워진다는 것이 중심극한정리이다.

$$Pr(a \le \bar{X} \le b) = Pr\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}} \le Z \le \frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

μ(=Mu, 뮤)

σ(=Sigma, 씨그마)

- 중심극한정리의 놀라운 점은 본래의 분포 형태와 상관없이,
- 평균들을 모으면 정규분포가 된다는 것이고,
- 표본 평균의 추정량 $\bar{X}$의 분포를 알려면 모집단의 평균과 표준편차만 알면 된다.



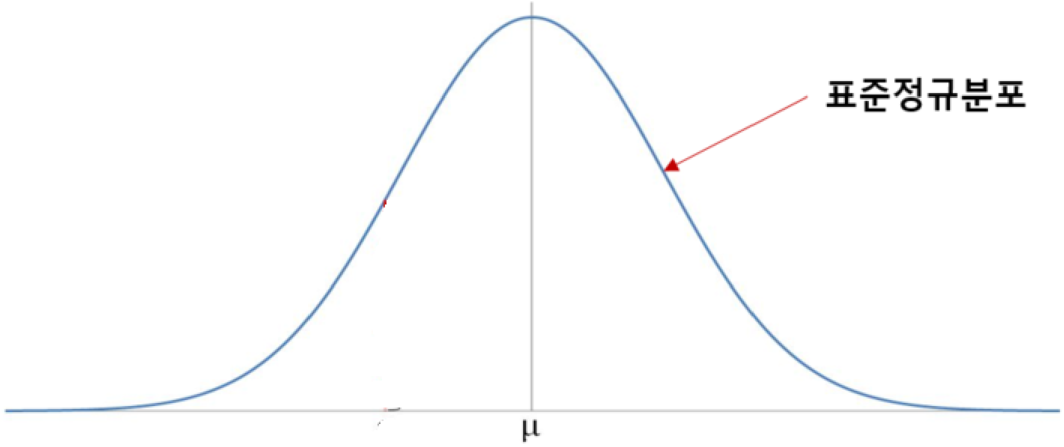
연속확률분포 - 정규분포

- 드무아브르는 좌우 대칭인 종모양의 그래프를 아래와 같은 간단한 식으로 나타냈다.

드무아브르의 식 (de Moivre's formula)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- 평균이 0이고 표준편차가 1인 정규분포 N(0, 1)을 표준정규분포라고 부름(e는 유용한 수학적 상수로서 약 2.718 임)



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

연속확률분포 - 정규분포

- 드무아브르의 식이 정말 종모양인가?

- z가 0에서 멀리 떨어진 값일수록 f(z)는 0에 가깝고,
- f(z)=f(-z)이므로 대칭적,
- z=0에서 최대값을 가지며,
- 이 분포를 평균이 0, 표준편차가 1인 성질을 갖도록 특별히 조정한 것으로 표준정규분포라는 이름이 붙여졌다.

- 요약하면, p=1/2인 이항분포를 '정규화' 하면,
- 즉 중심이 0이고 표준편차가 1이 되도록 만들면, 표준정규분포로 근사 된다는 것

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

  - 다음은 x_i 의 Z-점수를 구하는 식

$$Z_i = \frac{x_i - \overline{x}}{s}$$

x_i: i 번째 x 값,　　$\overline{x}$: x의 평균,
s : x의 표준편차,　Z_i : x_i번째 Z-점수

연속확률분포 - 정규분포

- 정규분포(normal distribution)는 추정과 검정을 하는 추정통계학, 회귀분석과 같은 모형 적합 시 근간이 되는 확률 분포이다.

- 일상 주변에서 흔히 접할 수 있는 확률분포이며,
- 중심 극한의 정리(Central Limit Theorem)에 따라 샘플의 갯수 n이 증가하면 이항분포, 초기하분포, 포아송분포 등의 이산형 확률분포와 t-분포, F-분포 등의 연속형 확률분포가 정규분포로 근사하게 된다.

- 따라서 정규분포는 통계에 있어서 정말 중요하고 많이 사용되는 확률분포라고 할 수 있다.

연속확률분포 - 정규분포

- 다음은 정규분포의 확률 밀도 함수, 누적 분포 함수, 분위수 함수, 난수 발생 등을 위한 R 함수 및 모수는 다음과 같다.

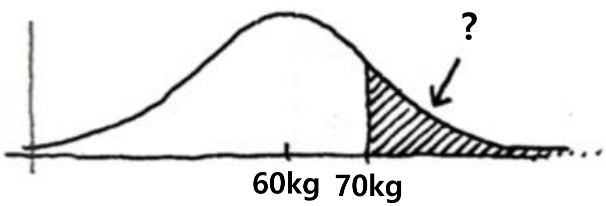
구분		정규분포 R 함수/모수
밀도 함수 Density function	d	dnorm(x, mean=0, sd=1)
누적 분포 함수 Cumulative distribution function	p	pnorm(q, mean=0, sd=1, lower.tail=TRUE/FALSE)
분위수 함수 Quantile function	q	qnorm(p, mean=0, sd=1, lower.tail=TRUE/FALSE)
난수 발생 Random number generation	r	rnorm(n, mean=0, sd=1)

- ✓ lower.tail=TRUE 인 경우, probabilities is $P[X \le x]$.
- ✓ lower.tail=FALSE 인 경우, probabilities is $P[X > x]$.

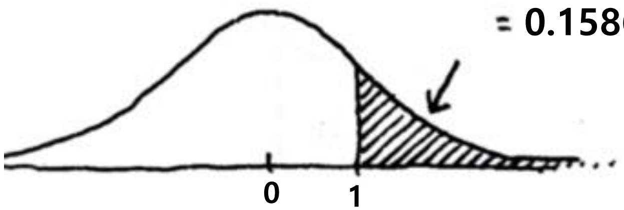
연속확률분포 - 정규분포

- 다음은 학생들의 몸무게가 평균μ=60, 표준편차σ=10인 정규분포라고 가정, 몸무게가 70kg 큰 확률을 계산하는 예제이다.

$$\Pr(X>70) = \Pr\left(\frac{x-\mu}{\sigma} > \frac{70-60}{10}\right) = \Pr\left(z>\frac{10}{10}\right) = \Pr(z>1)$$



Z변환




```
pnorm(1, mean = 0, sd = 1, lower.tail = TRUE)    # 정규분포의 누적분포함수
#[1] 0.8413447
1 - pnorm(1, mean = 0, sd = 1, lower.tail = TRUE)
#[1] 0.1586553
```

- ✓ lower.tail=TRUE 인 경우, probabilities is *P[X ≤ x]*.
- ✓ lower.tail=FALSE 인 경우, probabilities is *P[X > x]*.

$$z = \frac{x-\mu}{\sigma}$$

x : 정규확률변수 X의 실제 값,
μ : 정규확률변수 X의 평균,
σ : 정규확률변수 X의 표준편차,
Z : x의 Z-점수

연속확률분포 - t-분포

- 중심극한정리가 놀랍긴 하지만 최소한 두 가지 문제점이 있다.
  - 첫째: 표본의 크기가 커야 한다.
  - 둘째: 이를 이용하려면 표준편차 σ를 알아야 한다.

- 하지만 표본은 작을 때가 더 많고 σ도 통상 알려져 있지 않을 때가 많으므로,
- 이 경우 표본의 표준편차를 통해 σ를 추정해보는 것이며 표본의 표준편차는 다음 식과 같다.

$$z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

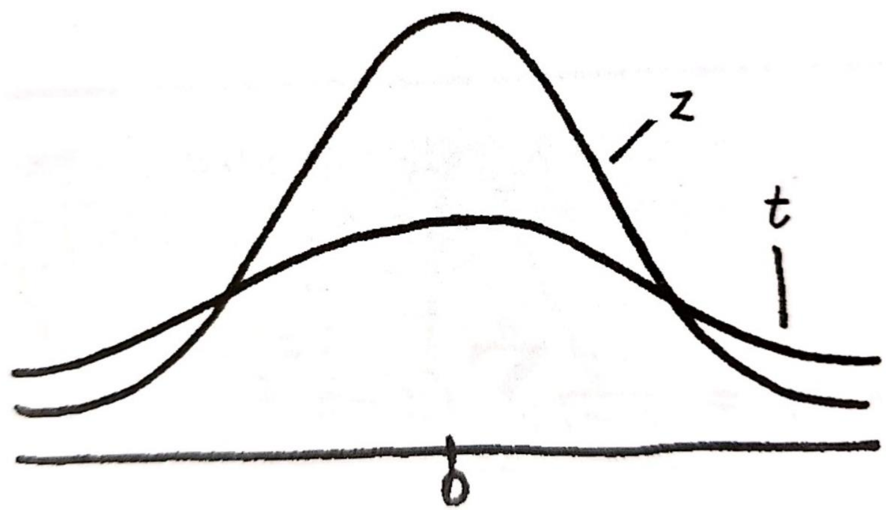
  - Z의 확률변수에서 σ대신 s를 바꿔 넣어 새로운 확률변수 t를 정의한다.

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

연속확률분포 - t-분포

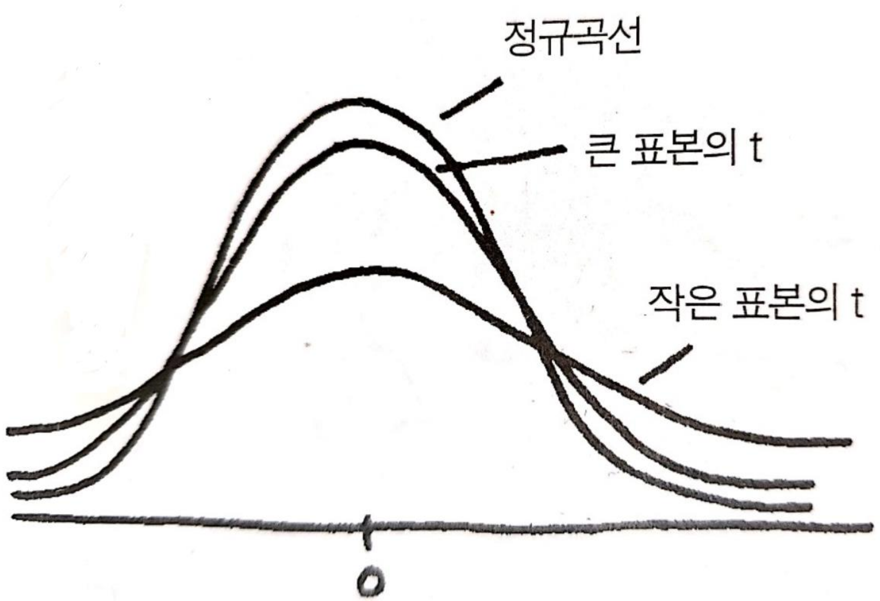
- 확률변수 t를 스튜던트(student) t분포라고도 하는데, 이 개념을 발견한 윌리엄 고셋이 'student' 라는 가명으로 발표했기 때문이다.

- 스튜던트 t는 모집단이 원래 정규분포 또는 정규분포에 가까운 분포였다고 가정, t는 z보다 퍼져 있고, 정규곡선보다 평평하며,
- s를 사용하니까 불확실성이 커져서 t가 z보다 더 느슨해졌기 때문이다.



연속확률분포 - t-분포

- 평평한 정도는 표본 크기에 좌우되고,
- 표본 크기가 클수록 s는 σ에 가까워지고, t도 정규분포인 z에 가까워진다.



연속확률분포 – t-분포

- 다음은 **t분포의 확률 밀도 함수, 누적 분포 함수, 분위수 함수, 난수 발생** 등을 위한 R 함수 및 모수는 다음과 같다.

구분		정규분포 R 함수/모수
밀도 함수 Density function	d	dt(x, df)
누적 분포 함수 Cumulative distribution function	p	pt(q, df, lower.tail=TRUE/FALSE)
분위수 함수 Quantile function	q	qt(p, df, lower.tail=TRUE/FALSE)
난수 발생 Random number generation	r	rt(n, df)

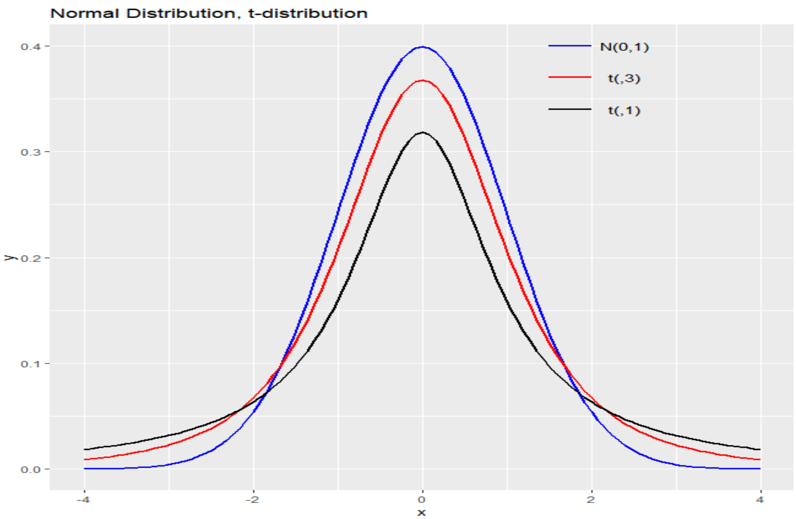
✓ lower.tail=TRUE 인 경우, probabilities is *P[X ≤ x]*.

✓ lower.tail=FALSE 인 경우, probabilities is *P[X > x]*.

연속확률분포 - t-분포

- 다음은 구간이 [-4, 4]인 정규분포의 밀도함수, 자유도가 3인 t분포의 밀도함수, 자유도가 1인 t분포의 밀도함수의 그래프 예제이다.

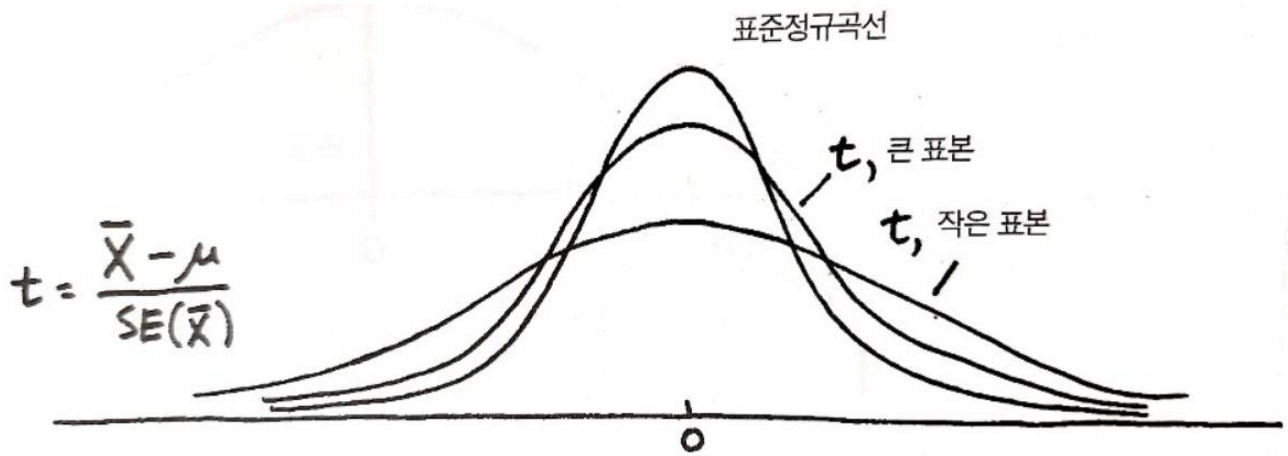
```
library(ggplot2)
ggplot(data.frame(x=c(-4,4)), aes(x=x)) +
  stat_function(fun=dnorm, colour="blue", size=1) +
  stat_function(fun=dt, args=list(df=3), colour="red", size=1) +
  stat_function(fun=dt, args=list(df=1), colour="black", size=1) +
  annotate("segment", x=1.5, xend=2, y=0.4, yend=0.4, colour="blue", size=1) +
  annotate("segment", x=1.5, xend=2, y=0.37, yend=0.37, colour="red", size=1) +
  annotate("segment", x=1.5, xend=2, y=0.34, yend=0.34, colour="black", size=1) +
  annotate("text", x=2.4, y=0.4, label="N(0,1)") +
  annotate("text", x=2.4, y=0.37, label="t(,3)") +
  annotate("text", x=2.4, y=0.34, label="t(,1)") +
  ggtitle("Normal Distribution, t-distribution")
```



연속확률분포 - t-분포

- 통계량은 큰 표본일 때에만 근사적으로 정규분포를 따르고, 작은 표본(n=5, 10, 25 …)일 때는 스튜던트 t를 사용해야 한다.

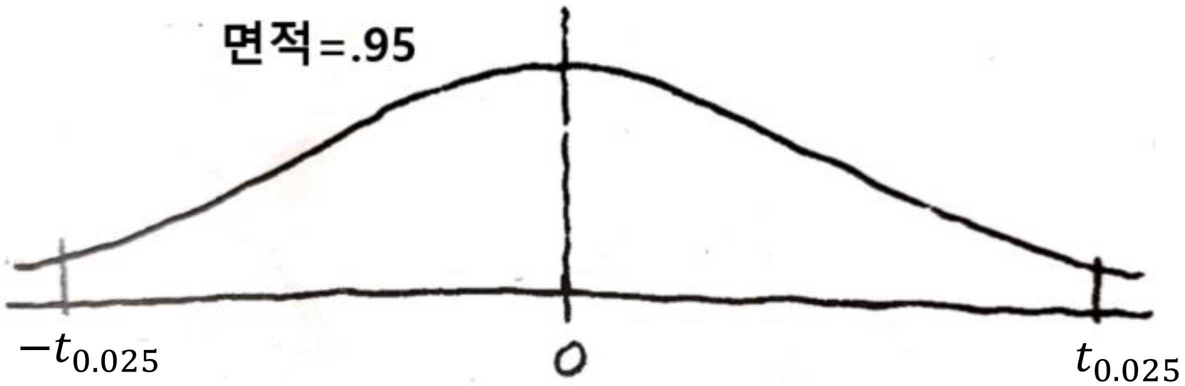
- t분포는 정규분포보다 평평하고, 그 정도는 표본의 크기에 좌우된다.



  - 고셋은 n이 표본 크기일 경우 n-1을 표본의 자유도 라고 부른다.
  - n개의 자료 $x_1, x_2,\ x_3,\ldots,\ x_n$이 주어질 때 $\bar{x}$를 계산하면 n-1개의 정보를 남겨놓고 1개의 '자유도'를 사용한다.

연속확률분포 - t-분포

- 고셋은 표본 크기, 즉 자유도에 따른 t분포표를 계산하였고, 자유도가 많을수록 t분포는 표준정규분포에 가까워진다.



  - 표본 크기 n을 알면, 자유도가 n-1인 t분포를 택하면 된다.
  - z분포(즉, 표준정규분포)에서처럼, t분포에서도 임계값 $t_{0.025}$를 찾아서 95% 신뢰수준을 얻을 수 있고, 임계값보다 큰 부분의 곡선 아래 면적은 0.025이다.

연속확률분포 - t-분포

- (1-α)·100% 신뢰구간에 대해서 $Pr\left(t \geq t_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$가 되는 $t_{\frac{\alpha}{2}}$를 찾는다.
- 아래 표는 t분포의 일부 임계점이다.

	1-α	.80	.90	.95	.99
	α	.20	.10	.05	.01
	$\alpha/2$	.10	.05	.025	.005
자유도	1	3.09	6.31	12.71	63.66
	10	1.37	1.81	2.23	4.14
	30	1.31	1.70	2.04	2.75
	100	1.29	1.66	1.98	2.63
	∞	1.28	1.65	1.96	2.58

  - 각 칸은 동일한 신뢰수준에서 자유도에 따른 값을 나타내고, 자유도가 높을수록 임계값은 정규분포의 임계값 $z_{\frac{\alpha}{2}}$에 가깝다.

연속확률분포 - t-분포

- 다음은 R의 t분포의 분위수 함수 qt()를 이용하여 자유도 4인 경우의 $t_{0.025}$ 임계값 구하는 예제이다.

```
# t분포 분위수 함수 : qt(p, df, lower.tail = TRUE/FALSE)
# Pr(t > 0.025)이고, 자유도가 4
# lower.tail  logical; if TRUE (default),
#                         probabilities are P[X ≤ x] otherwise, P[X > x].
#
qt(p=0.025, df=4, lower.tail = FALSE)
# [1] 2.776445
```

[t분포의 일부 임계점]

	$1-\alpha$	.80	.90	.95	.99
	α	.20	.10	.05	.01
	$\alpha/2$	.10	.05	.025	.005
자유도	1	3.09	6.31	12.71	63.66
	2	1.89	2.92	4.30	9.92
	3	1.64	2.35	3.18	5.84
	4	1.53	2.13	$\boxed{2.78}$	4.60
	5	1.48	2.01	2.57	4.03