표본 분산을 n-1로 나누는 이유

베르누이분포의 확률 구하는 법

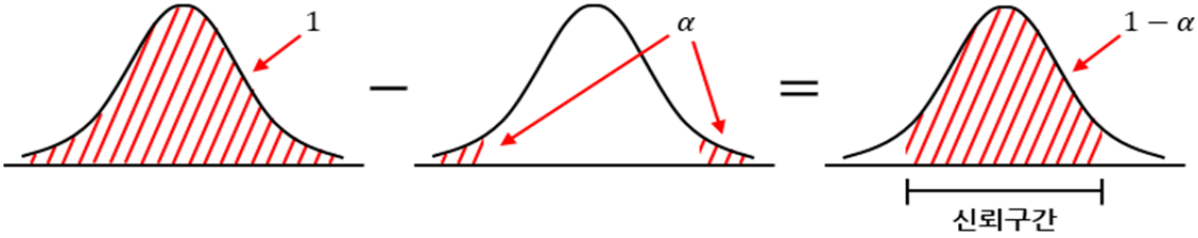
신뢰구간이란?

점추정은 그 특성상 신뢰도가 떨어지기에, 신뢰도를 높이기 위해서 일정구간을 활용한 구간추정을 해야 한다. 그런데 구간추정을 할 때 "과연 구간의 길이를 어느 정도로 할 것이냐?"라는 문제가 있다. 예를 들어 전 세계 성인 남자의 평균 키를 150cm ~ 190cm로 구간추정 했다고 해보자. 그럼 이 구간의 길이는 너무 길어서, 평균 키(모수)가 해당 구간 안에 들어가는 것은 너무나도 당연하다. 그래서 이 구간은 추정치로서 값어치가 떨어진다. 반대로 전 세계 성인 남자의 평균 키를 170cm~172cm로 구간추정 해보면, 구간의 길이가 너무 짧아서, 평균 키(모수)가 해당 구간 안에 들어갈 확률은 매우 낮아진다. 그래서 이 구간은 신뢰도가 떨어진다.

그래서 둘 다 신뢰하기에는 구간이 너무 막 잡혔는데, 이렇듯 구간의 길이는 너무 길어서 좋을 것이 없고, 반대로 너무 짧아서도 좋을 것이 없다. 그러므로 구간추정으로 구간을 만들 때는, 너무 길지도 너무 짧지도 않은 적당한 구간을 만들 필요가 있는데, **나름의 기준을 통해서 신뢰할 수 있는 구간의 길이를 만든 것이 신뢰구간이다.**

그런데 신뢰구간이라고 해서 완벽한 것은 아니다. 아무리 신뢰할 수 있는 구간이라도 **모수가 신뢰구간 안에 포함되지 않을 확률은 항상 존재하는데, 이 확률을 보통 α(알파)라고 부른다**. 그런데 신뢰구간은 양쪽으로(왼쪽과 오른쪽) 다루어야 하므로, α가 둘로 나뉘어서 α/2가 된다. 그래서 신뢰구간을 추정할 때는 α/2가 많이 나오는데, 정규분포 그래프에 대입해서 이해하면 편하다.(신뢰구간은 크게 "모평균"과 "모비율"과 "모분산"의 신뢰구간을 많이 구하는데, 각 신뢰구간을 추정할 때는 각각에 맞는 확률분포를 사용해서 구한다)

그런데 확률의 총합은 1이므로, 그래프의 총면적도 1이다. 그래서 모수가 신뢰구간 안에 포함되지 않을 확률이 α이므로, **모수가 신뢰구간 안에 포함될 확률은 1-α가 된다**. 그리고 1-α를 "신뢰수준"이라고 부르는데, 보통 90%와 95%와 99%의 확률을 많이 사용한다. 그리고 신뢰수준을 기반으로 설정된 구간이 신뢰구간인데, 이렇게 구간을 설정할 때는 임의대로 막 잡는 것이 아니라, **확률분포의 1-α를 기준으로** 구간을 설정한다.



연속확률분포의 공식

이전까지 다루었던 이산확률분포와는 다르게, 연속확률분포는 셀 수가 없기 때문에, 확률을 구할 때 그래프를 사용한다. 그런데 그래프를 어떻게 사용하는지에 따라서, 연속확률분포의 공식은 크

게 2가지로 나뉘는데, 먼저 균등분포와 지수분포처럼 그래프의 면적을 구하는 공식이 있고, 정규분포와 t분포 그리고 카이제곱분포와 F분포처럼 그래프의 x축 좌표를 구하는 공식이 있다.(참고로 이전까지 다루었던 이산확률분포의 공식은 모두 확률을 구하는 데 사용하지만, 연속확률분포의 공식은 그렇지가 않다. 그래서 통계를 처음 접할 때 많이 헷갈릴 수 있으므로, 연속확률분포의 공식에 대해서 한 번 알아보고 넘어가는 것이 좋다)

1. 그래프의 면적을 구함
2. 그래프의 x축 좌표를 구함

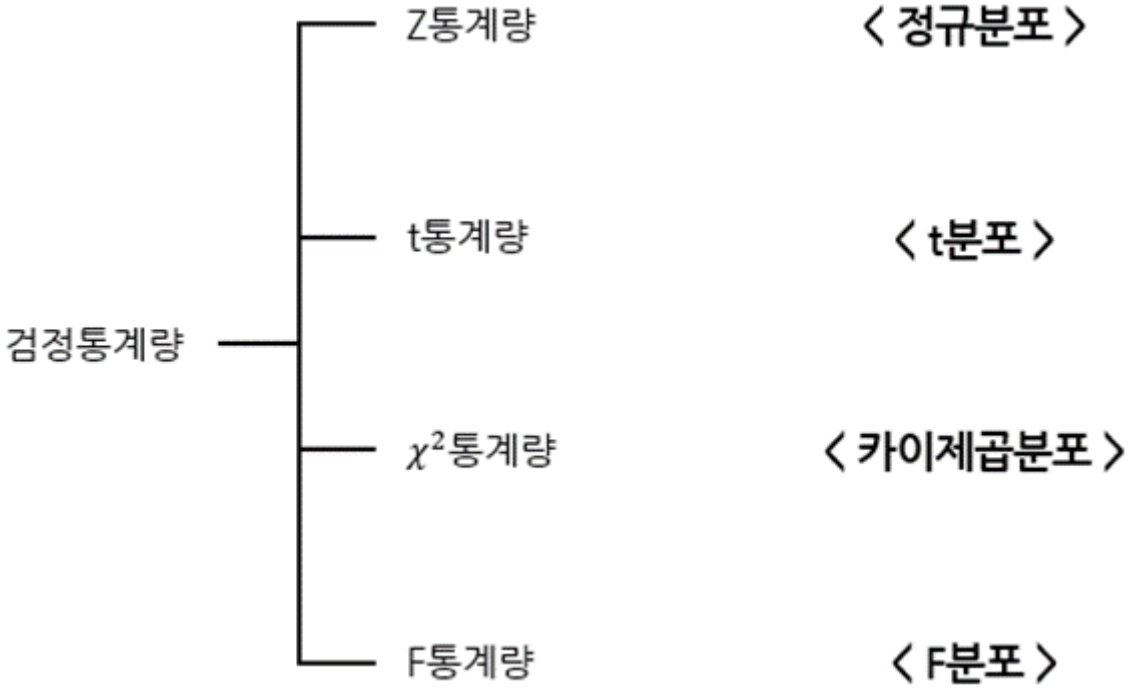
먼저 균등분포와 지수분포는 공식을 누적분포함수인 F(x)로 표기하는데, F(x)는 그래프의 면적을 구하는 공식이다. 연속확률분포에서는 그래프의 면적이 곧 확률값이라서, 면적의 넓이로 확률을 구하는데, F(x)를 사용하면 바로 면적의 넓이를 구할 수 있다.

다음으로 정규분포와 t분포 그리고 카이제곱분포와 F분포는 공식을 Z와 t와 χ2과 F로 표기하는데, Z와 t와 χ2과 F는 그래프의 x축 좌표를 구하는 공식이다. 단지 공식으로 그래프의 x축 좌표인 Z값, t값, χ2값, F값을 구할 뿐, 확률을 구하는 공식은 아니다.

검정통계량이란?

귀무가설과 대립가설은 "모집단의 모수가 이럴 것이다"라는 것을 나타내기 때문에, 모수인 μ와 p와 σ2을 사용해서 설정하였다. 그런데 어떤 가설이 더 타당한지를 파악하기 위해서, 계산을 할 때는 모수를 사용할 수가 없다. 왜냐하면 시간과 비용이 너무 많이 들기 때문에, 현실적으로 모집단 전체를 조사하기는 힘들다. 그래서 통계에서는 표본을 뽑아서 표본통계량 으로 계산하는데, 이 표본통계량을 가설검정에서는 검정통계량이라고 부른다.

단지 "가설검정"에서 사용하는 통계량이기에 "검정통계량"이라고 부를 뿐, 별다른 큰 의미는 없다. 그냥 표본통계량이라고 생각해도 된다. 그리고 가설검정을 할 때는 가장 처음으로 귀무가설과 대립가설을 설정한다고 했었는데, 그다음에 보통 하는 것이 검정통계량 계산이다.(사람에 따라서 "기각역"을 먼저 구하기도 한다) 그리고 가설검정을 할 때는 그냥 막 하는 것이 아니라 확률분포를 활용하는데, 신뢰구간을 구할 때와 마찬가지로 정규분포와 t분포와 χ2분포와 F분포를 활용한다. 그래서 검정통계량도 확률분포에 따라서 "Z통계량" "t통계량" "χ2통계량" "F통계량"으로 세분화할 수 있다



그런데 한 가지 주의할 점은 검정통계량으로 확률을 구하는 것이 아니라, **확률분포 그래프의 x축 좌표를 구한다는 점이다.** **그래서 검정통계량의 공식을 한 번 살펴보면, 그래프의 x축 좌표를 구하는 공식이라는 것을 알 수 있다.**

그래프의 x축 좌표

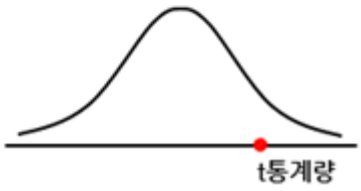
$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ 〈 정규분포 〉

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 〈 t분포 〉

$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ 〈 카이제곱분포 〉

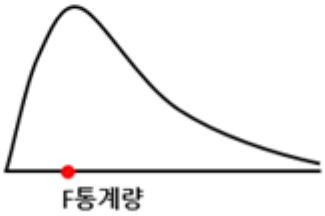
$F = \frac{s_1^2}{s_2^2}$ 〈 F분포 〉

〈 정규분포 〉



〈 t분포 〉

〈 카이제곱분포 〉

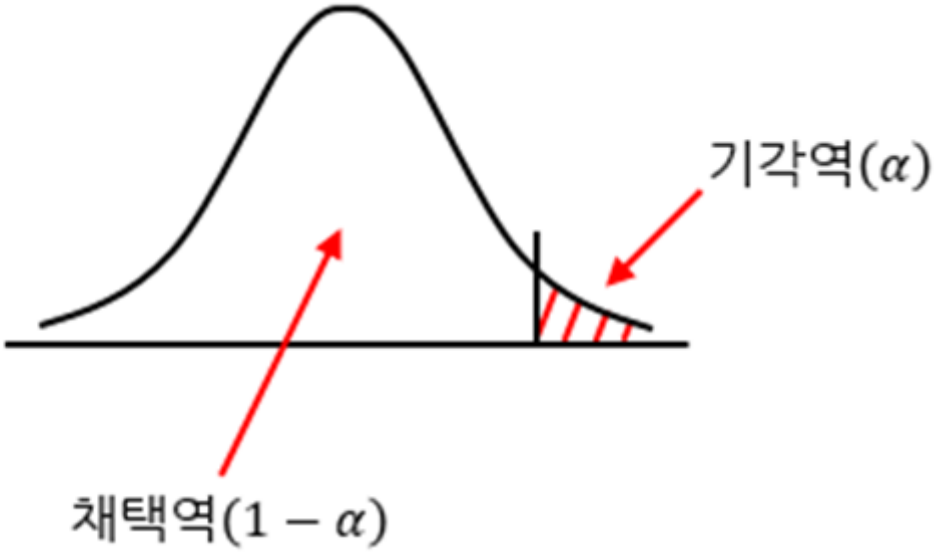


〈 F분포 〉

기각역이란?

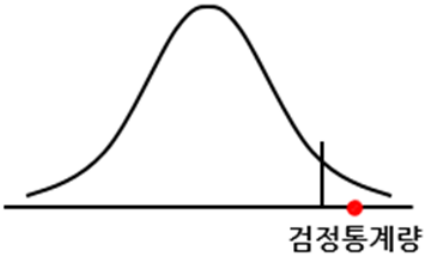
가설검정은 귀무가설과 대립가설 중에서 하나의 가설을 양자택일한다고 했었다. 그래서 귀무가설을 옳다고 채택하면, 자동으로 대립가설은 탈락(기각)하게 되고, 반대로 대립가설을 옳다고 채택하면, 자동으로 귀무가설은 탈락하게 된다. 그리고 가설검정은 나름의 기준을 통해서 이렇게 채택과 탈락 여부를 결정하는데, 한 가지 명심할 것은 100%의 정답이 아니라 항상 어느 정도의 오차는 존재한다는 점이다. 그래서 가설검정도 틀릴 확률은 항상 존재한다.

예를 들어 귀무가설과 대립가설 중에서, 귀무가설이 더 옳다면 귀무가설을 채택해야 한다. 하지만 오차에 의해서 "귀무가설이 더 옳은데도 불구하고, 귀무가설을 탈락시키는 확률"이 생기는데, **이 확률을 보통 α(알파)라고 부른다.**(확률 α를 "유의수준"이라고 부르는데, 보통 1%, 5%, 10%를 많이 사용한다) 그렇다면 확률의 총합은 1이기 때문에, "귀무가설이 더 옳기에, 귀무가설을 채택할 확률"은 1-α가 된다. 그래서 1-α는 귀무가설을 채택시키므로, 1-α의 영역을 "채택역"이라고 부르고, 반대로 α는 귀무가설을 기각(탈락)시키므로, α의 영역을 "기각역"이라고 부른다.



그럼 이 채택역과 기각역으로 귀무가설의 채택과 기각 여부를 판단하는데, 이전 글에서 다루었던 검정통계량을 활용한다. 그래서 검정통계량이 "채택역"안에 위치하면 귀무가설이 채택되고, 반대로 검정통계량이 "기각역"안에 위치하면 귀무가설이 기각(탈락)된다. 여기서 채택과 탈락의 여부는 귀무가설만 기준으로 해서 생각하는 것이 편할 것이다. 어차피 가설검정은 양자택일이므로, 대립가설은 정반대로 판단하면 된다.

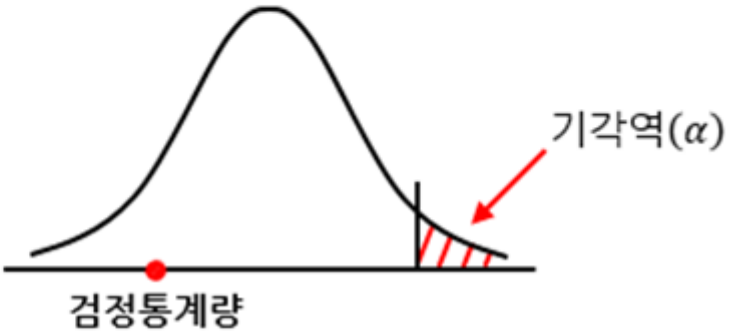
〈 귀무가설 채택 〉



〈 귀무가설 탈락 〉

기각역 설정하는 법(단측, 양측 검정)

이전 글에서 검정통계량이 "채택역"안에 위치하면 귀무가설을 채택하고, 검정통계량이 "기각역"안에 위치하면 귀무가설을 기각(탈락)한다고 했었다. 이렇게 가설검정은 표본으로 뽑은 통계량이(검정통계량) 어디에 위치하느냐에 따라서 귀무가설의 채택과 기각 여부를 판단한다. 그런데 만약 표본으로 뽑은 검정통계량의 값이 작아서 그래프의 왼쪽에 위치하고 있을 때, 기각역이 오른쪽에 있다면 어떻게 될까? 아마도 상황이 이상할 것이다.



〈 상황이 이상하다 〉

검정통계량이 그래프의 왼쪽에 위치하고 있으면, 굳이 멀리 떨어져 있는 오른쪽 기각역 말고, 그냥 가까이에 있는 왼쪽 기각역과 비교를 하면 된다. 그리고 오른쪽 기각역은 모수가 얼마나 큰지를 판단하기 때문에, 값이 작은 검정통계량으로 모수가 얼마나 작은지는 판단을 못 한다. 그래서 검정통계량의 값이 작아서 그래프의 왼쪽에 위치하고 있으면 기각역을 왼쪽에 설정한다. 반대로 검정통계량의 값이 커서 그래프의 오른쪽에 위치하고 있으면 기각역을 오른쪽에 설정한다. 이렇게 기각역을 왼쪽에 설정한 것을 "좌측검정"이라고 하고, 기각역을 오른쪽에 설정한 것을 "우측검정"이라고 하는데, 한쪽 방향만 검정하기 때문에 둘 다 "단측검정"이라고 한다.

〈 좌측검정 〉 〈 우측검정 〉

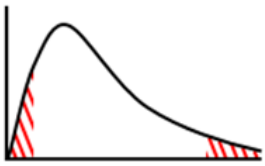
정리 해 보면 다음과 같다

1. 양측검정

$H_0: \mu = 7$ $H_0: p = 13\%$ $H_0: \sigma^2 = 4$

$H_1: \mu \neq 7$ $H_1: p \neq 13\%$ $H_1: \sigma^2 \neq 4$

같지 않다.

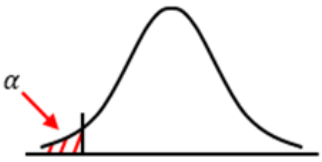


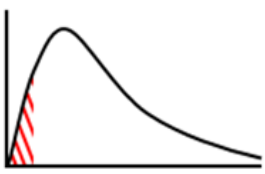
2. 좌측검정

$H_0: \mu \geq 7$ $H_0: p \geq 13\%$ $H_0: \sigma^2 \geq 4$

$H_1: \mu < 7$ $H_1: p < 13\%$ $H_1: \sigma^2 < 4$

작다.



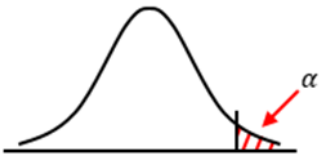


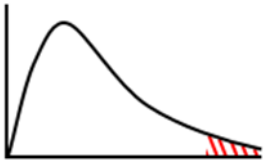
3. 우측검정

$H_0: \mu \leq 7$ $H_0: p \leq 13\%$ $H_0: \sigma^2 \leq 4$

$H_1: \mu > 7$ $H_1: p > 13\%$ $H_1: \sigma^2 > 4$

크다.





참고로 위의 3가지 상황 모두 검정통계량으로 귀무가설의 채택과 기각 여부를 판단한다. 그래서 검정통계량이 하얀색 "채택역"안에 위치하면 귀무가설이 채택되고, 반대로 검정통계량이 빗금 친 "기각역"안에 위치하면 귀무가설이 기각(탈락)된다.

두 모집단에 대한 추론

두 모집단의 귀무가설과 대립가설 설정하는 법

두 모집단의 가설검정은 두 모수가 서로 어떠한 관계에 있는지를 비교하는 것이다. 두 모수의 관계는 "같다" "크다" "작다" 이렇게 3가지로 나타낸다. 그래서 귀무가설과 대립가설을 설정할 때는 이러한 특징을 나타내면 되는데, 예를 들어 "평균이 서로 같다" "비율1이 더 크다" "분산1이 더 작다"처럼 두 집단의 모수를 서로 비교해서 가설을 설정하면 된다.

평균이 서로 같다. 비율1이 더 크다. 분산1이 더 작다.

$H_0: \mu_1 = \mu_2$ $H_0: p_1 \geq p_2$ $H_0: \sigma_1^2 \leq \sigma_2^2$

$H_1: \mu_1 \neq \mu_2$ $H_1: p_1 < p_2$ $H_1: \sigma_1^2 > \sigma_2^2$

그리고 두 모수의 관계를 파악할 때는 "뺄셈"과 "나눗셈"을 사용하기에, 귀무가설과 대립가설을 뺄셈과 나눗셈을 활용해서 설정하기도 한다.

$$H_0: \mu_1 - \mu_2 = 0 \qquad H_0: p_1 - p_2 \geq 0 \qquad H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq 1$$

$$H_1: \mu_1 - \mu_2 \neq 0 \qquad H_1: p_1 - p_2 < 0 \qquad H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$$

〈 빨셈과 나눗셈으로 가설을 설정하기도 한다 〉

정규 분포에서의 값 구하기

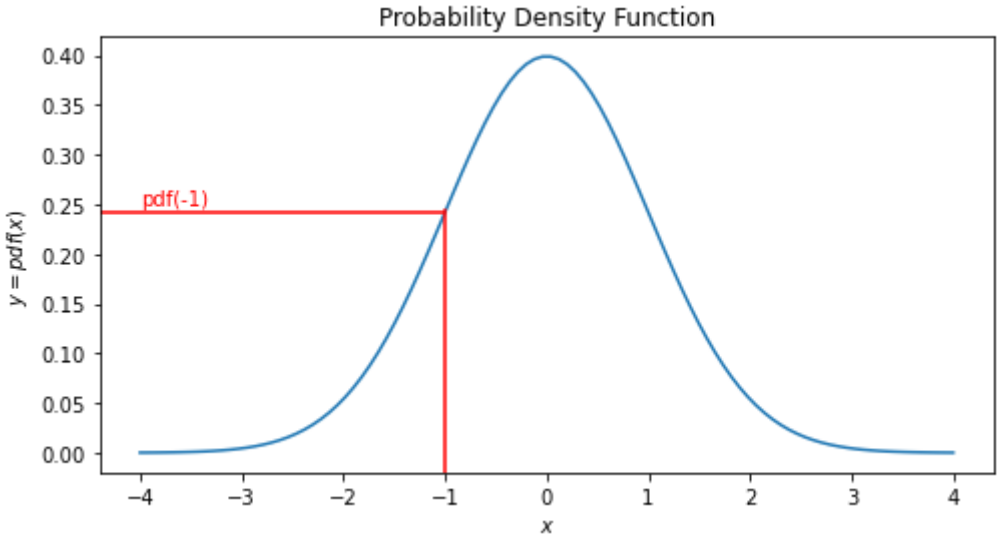
scipy.stats.norm에 있는 함수를 이용해서, 정규 분포 그래프에서의 값들을 구할 수 있다.

- **pdf (probability density function, 확률밀도함수)**
- **cdf (cumulative distribution function, 누적분포함수)**
- **ppf (percent point function, 누적분포함수의 역함수)**

$pdf(probability density function$, 확 률 밀 도 함 수$)$

pdf는, 평균 μ와 표준편차 σ 인 정규분포에서, 어떤 값 x에 대한 확률 값을 구하는 함수이다.

In [41]:
```
rv = stats.norm(loc=0, scale=1)
xx = np.linspace(-4, 4, 100)
plt.figure(figsize=(8,4))

pdf = rv.pdf(xx)
plt.plot(xx, pdf)

plt.title("Probability Density Function")
plt.xlabel("$x$"); plt.ylabel("$y=pdf(x)$")

plt.axvline(x=-1, ymin=0, ymax=0.6, color='red')
plt.axhline(y=rv.pdf(-1), xmin=0, xmax=0.385, color='red')
plt.text(-4,0.25,'pdf(-1)',color='red')

plt.show()
```



cdf(cumulative distribution function, 누적 분포 함수)

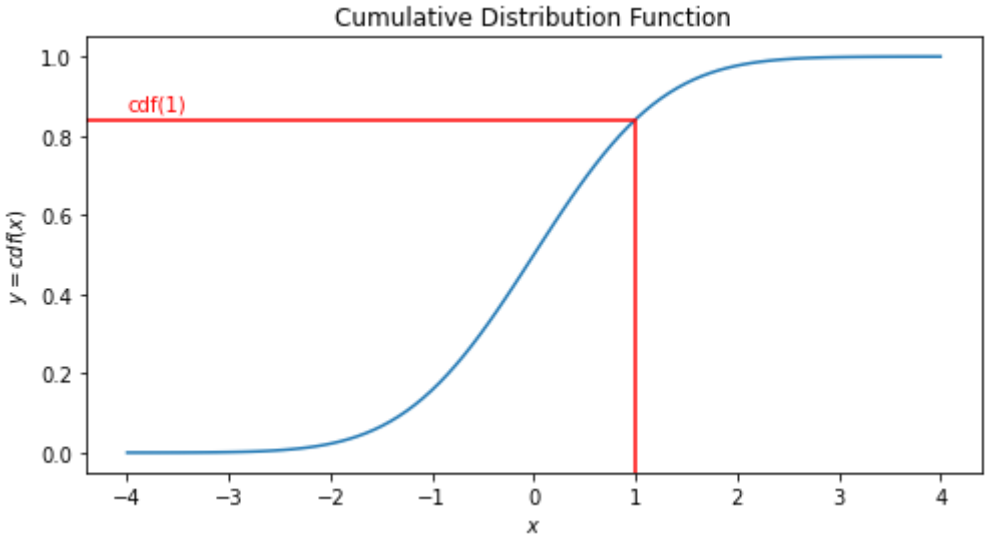
cdf는, 정규 분포 곡선에서 x의 값에 따른 누적 확률을 구하는 값이다.

pdf가 x의 값에 따른 정규분포 확률값을 나타내기에, pdf 곡선은 우리가 알고 있는 종모양의 곡선이다. 이 곡선의 면적을 구하는 것은, x가 나올 수 있는 모든 경우에 대한 확률이기에, 전체 면적을 더하면 1이 된다.

cdf는 이러한 x가 증가함에 따른 누적 확률이기에, 누적 확률 값은 0에서 시작해서 1까지 증가하게 된다.

In [42]:

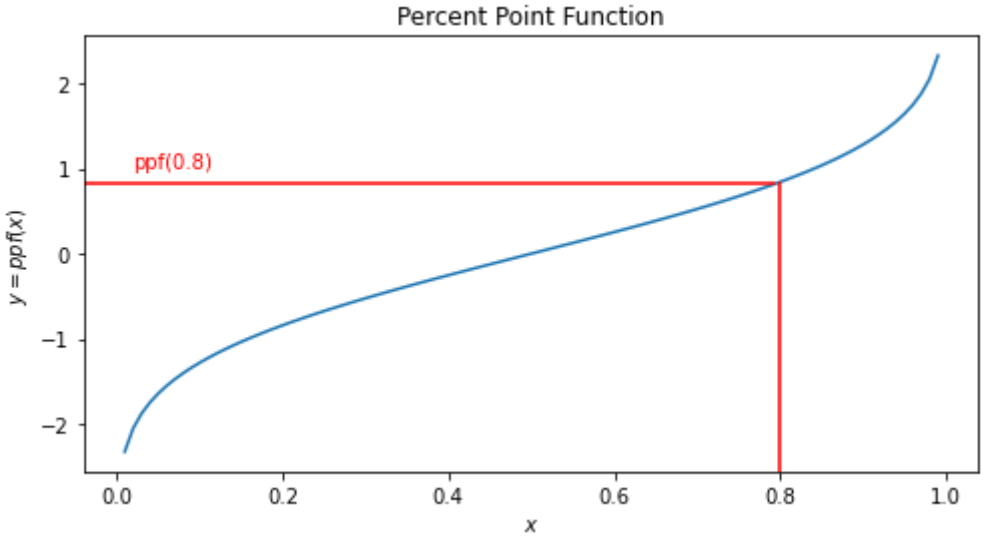
```
rv = stats.norm(loc=0, scale=1)
xx = np.linspace(-4, 4, 100)
plt.figure(figsize=(8,4))

cdf = rv.cdf(xx)
plt.plot(xx, cdf)

plt.title("Cumulative Distribution Function")
plt.xlabel("$x$"); plt.ylabel("$y=cdf(x)$")

plt.axvline(x=1, ymin=0, ymax=0.81, color='red')
plt.axhline(y=rv.cdf(1), xmin=0, xmax=0.61, color='red')
plt.text(-4, rv.cdf(1.1),'cdf(1)',color='red')

plt.show()
```



$$ppf\left(percentp\oint function,\ 누\ 적\ 분\ 포\ 함\ 수\ 의\ 역\ 함\ 수\right)$$

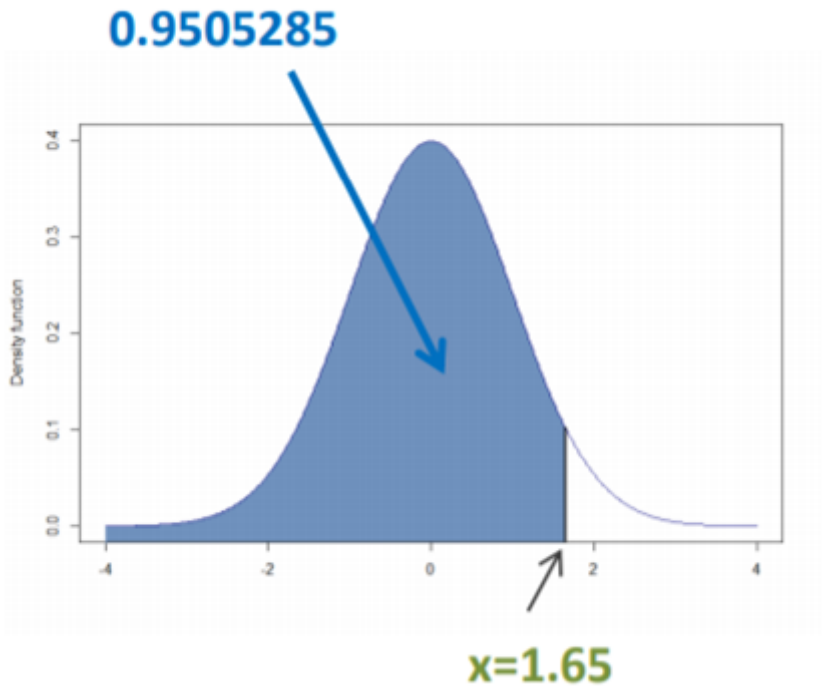
ppf는 누적분포함수에서 해당 확률일 때의 x를 구하는 함수이다.

In [43]:
```
rv = stats.norm(loc=0, scale=1)
xx = np.linspace(0, 1, 100)
plt.figure(figsize=(8,4))

ppf = rv.ppf(xx)
plt.plot(xx, ppf)

plt.title("Percent Point Function")
plt.xlabel("$x$"); plt.ylabel("$y=ppf(x)$")

plt.axvline(x=0.8, ymin=0, ymax=0.66, color='red')
plt.axhline(y=0.82, xmin=0, xmax=0.77, color='red')
plt.text(0.02, 1.0,'ppf(0.8)',color='red')

plt.show()
```



In [1]:
```
from scipy.stats import norm
print(norm.ppf(0.95))
print(norm.cdf(1.6448536269514722))
```

```
1.6448536269514722
0.95
```



두 모평균 차이의 검정(σ를 아는 경우)

참고

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

〈모분산〉

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

〈표본분산〉

가설 표현

두 모평균의 가설검정은 크게 "σ를 아는 경우"와 "σ를 모르는 경우"로 나뉘는데, 이번에는 σ를 아는 경우에 대해서 알아보자. 일단 두 모평균의 가설검정은 2개의 모평균이 서로 어떠한 관계에 있는지를 비교하는 것인데, 두 모평균의 차이를 비교할 때는 보통 "뺄셈"을 활용한다. 그래서 뺄셈을 활용해서 귀무가설과 대립가설을 설정하기도 하기에, 가설을 표현하는 방법은 총 2가지이다.

$\mu_1 = \mu_2$

$\mu_1 - \mu_2 = \mu_2 - \mu_2$ ← 양변에 $-\mu_2$를 더한다.

$\mu_1 - \mu_2 = 0$ ← 뺄셈으로 설정

〈방법1〉

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

or

〈방법2〉

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

예를 들어서

한국인과 일본인의 평균키(모평균)는 얼마나 차이 나는가?

대기업과 중소기업의 평균연봉(모평균)은 얼마나 차이 나는가?

동양인과 백인의 평균 IQ(모평균)는 얼마나 차이 나는가?

남성과 여성의 평균수명(모평균)은 얼마나 차이 나는가?

도시와 농촌의 평균소득(모평균)은 얼마나 차이 나는가?

두 모평균 차이의 신뢰구간 구하는 법(σ를 아는 경우)

정규 분포 공식(Z)

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

1개 추가

〈 집단 1개 〉 〈 집단 2개 〉

$$-Z\alpha_{/2} \le \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z\alpha_{/2}$$

$$-Z\alpha_{/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) \le Z\alpha_{/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$-(\bar{x}_1 - \bar{x}_2) - Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le -(\mu_1 - \mu_2) \le -(\bar{x}_1 - \bar{x}_2) + Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) + Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \ge \mu_1 - \mu_2 \ge (\bar{x}_1 - \bar{x}_2) - Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) - Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm Z\alpha_{/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leftarrow \text{신뢰구간 공식}$$

두 모평균 차이의 신뢰구간 예제

1. **대기업과 중소기업의 신입사원 평균연봉이 얼마나 차이 나는지를 비교하기 위하여, 대기업 15개와 중소기업 20개를 조사하였더니, 평균연봉은 각각 4100만 원과 2800만 원이 나왔다. 그리고 각 회사의 과거 자료를 분석해보니 분산은 각각 400만 원과 500만 원이라고 한다. 이 때 대기업과 중소기업의 신입사원 평균연봉의 차이에 대한 90%의 신뢰구간을 구하시오.**

In [2]:
```
from scipy.stats import norm
import numpy as np
alpha = 0.1
alpha  = alpha / 2
z = norm.ppf(1-alpha)
```

```
x1 = 4100
x2 = 2800
n1 = 15
n2 = 20
s1 = 400
s2 = 500

up = (x1-x2) + (z * np.sqrt(s1/ n1 + s2/n2))
down = (x1-x2) - (z * np.sqrt(s1/ n1 + s2/n2))

print('신뢰 구간 %6.4f <= u1 - u2 <= %6.4f'%(down,up))
```

```
신뢰 구간 1288.1769 <= u1 - u2 <= 1311.8231
```

1. 스마트폰을 생산하는 A와 B 두 회사가 있는데, 두 회사에서 생산하는 스마트폰의 평균수명이 얼마나 차이 나는지를 비교하려고 한다. 그래서 각각 25개와 30개의 표본을 뽑았더니, 평균수명은 각각 750일과 700일이 나왔고, 각 회사의 과거 데이터를 분석해보니 분산은 각각 40일과 45일이라고 한다. 이때 두 스마트폰의 평균수명의 차이에 대한 95%의 신뢰구간을 구하시오.

In [18]:

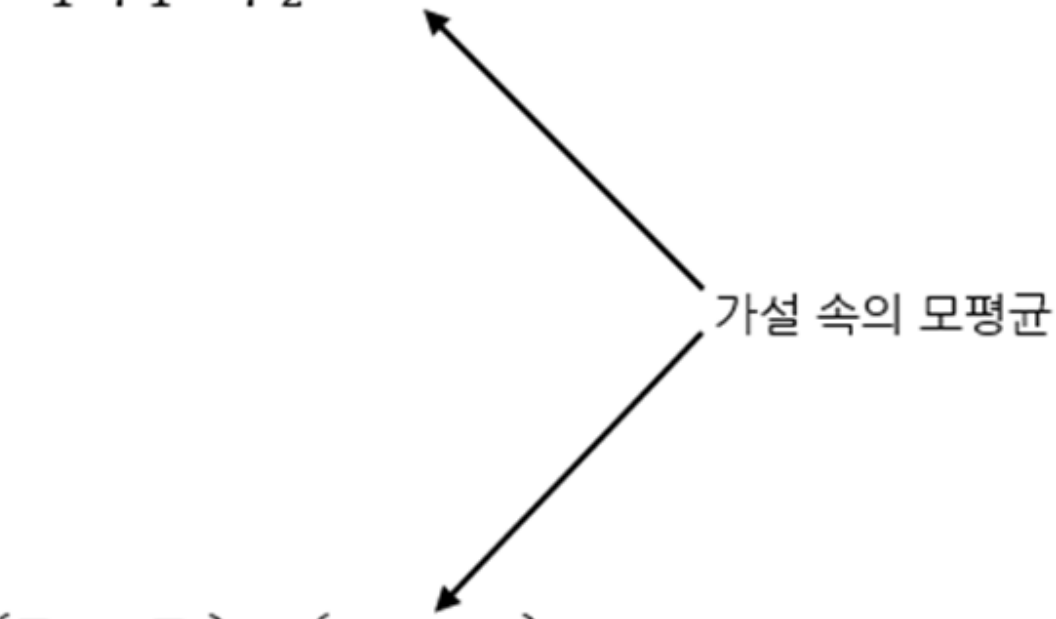
```
alpha = 0.05
alpha  = alpha / 2
z = norm.ppf(1-alpha)

x1 = 750
x2 = 700
n1 = 25
n2 = 30
s1 = 40
s2 = 45

up = (x1-x2) + (z * np.sqrt(s1/ n1 + s2/n2))
down = (x1-x2) - (z * np.sqrt(s1/ n1 + s2/n2))

print('신뢰 구간 %6.4f <= u1 - u2 <= %6.4f'%(down,up))
```

```
신뢰 구간 46.5491 <= u1 - u2 <= 53.4509
```

두 모평균 차이의 검정

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$



$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

표준정규분포표에서 Z값

α	Z값
0.1	1.28
0.05	1.645
0.025	1.96
0.01	2.33
0.005	2.575

두 모평균 차이의 검정 예제

1. **도시와 농촌의 월평균소득은 서로 비슷하다고 알려져 있는데, 일부 농민운동단체에서는 도시의 월평균소득이 더 많다고 주장한다. 그래서 실제로 그러한지를 알아보기 위해 각각 100개와 90개의 가구를 조사하였더니, 도시가구의 월평균소득은 240만 원이 나왔고 농촌가구의 월평균소득은 230만 원이 나왔다. 그럼 도시가구의 모표준편차는 60만 원이고 농촌가구의 모표준편차는 70만 원이라고 가정했을 때, 도시의 월평균소득이 더 크다고 할 수 있는지 유의수준 10%에서 검정하시오.**

   귀무가설 : u1 <= u2 도시 월평균 소득은 농촌 월평균 소득과 같거나 작다

   대립가설 : u1 > u2 도시 월평균 소득은 농촌 월평균 소득보다 많다

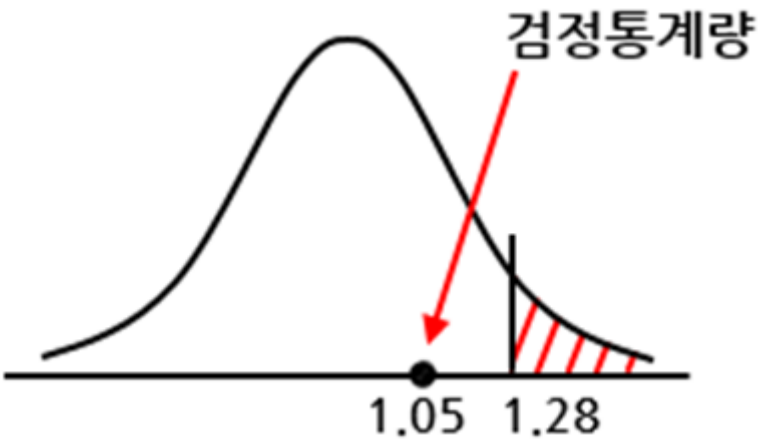
   ==> 대립가설로 도시의 월평균소득이 더 많다는 주장이 나왔으므로, 대립가설을 μ1이 더 "크다"로 설정한다

In [10…

```
alpha = 0.1
z = norm.ppf(1-alpha)

x1 = 240
x2 = 230
n1 = 100
n2 = 90
s1 = 60**2
s2 = 70**2

statistic = (x1-x2) / np.sqrt(s1 / n1 + s2 / n2)
z, statistic
```

Out[10… (1.2815515655446004, 1.0514994558491724)

그다음 유의수준 α=0.1이므로 기각역은 1.28이다. 그럼 검정통계량이 "채택역"안에 위치하므로 귀무가설이 채택된다. **그래서 도시의 월평균소득이 더 많다고 할 수 없다.**



1. 체크카드를 사용하면 신용카드를 사용할 때보다, 카드 사용액이 더 낮아진다는 의견이 나오고 있다. 그래서 실제로 어떠한지를 알아보기 위해 각 카드사용자 55명과 60명의 카드 사용액을 조사하였더니, 체크카드의 평균 사용액은 50만 원이었고 신용카드의 평균 사용액은 100만 원이었다. 그럼 체크카드의 모표준편차는 25만 원이고 신용카드의 모표준편차는 30만 원이라고 가정했을 때, 체크카드를 사용하면 카드 사용액이 더 낮아진다고 할 수 있는지 유의수준 1%에서 검정하시오.

   귀무가설 : u1 - u2 >= 0 체크카드 사용금액이 신용카드 사용금액보다 크거나 같다

   대립가설 : u1 - u2 < 0 체크카드 사용금액이 신용카드 사용금액보다 작다

In [3]:
```
from scipy.stats import norm
import numpy as np
alpha = 0.01
z = norm.ppf(1-alpha)

x1 = 50
x2 = 100
n1 = 55
n2 = 60
s1 = 25**2
s2 = 30**2

statistic = (x1-x2) / np.sqrt(s1 / n1 + s2 / n2)
-z, statistic
```

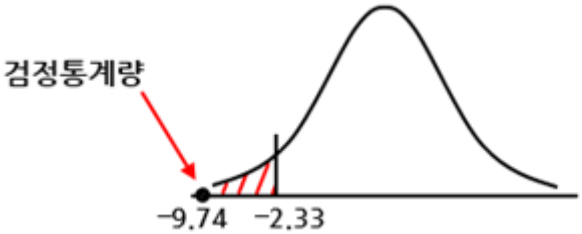
Out[3]: (-2.3263478740408408, -9.737945687202027)

In [31]:
```
alpha = 0.01
z = norm.ppf(1-alpha)

x1 = 50
x2 = 100
n1 = 55
n2 = 60
s1 = 25**2
s2 = 30**2

up = (x1-x2) + (z * np.sqrt(s1/ n1 + s2/n2))
down = (x1-x2) - (z * np.sqrt(s1/ n1 + s2/n2))
```

```
print('신뢰 구간 %6.4f <= u1 - u2 <= %6.4f'%(down,up))
```

```
신뢰 구간 -61.9448 <= u1 - u2 <= -38.0552
```

이미 판단은 끝났지만 그럼에도 기각역을 한 번 구해보면, 일단 유의수준 α=0.01인데 0.01에 해당하는 Z값은 2.33이다. 그런데 그래프의 왼쪽 좌표라서 -값을 붙이면 기각역은 -2.33이 된다. 그럼 검정통계량이 "기각역"안에 위치하므로 귀무가설이 기각(탈락)된다. 그래서 체크카드를 사용하면 카드 사용액이 더 낮아진다고 할 수 있다



1. 두 개의 건전지 A와 B가 있는데, A건전지의 평균수명이 100일 더 길다고 알려져 있다. 하지만 일부에서는 아닐 수도 있다는 의견이 나오고 있어서, 실상을 파악하기 위해 각 건전지 50개와 60개 뽑았더니, A건전지의 평균수명은 470일이 나왔고 B건전지의 평균수명은 360일이 나왔다. 그리고 과거에 수집한 데이터를 분석해보니, A건전지의 모표준편차는 20일이고 B건전지의 모표준편차는 30일이라고 한다. 이때 A건전지의 평균수명이 100일 더 길다고 할 수 있는지 유의수준 5%에서 검정하시오.

   귀무가설: u1 = u2 + 100

   대립가설: u1 != u2 + 100

In [4]:
```
import numpy as np
import scipy.stats as stats

alpha = 0.05

x1 = 470
x2 = 360
n1 = 50
n2 = 60
s1 = 20**2
s2 = 30**2

statistic = stats.norm.ppf(1 - alpha/2)
z = ((x1- x2) - 100)  / np.sqrt(s1/n1 + s2/n2)
statistic, z
```

Out[4]: (1.959963984540054, 2.0851441405707476)

$H_0: \mu_1 = \mu_2 + 100$

$H_1: \mu_1 \neq \mu_2 + 100$

$H_0: \mu_1 - \mu_2 = 100$

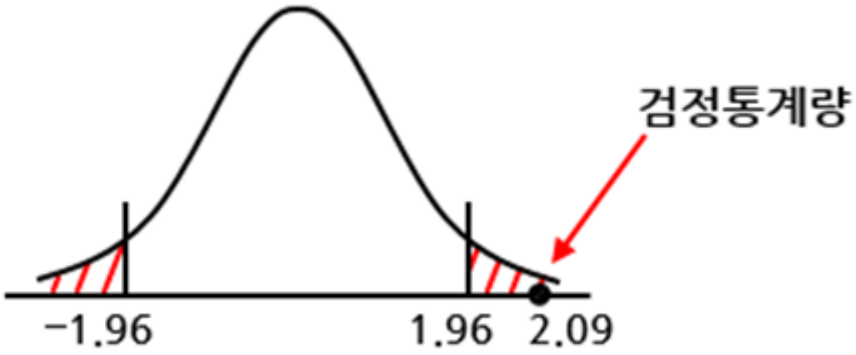
$H_1: \mu_1 - \mu_2 \neq 100$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(470 - 360) - 100}{\sqrt{\frac{20^2}{50} + \frac{30^2}{60}}}$$

$$= 2.09$$

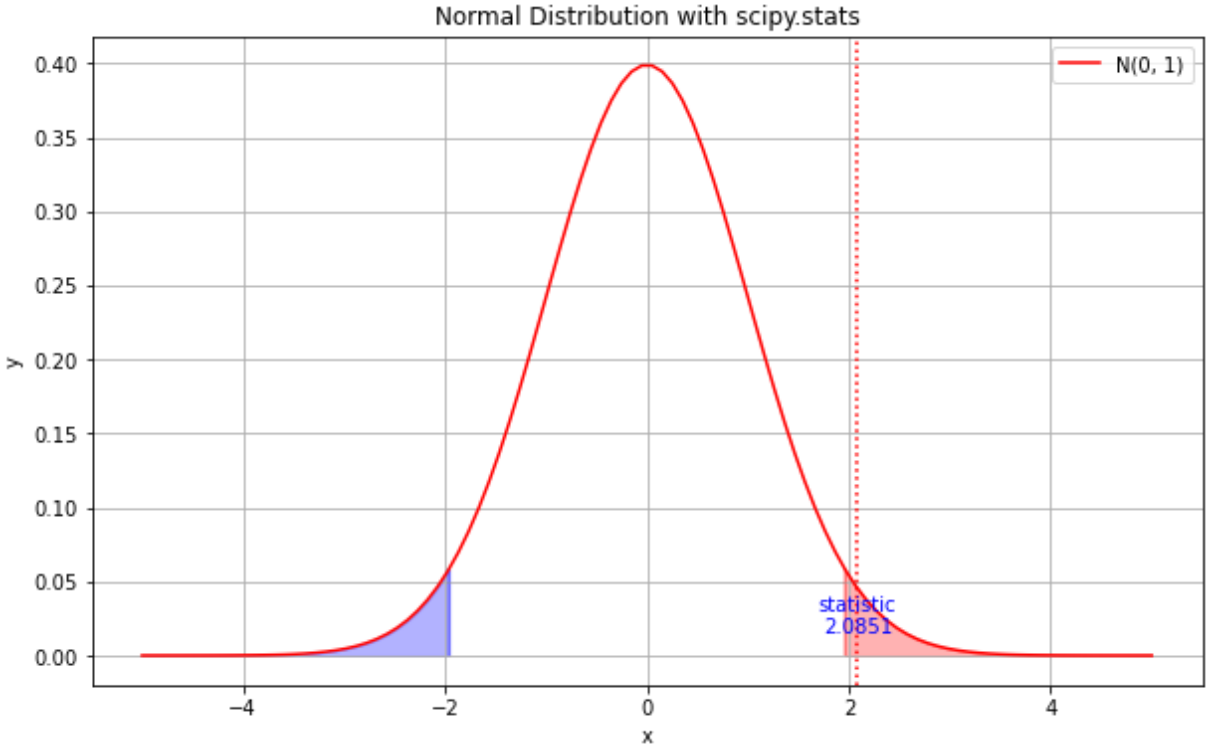
다음으로 유의수준 α=0.05인데, 양측검정이므로 α/2=0.025에 해당하는 Z값 1.96이다. 그런데 양쪽으로 설정해야 하므로 기각역은 ±1.96이기에, 검정통계량은 "기각역"안에 위치하게 된다. 그래서 귀무가설이 기각(탈락)이기에 A건전지의 평균수명이 100일 더 길다고 할 수 없다.



In [38]:

```
import matplotlib.pyplot as plt
x = np.linspace(-5,5,100) # 동일 간격으로 -5부터 5까지 100개 생성
rv = stats.norm(0, 1)
y1 =rv.pdf(x)

z_025 = rv.ppf(0.025)
z_975 = rv.ppf(0.975)

z_025_ = np.linspace(-5,z_025,1000)
z_975_ = np.linspace(z_975,5,1000)

plt.figure(figsize=(10, 6))          # 플롯 사이즈 지정

plt.vlines(z_025, 0, rv.pdf(z_025), colors="b", alpha=0.3)
plt.vlines(z_975, 0, rv.pdf(z_975), colors="r", alpha=0.3)
plt.plot(x, y1, color="red")          # 선을 빨강색으로 지정하여 plot 작성
plt.fill_between(z_025_, 0,rv.pdf(z_025_), color ="b", alpha=0.3)
plt.fill_between(z_975_, 0,rv.pdf(z_975_), color ="r", alpha=0.3)

plt.axvline(x=z, color='r', linestyle=':')
plt.text(z , .015, 'statisticWn' + str(round(z,4)),
             horizontalalignment='center', color='b')

plt.xlabel("x")                       # x축 레이블 지정
plt.ylabel("y")                       # y축 레이블 지정
plt.grid()                            # 플롯에 격자 보이기
```

```
plt.title("Normal Distribution with scipy.stats")      # 타이틀 표시
plt.legend(["N(0, 1)"])                                # 범례 표시
plt.show()
```



In [ ]:

두 모평균 차이의 검정(σ를 모르는 경우)

두 모평균 차이의 신뢰구간 구하는 법(σ를 모르는 경우)

두 모평균의 신뢰구간은 대부분 "σ1과 σ2를 모르는 경우"에 해당하는데, 신뢰구간을 구할 때는 t 분포를 사용한다. 그리고 신뢰구간 구하는 공식은 아래와 같다.

$$(\bar{x}_1 - \bar{x}_2) \pm t\alpha/_2(n_1+n_2-2) \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

그럼 공식을 한 번 살펴보면, 일단 공식에 있는 **sp는 "합동표준편차"라고 부르는데**, 두 집단의 표본을 모아서 한 번에 계산한 표준편차이다. 보통 추정을 할 때 표본의 수가 너무 적으면 추정값의 결과를 신뢰하기가 힘들다. 그래서 신뢰도를 높이기 위해서 두 집단의 표본을 모아서 한 번에 합동표준편차를 계산한다는 말이 있는데, 사실 별로 효과는 없다. 그래서 이전 글에서 다루었던 "σ1과 σ2를 아는 경우"처럼 따로따로 구했을 때와 합동표준편차로 구했을 때의 값을 서로 비교해보면, 값의 차이는 그렇게 크지가 않다. 그래서 쓸데없이 공식만 1개 늘어난 셈인데, 개인적인 생각으로는 "σ1과 σ2를 아는 경우"처럼 따로따로 계산해도 된다고 생각하지만, 일반적으로 통용되는 공식이 이것이기에, 그냥 합동표준편차를 사용하려고 한다. 합동표준편차 구하는 법은 아래와 같다.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우)

1. 건전지를 생산하는 두 회사가 있는데, 두 회사 건전지의 평균수명이 얼마나 차이 나는지를 비교하려고 한다. 그래서 각각 16개와 15개의 건전지를 표본으로 뽑아 실험하였더니, 표본평균은 각각 140일과 120일이 나왔고, 표본분산은 10일과 15일이 나왔다고 한다. 이때 두 건전지의 평균수명의 차이에 대한 95%의 신뢰구간을 구하시오.

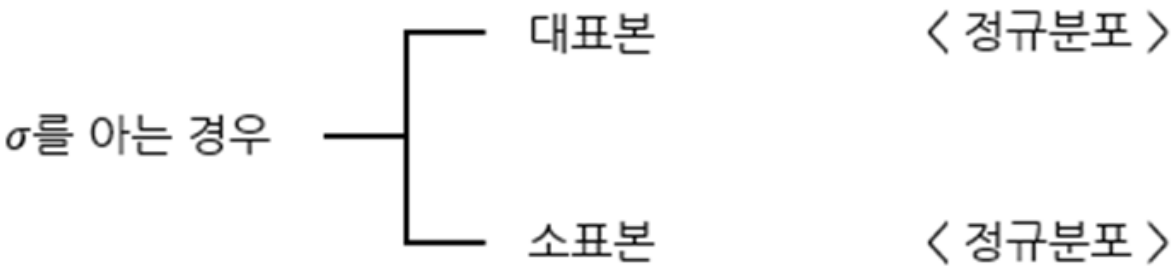
In [59]:
```
import numpy as np
import scipy.stats as stats

alpha = 0.05

x1 = 140
x2 = 120
n1 = 16
```

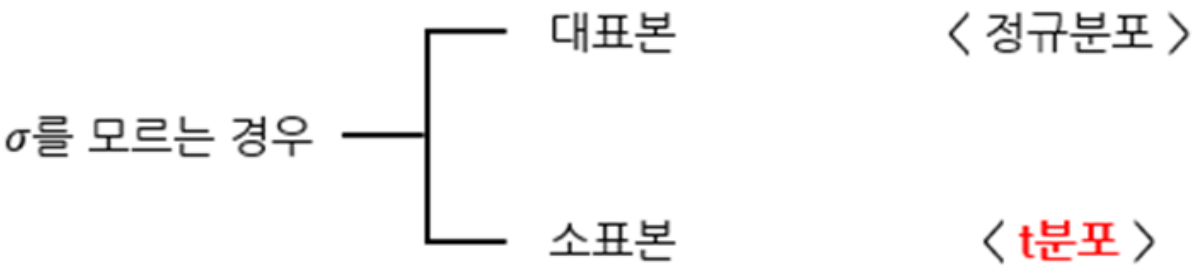
```
n2 = 15
s1 = 10
s2 = 15

statistic = stats.t.ppf(1 - alpha/2, n1+n2-2)
sp = np.sqrt(((n1 - 1)* s1 + (n2 - 1)* s2) / (n1+n2-2))

up = (x1 - x2) + statistic * sp * np.sqrt((1 / n1) + (1 / n2))
down = (x1 - x2) - statistic * sp * np.sqrt((1 / n1) + (1 / n2))

print('두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) %6.4f <= u1 - u2 <= %6
print('두 건전지의 평균수명의 차이는 17.4102일에서 22.5898일 사이라고 추정할 수
```

```
두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) 17.4102 <= u1 - u2 <= 22.5898
두 건전지의 평균수명의 차이는 17.4102일에서 22.5898일 사이라고 추정할 수 있다.
```

두 집단의 자유도가 30이 넘는경우

1. **우유를 생산하는 두 회사 A와 B가 있는데, 두 회사 우유의 평균용량이 얼마나 차이 나는지를 비교하려고 한다. 그래서 각각 60개와 70개의 우유를 표본으로 뽑아서 조사하였더니, 평균용량은 250mL와 210mL가 나왔고, 표본분산은 13mL와 9mL가 나왔다고 한다. 이때 두 회사 우유의 평균용량의 차이에 대한 99%의 신뢰구간을 구하시오.**





그리고 "두 모평균의 신뢰구간"도 단일 "모평균의 신뢰구간"과 마찬가지로 표본의 수가 많아지면 정규분포를 사용한다. 그런데 단일 모평균에서는 n≥30이라는 명확한 기준이 있었는데, 두 모평균의 신뢰구간에서는 "대표본"과 "소표본"이라고만 나타낼 뿐, 명확한 기준이 없다. 왜냐하면 그 이유는 t분포표에 있는데, 보통 t분포는 표본이 적을 때 사용하려고 만든 분포이기에, 표본의 수가 31개(자유도 기준으로 30)를 넘어가면 사용할 수가 없다.

==> 라고 보통 통계책에서 나오는데.. 그이유는 t 분포가 31개 이상이 없기 때문인데.. 지금 컴퓨터 발달로 31개 이상에도 값을 알수 있다.

$$(\bar{x}_1 - \bar{x}_2) \pm t\alpha_{/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm Z\alpha_{/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

〈소표본〉 〈대표본〉

In [73]:
```
import numpy as np
import scipy.stats as stats

alpha = 0.01

x1 = 250
x2 = 210
n1 = 60
n2 = 70
s1 = 13
s2 = 9

statistic = stats.t.ppf(1 - alpha/2, n1+n2-2)
sp = np.sqrt(((n1 - 1)* s1 + (n2 - 1)* s2) / (n1+n2-2))

up = (x1 - x2) + statistic * sp * np.sqrt((1 / n1) + (1 / n2))
down = (x1 - x2) - statistic * sp * np.sqrt((1 / n1) + (1 / n2))

print('두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) %6.4f <= u1 - u2 <= %6
print('두 건전지의 평균수명의 차이는 %6.4f일에서 %6.4f일 사이라고 추정할 수 있
```

```
두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) 38.4851 <= u1 - u2 <= 41.5149
두 건전지의 평균수명의 차이는 38.4851일에서 41.5149일 사이라고 추정할 수 있다.
```

In [86]:
```
statistic = stats.norm.ppf(1 - alpha/2)
up = (x1 - x2) + statistic *  np.sqrt((s1 / n1) + (s2 / n2))
down = (x1 - x2) - statistic *  np.sqrt((s1 / n1) + (s2 / n2))

print('두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) %6.4f <= u1 - u2 <= %6
print('두 건전지의 평균수명의 차이는 %6.4f일에서 %6.4f일 사이라고 추정할 수 있
```

```
두 모평균 차이의 신뢰구간 예제(σ를 모르는 경우) 38.4865 <= u1 - u2 <= 41.5135
두 건전지의 평균수명의 차이는 38.4865일에서 41.5135일 사이라고 추정할 수 있다.
```

두 모평균 차이의 가설검정(σ를 모르는 경우)

두 모평균 차이의 가설검정은 "σ를 모르는 경우"가 대부분인데, σ를 모르는 경우에는 t분포를 사용한다. 그리고 검정통계량 구하는 공식은 아래와 같은데, 신뢰구간에서 사용한 공식을 그대로 사용한다.

검정통계량 공식에 있는 sp는 합동표준편차라고 부르는데, 두 집단의 표본을 모아서 한 번에 계산한 표준편차이다.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

귀무가설과 대립가설

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

〈 t분포 〉

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

〈 정규분포 〉

1. 공대생과 인문대생의 월평균독서량이 차이가 나는지를 알아보려고 한다.

그래서 공대생 10명과 인문대생 11명을 뽑아 월평균독서량을 조사하였더니, 평균은 각각 5권과 7권이 나왔고, 표본분산은 각각 0.5권과 2권이 나왔다. 그럼 공대생과 인문대생의 월평균독서량이 차이가 나는지를 유의수준 5%에서 검정하시오.

귀무가설 : $u_1 - u_2 = 0$

대립가설 : $u_1 - u_2 \neq 0$

In [10…

```
import numpy as np
import scipy.stats as stats

alpha = 0.05

x1 = 5
x2 = 7
n1 = 10
n2 = 11
s1 = 0.5
s2 = 2

statistic = stats.t.ppf(1 - alpha/2, n1+n2-2)
sp = np.sqrt(((n1 - 1)* s1 + (n2 - 1)* s2) / (n1+n2-2))

t = (x1 - x2) / (sp * np.sqrt(1/n1 + 1/n2))

p_value = stats.t.cdf(t, n1+n2-2) *2

statistic, t, p_value
```

Out[10… (2.093024054408263, -4.0309781974574195, 0.0007138501522682035)

기각역이 $\pm$ 2.093 이고 t 값이 -4.03 이기 때문에 기각역 안에 위치 하므로 귀무가설 기각

공대생과 인문대생의 월평균 독서량은 차이가 난다고 할수 있다

1. **두 개의 진통제 A와 B가 있는데, 진통제 B의 지속시간이 7시간 더 길다고 알려져 있다. 하지만 일부에서는 지속시간의 차이가 7시간보다는 작을 것이라는 소리를 하고 있다. 그래서 실제로 어떠한지를 알아보기 위해 각각 11개와 9개의 표본을 뽑았더니, 평균 지속시간은 각각 17시간과 23시간이 나왔고, 표본분산은 각각 6시간과 7시간이 나왔다. 그럼 두 진통제의 지속시간의 차이가 7시간보다 작다고 할 수 있는지, 유의수준 10%에서 검정하시오.**

귀무가설 : $u_1 \leqq u_2 - 7$ ==> $u_1 - u_2 \leqq -7$

대립가설 : $u_1 > u_2 - 7$ ==> $u_1 - u_2 > -7$

In [10…

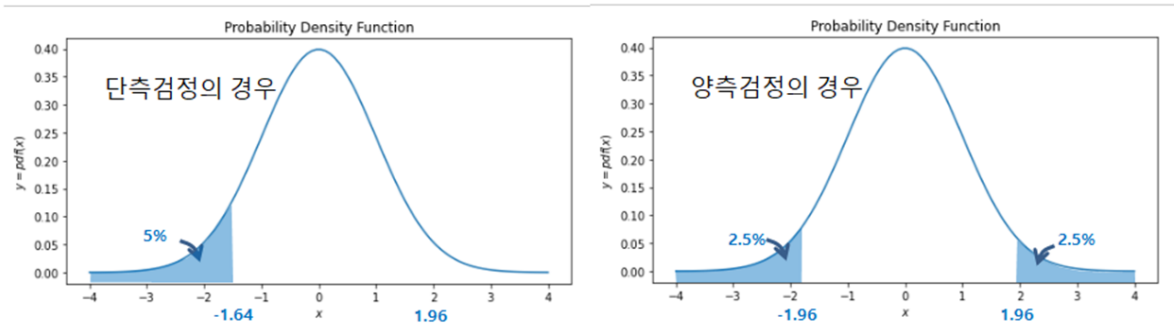
```
import numpy as np
import scipy.stats as stats

alpha = 0.1
```

```
x1 = 17
x2 = 23
n1 = 11
n2 = 9
s1 = 6
s2 = 7

statistic = stats.t.ppf(1 - alpha, n1+n2-2)
sp = np.sqrt(((n1 - 1)* s1 + (n2 - 1)* s2) / (n1+n2-2))

t = ((x1 - x2) + 7)  / (sp * np.sqrt(1/n1 + 1/n2))

p_value = 1- stats.t.cdf(t, n1+n2-2)

statistic, t, p_value
```

Out[10… (1.330390943569909, 0.8764151118481825, 0.196175430579401)

유의수준 α=0.1이고 자유도는 11+9-2=18이므로, t분포표에서 해당하는 값을 찾으면 1.33이 나온다. 그래서 기각역은 1.33이므로, 검정통계량은 "채택역"안에 위치하게 된다. 그래서 귀무가설이 채택되므로 진통제 B의 지속시간이 7시간 더 길다고 할 수 있다.

Pvalue 구하는 법



ex) 라인 1에서 생산된 초콜릿 표본 25개의 무게는 평균 198.5g , 표준편차 4.8g, 라인 2에서 생산된 초콜릿 표본 34개의 무게는 평균 201.3g, 표준편차 5.1g으로 측정되었다. 모평균의 차이가 있는지 유의 수준 5% 검정

귀무가설은 u1-u2 = 0

대립가설은 u1-u2 != 0

In [11…

```
import numpy as np
import scipy.stats as stats

alpha = 0.05

x1 = 198.5
x2 = 201.3
n1 = 25
n2 = 34
s1 = 4.8**2
s2 = 5.1**2
```

```
statistic = stats.t.ppf(1 - alpha/2, n1+n2-2)
sp = np.sqrt(((n1 - 1)* s1 + (n2 - 1)* s2) / (n1+n2-2))

t = ((x1 - x2) )  / (sp * np.sqrt(1/n1 + 1/n2))

'''
양측 검정이기 때문에 곱하기 2를 해주는거고
만약에 t 통계량이 양수로 나올경우는 (1 - stats.t.cdf(t, n1+n2-2) ) * 2 를 해0
'''
p_value = stats.t.cdf(t, n1+n2-2) * 2

statistic, sp, t, p_value
```

Out[11… (2.0024654580545986,
4.975889235524694,
-2.135850172171308,
0.03699983287879092)

In [ ]:

두 모비율 차이의 가설 검정

두 모비율의 신뢰구간은 이전 글에서 다루었던 “두 모평균”과 기본적인 개념은 비슷한데, “두 모평균”과 마찬가지로 각 집단의 모비율을 따로따로 추정하는 것이 아니라, 두 집단의 모비율이 서로 얼마나 차이 나는지를 추정하는 것이다. 참고로 두 집단의 모비율을 서로 어떻게 비교하는지에 대해서 몇 가지 예를 들면 아래와 같다.

여당과 야당의 지지율(모비율)은 얼마나 차이 나는가?

대졸과 고졸의 취업률(모비율)은 얼마나 차이 나는가?

선진국과 개발도상국의 출산율(모비율)은 얼마나 차이 나는가?

흡연자와 비흡연자의 폐암 발생률(모비율)은 얼마나 차이 나는가?

동양인과 서양인의 비만율(모비율)은 얼마나 차이 나는가?

Z 통계량

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

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

집단 1개 추가

〈 집단 1개 〉 〈 집단 2개 〉

신뢰구간

$$-Z\alpha/_2 \le \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \le Z\alpha/_2$$

$$-Z\alpha/_2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le (\hat{p}_1-\hat{p}_2)-(p_1-p_2) \le Z\alpha/_2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$-(\hat{p}_1-\hat{p}_2)-Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le -(p_1-p_2) \le -(\hat{p}_1-\hat{p}_2)+Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(\hat{p}_1-\hat{p}_2)+Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \ge p_1-p_2 \ge (\hat{p}_1-\hat{p}_2)-Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(\hat{p}_1-\hat{p}_2)-Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(\hat{p}_1-\hat{p}_2) \pm Z\alpha/_2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \longleftarrow \text{신뢰구간 공식}$$

두 모비율 차이의 신뢰구간 풀이

1. 어느 제조업 공장에 제품을 생산하는 두 대의 기계 A와 B가 있는데, 각 기계에서 생산하는 제품의 불량률이 서로 얼마나 차이 나는지를 조사한다고 한다. 그래서 각 기계에서 100개와 150개의 표본을 뽑아 불량품의 개수를 체크하였더니, 각각 9개와 3개가 불량품이었다. 이때 두 기계의 불량률의 차이에 대한 90%의 신뢰구간을 추정하시오.

In [13…

```
alpha = 0.1

n1 = 100
n2 = 150

p1 = 9 / n1
p2 = 3 / n2


statistic = stats.norm.ppf(1-alpha/2)

up = (p1 - p2) + (statistic * np.sqrt( (p1*(1-p1) / n1) + (p2* (1-p2) / n2)
```

```
down = (p1 - p2) - (statistic * np.sqrt( (p1*(1-p1) / n1) + (p2* (1-p2) / n
print('두 모비율 차이의 신뢰구간 예제 %6.4f <= p1 - p2 <= %6.4f'%(down,up))
```

```
두 모비율 차이의 신뢰구간 예제 0.0193 <= p1 - p2 <= 0.1207
```

1. 선거에 앞서 특정 정당에 대한 A지역과 B지역의 지지율이 서로 얼마나 차이 나는지를 조사한다고 한다. 그래서 각 지역에서 40명과 50명을 대상으로 특정 정당의 지지율을 조사하였더니, 각각 75%와 46%가 해당 정당을 지지하는 것으로 나왔다. 이때 두 지역의 지지율의 차이에 대한 95%의 신뢰구간을 구하시오.

In [13…

```
alpha = 0.05

n1 = 40
n2 = 50

p1 = 0.75
p2 = 0.46


statistic = stats.norm.ppf(1-alpha/2)

up = (p1 - p2) + (statistic * np.sqrt( (p1*(1-p1) / n1) + (p2* (1-p2) / n2)
down = (p1 - p2) - (statistic * np.sqrt( (p1*(1-p1) / n1) + (p2* (1-p2) / n
print('두 모비율 차이의 신뢰구간 예제 %6.4f <= p1 - p2 <= %6.4f'%(down,up))
```

```
두 모비율 차이의 신뢰구간 예제 0.0974 <= p1 - p2 <= 0.4826
```

두 모비율 차이의 가설검정

두 모비율의 가설검정은 "두 모비율의 관계가 이럴 것이다"라는 두 개의 가설 중, 어느 가설이 더 타당한지를 판단하는 것이다. 즉 모비율 p1과 p2를 모르는 상태인데, 그래서 검정통계량에 나와 있는 p1과 p2는 실제의 모비율이 아니라 가설 속의 모비율이다.(귀무가설과 대립가설 속에 나오는 모비율) 그리고 p1-p2=0이라고 나와 있으므로, 검정통계량의 p1-p2에는 0을 대입하면 된다. (문제를 응용하면, 0 이외에 다른 수치도 사용할 수 있다)

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

가설 속의 모비율

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

합동 표본비율

구하는 표본 수

$$\hat{p}_0 = \frac{x_1 + x_2}{n_1 + n_2}$$

전체 표본 수

1. 동일한 제품을 생산하는 2대의 기계가 있는데, 두 기계에서 생산하는 제품의 불량률은 서로 같은 것으로 알려져 있다. 하지만 최근 품질관리팀에서 불량률이 서로 다를 수도 있다는 의견이 나오고 있어서, 실제로 어떠한지를 알아보기 위해 각각 표본 120개와 130개를 뽑았더니, 불량품은 각각 6개와 4개가 나왔다. 그럼 두 기계에서 생산하는 제품의 불량률은 서로 다르다고 할 수 있는지 유의수준 5%에서 검정하시오.

귀무가설 : p1 - p2 = 0

대립가설 : p1 - p2 != 0

In [16…

```
alpha = 0.05

n1 = 120
n2 = 130
```

```
p1 = 6 / n1
p2 = 4 / n2

x1 = 6
x2 = 4

statistic = stats.norm.ppf(1-alpha/2)
p0 =  (x1 + x2 ) / (n1 + n2)

z = (p1-p2) / np.sqrt(p0 *(1-p0) *(1/n1 + 1/n2))
p_value = (1- norm.cdf(z)) *2

statistic, z, p_value
```

Out[16… (1.959963984540054, 0.775217091182553, 0.43821139060496894)

그다음 유의수준 α=0.05인데, 양측검정이므로 α/2=0.025에 해당하는 Z값 1.96이다. 그런데 양쪽으로 설정해야 하므로 기각역은 ±1.96이기에, 검정통계량은 "채택역"안에 위치하게 된다. 그래서 귀무가설이 채택이므로 두 기계에서 생산하는 제품의 불량률은 서로 같다고 할 수 있다.

1. **도시와 시골에 사는 사람들의 흡연율을 조사하고 있는데, 해당 조사팀에 의하면 도시에 사는 사람들의 흡연율이 더 높을 수도 있다는 의견이 나오고 있다. 이에 실제로 어떠한지를 알아보기 위해 각각 200명과 150명을 뽑아 흡연율을 조사하였더니, 흡연자는 총 80명과 50명이 나왔다. 그럼 도시에 사는 사람들의 흡연율이 더 높다고 할 수 있는지 유의수준 1%에서 검정하시오.**

In [17…

```
alpha = 0.01

n1 = 200
n2 = 150

p1 = np.round(80 / n1, 2)
p2 = np.round(50 / n2 , 2)

x1 = 80
x2 = 50

statistic = stats.norm.ppf(1-alpha)
p0 =  (x1 + x2 ) / (n1 + n2)

z = (p1-p2) / np.sqrt(p0 *(1-p0) *(1/n1 + 1/n2))
p_value = (1- norm.cdf(z))

statistic, z, p_value
```

Out[17… (2.3263478740408408, 1.3412498085558295, 0.08991967984623273)

다음으로 유의수준 α=0.01인데, 0.01에 해당하는 Z값은 2.33이다. 그래서 검정통계량이 "채택역"안에 위치하므로 귀무가설이 채택된다. 그러므로 도시에 사는 사람들의 흡연율이 더 높다고 할 수 없다.

두 모분산 차이의 가설 검정

두 모분산 차이의 신뢰구간 구하는 법

먼저 두 모분산의 신뢰구간 역시 이전에 알아보았던 "평균"과 "비율"이랑 마찬가지로, 두 집단의 모분산이 서로 얼마나 차이 나는지를 파악하는 것이다. 그런데 평균과 비율은 "뺄셈"을 활용해서 두 집단을 비교하였지만, 분산은 "나눗셈"을 활용해서 두 집단을 비교한다

$$\frac{4}{4} = 1$$

$$\frac{분산1}{분산2} = 1$$

〈 두 분산은 차이가 없다 〉

$$\frac{8}{4} = 2$$

$$\frac{분산1}{분산2} = 2$$

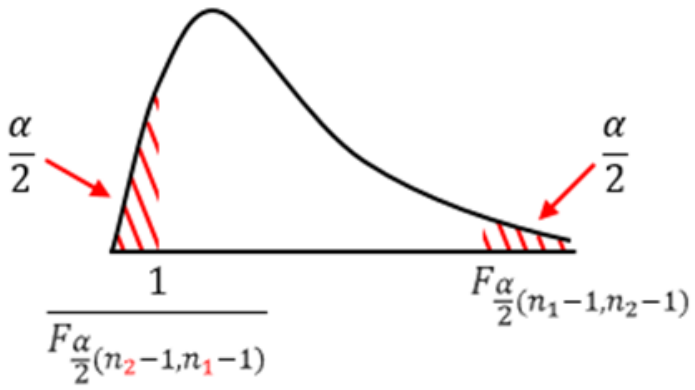
〈 두 분산은 차이가 크다 〉

그리고 뺄셈을 사용하지 않고 나눗셈을 사용하는 이유는 바로 "확률분포" 때문인데, 이전에 다루었던 평균과 비율은 뺄셈을 해도 정규분포나 t분포를 사용할 수 있었지만, 분산은 뺄셈을 하면 사용할 확률분포가 없다.

그래서 나눗셈을 하는 것이고, 2개의 분산을 나눴을 때 사용할 수 있는 분포가 F분포이다. 그래서 두 모분산 차이의 신뢰구간은 F분포를 사용하는데, 2개의 카이제곱분포 공식을 서로 나눠주면 F분포 공식이 나온다.

신뢰구간

두 모분산의 신뢰구간은 F분포 그래프의 α/2에 해당하는 양쪽 x축 좌표를 사용하는데, 한 가지 조심할 것은 왼쪽 x축 좌표는 분자와 분모의 자유도가 서로 바뀐다.



$$\frac{1}{F_{\frac{\alpha}{2}(n_2-1,\,n_1-1)}} \le \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \le F_{\frac{\alpha}{2}(n_1-1,\,n_2-1)}$$

$$\frac{s_2^2}{s_1^2} \times \frac{1}{F_{\frac{\alpha}{2}(n_2-1,\,n_1-1)}} \le \frac{\sigma_2^2}{\sigma_1^2} \le \frac{s_2^2}{s_1^2} \times F_{\frac{\alpha}{2}(n_1-1,\,n_2-1)} \quad \longleftarrow \text{각 변에 } \frac{s_2^2}{s_1^2} \text{을 곱한다.}$$

$$\frac{s_1^2}{s_2^2} \times F_{\frac{\alpha}{2}(n_2-1,\,n_1-1)} \ge \frac{\sigma_1^2}{\sigma_2^2} \ge \frac{s_1^2}{s_2^2} \times \frac{1}{F_{\frac{\alpha}{2}(n_1-1,\,n_2-1)}} \quad \longleftarrow \text{역수를 취한다.}$$

$$\frac{s_1^2}{s_2^2} \times \frac{1}{F_{\frac{\alpha}{2}(n_1-1,\,n_2-1)}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} \times F_{\frac{\alpha}{2}(n_2-1,\,n_1-1)} \quad \longleftarrow \text{양변을 서로 바꾼다.}$$

또는 아래 공식으로

$$\left[\frac{S_1^2/S_2^2}{F_{1-a/2;(n_1-1,n_2-1)}},\ \frac{S_1^2/S_2^2}{F_{a/2;(n_1-1,n_2-1)}}\right]$$

문제 1 두 라인에서 생산되는 초콜릿의 무게는 정규분포를 따른다고 한다. 라인1에서 생산된 초콜릿 표본 25개 무게와 라인 2에서 생산된 초콜릿 표본 34개의 무게는 아래 와 같다. 두 모집단 모분산 비율에 대한 95% 신뢰구간을 구하시오

In [34]:
```
import numpy as np
import scipy.stats as stats

line1 = np.array([200, 203, 201, 194, 195, 202, 200, 199, 204, 199, 195,
 196, 199, 200, 199, 198, 200, 198, 199, 199, 197, 194, 197, 193, 202])

line2 = np.array([204, 201, 196, 202, 205, 205, 197, 209, 197, 201, 187, 20
         203, 200, 207, 201,213, 198, 198, 208, 197, 197, 199, 194, 203, 20
```

```
alpha = 0.05

f = line1.var(ddof=1) / line2.var(ddof=1)

down = f / stats.f.ppf(1 - alpha/2, dfn = len(line1)-1, dfd = len(line2)-1
up = f/ stats.f.ppf(alpha/2, dfn = len(line1)-1, dfd = len(line2)-1 )

down, up
```

Out[34]: (0.132606116813897, 0.6055166568937755)

In [ ]:

In [38]:
```
down = f * stats.f.ppf(alpha/2, dfd = len(line1)-1, dfn = len(line2)-1 )
up = f* 1 / stats.f.ppf(alpha/2, dfn = len(line1)-1, dfd = len(line2)-1 )
down, up
```

Out[38]: (0.13260611681389697, 0.6055166568937755)

문제2. 어떤 특정실험을 하는 방법이 2가지가 있는데, 실험1과 실험2의 분산이 서로 얼마나 차이 나는지를 비교하려고 한다. 그래서 실험마다 표본 9개와 6개를 뽑았더니, 분산은 각각 14와 12가 나왔다. 이때 두 실험의 모분산 차이에 대한 90%의 신뢰구간을 구하시오.

In [68]:
```
import numpy as np
import scipy.stats as stats

n1 = 9
n2 = 6

s1 = 14
s2 = 12

alpha = 0.1

f = s1 / s2

'''
두번째 공식
'''
down = f / stats.f.ppf(1 - alpha/2, dfn = n1-1, dfd = n2-1 )
up = f/ stats.f.ppf(alpha/2, dfn = n1-1, dfd = n2-1 )

down, up
```

Out[68]: (0.24213144396776046, 4.302081777396698)

In [69]:
```
'''
첫번째 공식
'''
down = f * 1 / stats.f.ppf(1- alpha/2, n1-1, n2-1 )
up = f* stats.f.ppf(1-alpha/2, n2-1, n1-1 )
down, up
```

Out[69]: (0.24213144396776046, 4.302081777396698)

문제3. 감기약의 지속시간을 테스트하기 위해서 알약1과 알약2의 분산이 서로 얼마나 차이 나는지를 비교하려고 한다. 그래서 각각 표본 11개와 8개를 뽑았더니, 분산은 11과 15가 나왔다. 이때 두 알약의 모분산 차이에 대한 95%의 신뢰구간을 구하시오.

In [77]:
```
import numpy as np
import scipy.stats as stats

n1 = 11
n2 = 8

s1 = 11
s2 = 15

alpha = 0.05

f = s1 / s2

'''
두번째 공식
'''
down = f / stats.f.ppf(1 - alpha/2, dfn = n1-1, dfd = n2-1 )
up = f/ stats.f.ppf(alpha/2, dfn = n1-1, dfd = n2-1 )

down, up
```

Out[77]: (0.15402549871348065, 2.896537650555497)

두 모분산 차이의 가설 검정 하는법

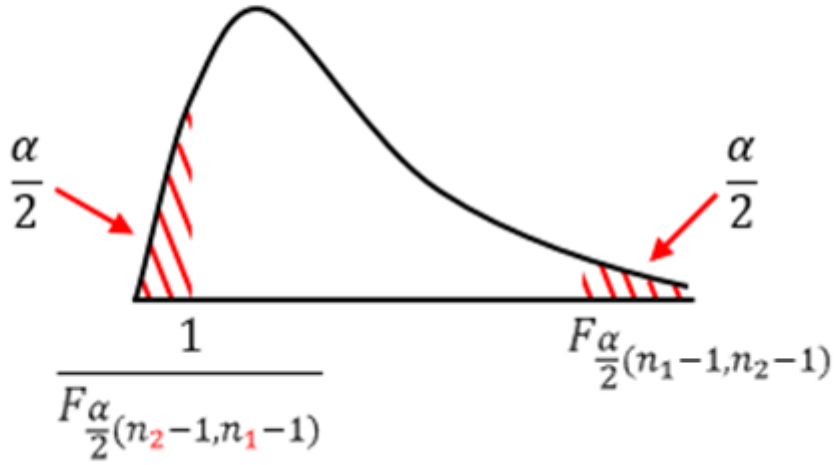
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

가설 속의 모분산

$$F = \frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2}$$

두 모분산의 가설검정을 할 때는 기본적으로 F분포를 사용하는데, F분포를 사용해서 검정통계량과 기각역을 구한다. 그리고 기각역은 아래와 같은데, 단지 **왼쪽 기각역을 구할 때는 자유도가 서로 바뀌며, 추가로 1/F를 해줘야 하기 때문에 조심해야 한다.** 그리고 양측검정을 할 때는 α/2를 하면 된다.



문제1. 동일한 제품을 생산하는 기계1과 기계2가 있는데, 두 기계에서 생산하는 제품의 분산은 서로 같은 것으로 알려져 있다. 하지만 일부에서는 기계1에서 생산한 제품의 분산이 더 큰 것 같다는 의견이 나오고 있어서, 이에 실상을 파악하기 위해 각각 표본 6개와 12개를 뽑았더니, 표본분산은 각각 30과 8이 나왔다. 그럼 기계1에서 생산하는 제품의 분산이 더 크다고 할 수 있는지 유의수준 10%에서 검정하시오.

귀무가설 : $\sigma_1^2 <= \sigma_2^2$

대립가설 : $\sigma_1^2 > \sigma_2^2$

In [91]:
```
alpha = 0.1
s1 = 30
s2 = 8
n1 = 6
n2 = 12

f = s1 / s2
value = stats.f.ppf(1 - alpha, dfn = n1-1, dfd = n2-1 )
p = 1 - stats.f.cdf(f, dfn = n1-1, dfd = n2-1)

value , f, p
```

Out[91]: (2.451184342974802, 3.75, 0.031581122675903 52)

문제2. 어떤 특정실험을 하는 방법이 2가지가 있는데, 실험1과 실험2의 분산은 서로 동일한 것으로 알려져 있다. 그런데 최근에는 실험1의 결과가 더 정확하게 나와서, 실험1의 분산이 더 작을 수도 있다는 의견이 나오고 있다.

이에 실제로 그러한지를 파악하기 위해 각각 5번과 7번의 실험을 하였더니, 표본분산은 각각 21과 25가 나왔다고 한다. 그럼 실험1의 분산이 더 작다고 할 수 있는지 유의수준 5%에서 검정하시오.

귀무가설 : $\sigma_1^2 >= \sigma_2^2$

대립가설 : $\sigma_1^2 < \sigma_2^2$

In [93]:
```
alpha = 0.05
s1 = 21
s2 = 25
n1 = 5
n2 = 7

f = s1 / s2
'''
왼쪽 기각역을 구해야 하기 때문에 자유도를 서로 바꿔줘야 한다.
F분포는 숫자 1을 기준으로 좌측검정과 우측검정을 결정한다
'''
value = 1/ stats.f.ppf(1 - alpha, n2-1, n1-1 )
p = stats.f.cdf(f, dfn = n1-1, dfd = n2-1)

value , f, p
```

Out[93]: (0.16225515762640028, 0.84, 0.4529253177409753)

문제3. 집단1과 집단2의 분산은 서로 동일한 것으로 알려져 있는데, 최근에는 두 집단의 분산이 서로 다를 수도 있다는 의견이 나왔다. 그래서 실제로 어떠한지를 알아보기 위해 각각 표본 11개와 8개를 뽑았더니, 표본분산은 각각 15와 10이 나왔다. 그럼 두 집단의 분산이 서로 다르다고 할 수 있는지 유의수준 5%에서 검정하시오.

귀무가설 : $\sigma_1^2 = \sigma_2^2$

대립가설 : $\sigma_1^2 != \sigma_2^2$

In [12…
```
alpha = 0.05
s1 = 15
s2 = 10
n1 = 11
n2 = 8

f = s1 / s2
'''
같지 않다 이기 때문에 양측 검정
'''

left = 1/ stats.f.ppf(1 - alpha/2, n2-1, n1-1 )
right = stats.f.ppf(1 - alpha/2, n1-1, n2-1 )

'''
cdf 누적 분포 함수
양측 검정이기 때문에 곱하기 2
'''
p = 2.0*(1.0 - stats.f.cdf(f, n1-1, n2-1))
```

```
print('F:',f)
print(left, right)
print('P value' , p)
'''
검정 통계량이 채택역 안에 있고 pvalue 가 0.05보다 크기 때문에
귀무가설을 채택 하여 두 집단의 분산은 서로 같다고 할 수 있다.
'''
```

```
F: 1.5
0.2531758332893394 4.761116434996814
P value 0.6066549462641624
```

Out[12… '₩n검정 통계량이 채택역 안에 있고 pvalue 가 0.05보다 크기 때문에₩n귀무가설을 채택 하여 두 집단의 분산은 서로 같다고 할 수 있다.₩n'

In [12…

```
import scipy.stats as stats

def f_test(x, y, alt="two_sided"):
    """
    Calculates the F-test.
    :param x: The first group of data
    :param y: The second group of data
    :param alt: The alternative hypothesis, one of "two_sided" (default), "
    :return: a tuple with the F statistic value and the p-value.
    """
    df1 = len(x) - 1
    df2 = len(y) - 1
    f = x.var(ddof=1) / y.var(ddof=1)
    if alt == "greater":
        p = 1.0 - stats.f.cdf(f, df1, df2)
    elif alt == "less":
        p = stats.f.cdf(f, df1, df2)
    else:
        # two-sided by default
        # Crawley, the R book, p.355
        p = 2.0*(1.0 - stats.f.cdf(f, df1, df2))
    return f, p
```