$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad np \qquad \hat{p} = \frac{x}{n}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad np(1-p)$$

$$\bar{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \qquad V(X) = E[X^2] - (E[X])^2$$

$$= \cancel{n^2p^2} - np^2 + np - \cancel{(np)^2}$$

$$= np(1-p)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$(n-1)s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad X \sim B(n,p)$$

$$E[\bar{X}] = E\left[\frac{1}{n}(x_1 + \cdots + x_n)\right] = \frac{1}{n}\Big(\underbrace{E[x_1]}_{\mu} + \underbrace{E[x_2]}_{\mu} + \cdots \underbrace{E[x_n]}_{\mu}\Big)$$

$$= \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot \mu = \mu$$

$$\mathrm{var}(\bar{X}) = \mathrm{Var}\left[\frac{1}{n}(x_1 + \cdots x_n)\right] = \frac{1}{n^2}\Big(\underbrace{V[x_1]}_{\sigma^2} + \cdots + \underbrace{V[x_n]}_{\sigma^2}\Big)$$

$$= \frac{1}{n^{\cancel{2}}} \cdot \cancel{n} \cdot \sigma^2 = \frac{\sigma^2}{n}$$

$$E[s^2] = \sigma^2 \qquad \mathrm{Var}(s^2) = 2\sigma^4/(n-1)$$

$$E[\hat{p}] = p \qquad \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

신뢰구간

1 proportion

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

2 proportion

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

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

2×2 일때

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

2 sample t

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$ 클때

paired t

$$\frac{\bar{d}}{s_d/\sqrt{n}}$$

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

작을때

$$Y = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

chi squared.

chi2_contingency

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \longrightarrow ?\ \left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right) \sim \chi^2_{[(r-1)(c-1)]}$$

$$[(x_i - \bar{x}) + (\bar{x} - \mu)]^2$$

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{\sigma^2} + \sum_{i=1}^{n}\frac{(\bar{x} - \mu)^2}{\sigma^2} - \sum_{i=1}^{n} 2(x_i - \bar{x})(\bar{x} - \mu)$$

$$\sum_{i=1}^{n} x_i - n\bar{x} = 0 \qquad \bar{x} = \frac{\sum x_i}{n}$$

$$\sum_{i=1}^{n} Z_i^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{n(\bar{x} - \mu)^2}{\sigma^2}$$

$$\frac{(\bar{x} - \mu)^2}{\sigma^2/n}$$

$$\therefore \underbrace{\frac{(n-1)s^2}{\sigma^2}}_{\sim \chi^2_{(n-1)}} = \underbrace{\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2}_{\sim \chi^2_{(n)}} - \underbrace{\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2}_{\sim \chi^2_{(1)}}$$

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1,\, n_2-1)} \quad \frac{Y_1/\nu_1}{Y_2/\nu_2}$$

ANOVA

k : level n : Data count

A, B, C (k)

ANOVA

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{k} y_{ij}^2 - (nk)\bar{y}_{..}^2$$

$$SSA = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$$

$\bar{y}_{i.}$: 레벨평균 $\bar{y}_{..}$: 전체 평균   between.

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(\underline{y_{ij}}-\bar{y}_{i.})^2$$

y_{ij} : value $\bar{y}_{i.}$: i레벨 평균   within

분산 $\sim \chi^2$

$MSA = SSA/(k-1)$ (각 레벨평균 − 전체평균)^{2}의 합 = var

$MSE = SSE/k(n-1)$ (각 값 − 레벨평균)^{2}의 합. = var

The ANOVA F statistic $= \dfrac{MSA}{MSE}$ (between / within) $\sim F^*$

$$F^* = MSA/MSE > f_{\alpha,\, k-1,\, k(n-1)}$$

reject H_0 ($\mu_1 = \mu_2 = \mu_3$)

.05

$f_{0.05,\, k-1,\, k(n-1)}$ MSA/MSE 의 p-value가 0.05보다 작다

$\sigma_0^2 = 0$ 이라는 가정.

$\chi^2_{(n)}$은 $(Z_1 + \cdots + Z_n)^2$ random Variables

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(X_i - \bar{X}) + (\bar{X} - \mu) \right]^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[(X_i - \bar{X})^2 - 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 - \frac{2}{\sigma^2} \underbrace{\sum_{i=1}^{n} (X_i - \bar{X})}_{\sum_{i=1}^{n} X_i - n\bar{X} = 0,\ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}} (\bar{X} - \mu) + \frac{1}{\sigma^2} \sum_{i=1}^{n} \underline{(\bar{X} - \mu)^2}_{\text{상수}}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}, \quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}$$

$$= \frac{(n-1) S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2 / n}$$

$$\therefore \frac{(n-1) S^2}{\sigma^2} = \underbrace{\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2}_{\sim \chi^2_{(n)}} - \underbrace{\left[\underbrace{\frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}}}_{Z} \right]^2}_{\sim \chi^2_{(1)}}$$

$$\therefore \frac{(n-1) S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

R.V 은 의 분포를 따른다.

이산 discrete

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \sum_{all\ x} \frac{x_n}{n} x \Rightarrow \sum_{all} x \cdot p(x)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{X})^2 \rightarrow \sum_{all\ x} (x-\bar{x})^2 \frac{x_n}{n-1} \Rightarrow \sum_{all\ x} (x-\mu)^2 p(x)$$

$$\Downarrow$$

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

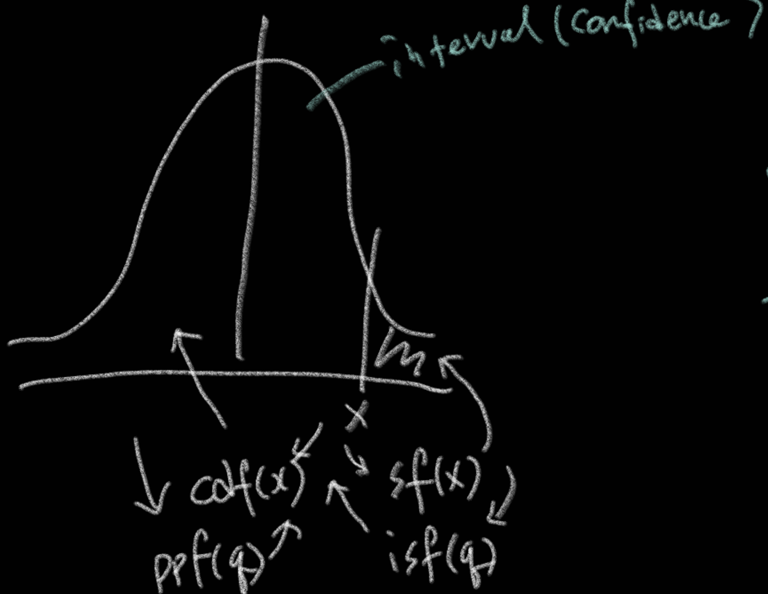
연속

continuous

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

※ continuity correction.

discrete → continuous.



∘ 분포 분위수 표현

$t_{(n-1), 1-\alpha/2}$: 양측, α 유의수준

예) $\alpha = 0.5 \rightarrow t_{n-1, 0.975}$

scipy.stats.t(n-1).ppf(0.975)

x: 검정값 (test statistics)

- norm, t, chi2, f
(cdf, ppf, sf, isf, interval)

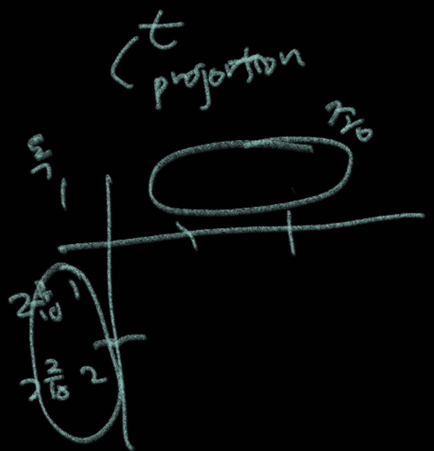
categorical (분류)

1 proportion — chisquare 1xm
1x2
2proportion — chi2_contingency
2x2 nxm

ttest_1samp (기준값)
ttest_ind   f_oneway   ols
ttest_rel
2group   2~   2~ 독립변수 (R)

- shapiro 정규성
- levene 등분산 (chi2, f)

2x2



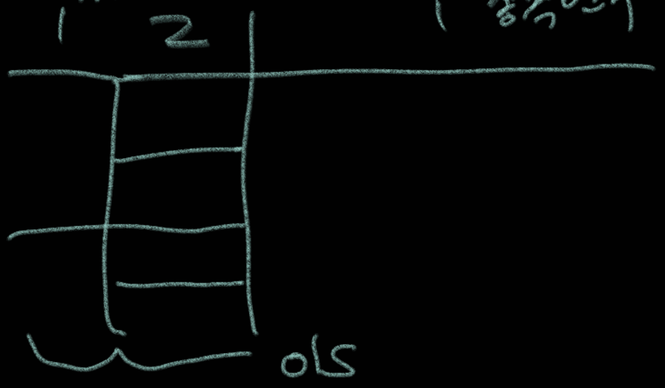
동질성

f_oneway, chi2_contingency

nxm

ANOVA.

$n_1, n_2 \times m$



$$z\ ;\quad \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})\cdot\frac{1}{n}}} \sim N(0,1),\qquad \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)\cdot\frac{1}{n}}}$$

p 신뢰구간 (p을 모르니까 σ 대신 s 쓰고 대신 $\hat{p}$ 사용)

검정 ($p=p_0$ 으로 가정했으니까)

2proportion

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

(p_1-p_2) 의 신뢰구간

$$\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2=0)}{\sqrt{\hat{p}_0(1-\hat{p}_0)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},\qquad \hat{p}_0=\frac{x_1+x_2}{n_1+n_2}$$

$$t\ ;\quad \frac{\bar{x}-\mu}{s\cdot\frac{1}{\sqrt{n}}} \sim N(0,1)\qquad\qquad \frac{\bar{x}-\mu_0}{s\cdot\frac{1}{\sqrt{n}}} \sim t(n-1)$$

검정 $\mu=\mu_0$

샘플이므니까 σ 대신 s

2sample t

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$\frac{\bar{X}_1-\bar{X}_2}{s_0\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},\qquad s_0=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

paired t

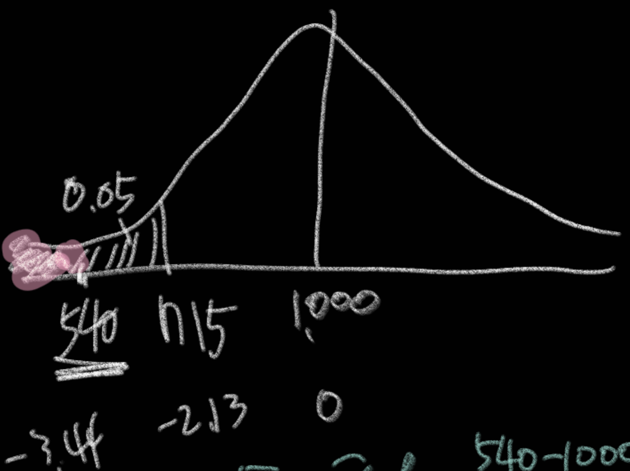
$$\frac{\bar{d}}{s_d\cdot\frac{1}{\sqrt{n}}},\qquad \bar{d} = \text{차이값 평균}$$

s_d = 차이값 s

$H_0 : \mu = 1{,}000$ 이라고 했을때. $\bar{X} = 540$

관측값이 이러니 혹시 $\mu < 1000$ 아닌지 확인해 보자

$H_a : \mu < 1{,}000$



$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{540-1{,}000}{299/\sqrt{5}}$

t(1000, 4, SE).cdf(540)

540이 1000 일때 나올확률

(SE) 오차 $299/\sqrt{5}$

t(1000, 4, SE).ppf(0.05)

= 1115 보다 관측값 540 작아.

(149 page)

$H_0 : p = .5$

$H_a : p > .5$

공평하다. $\hat{p} = 0.55$

관측값이 0.55이더나 그러면 혹시 $p > 0.5$ 아닌지 확인 해보자.

Norm(0.5, SE).sf(0.55)

0.55 보다 클 확률이 $p = 0.5$ 일때 나올확률

오차 (SE) $0.5\sqrt{1/1000}$

Norm(0.5, SE).isf(0.01) = 0.537

이므로 이보다 관측값 5.5 크다.

144 page

샘플갯수

$$E = \underbrace{Z_{\alpha/2} \cdot SE}$$ $$SE = \frac{s}{\sqrt{n}}$$

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

2 proportion.

Confidence Interval

$$-z_{\alpha/2} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_{\alpha/2}$$

test statistics.

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_0\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2}$$

2 sample t.

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

n 작을때.

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_0 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

- 평균
  - 1×m: ttest_1samp — 2×m: ttest_ind / ttest_rel — n×m: f_oneway — $n_1 \cdot n_2 \times m$: ols (2way)
- 비율
  - 1×2 → proportions_ztest (2×2) (관측, 전체)
  - chisquare (1×m) — 2×m → chi2_contingency (n×m) (연속1, 2...)
- 분산
  - chi2 (1) — f (2) — levene (2 ~)

Confusion Matrix

Ha 기준		예측(모델) 0	예측(모델) 1
True (실제)	0	TN	FP (α, 1종)
	1	FN (β, 2종)	TP

→ Ha 기준

실제	예측	
0	0	TN
0	1	FP
1	0	FN
1	1	TP

$\frac{TP}{FN + TP}$: Recall

$\frac{TN + TP}{TN + FN + FP + TP}$: accuracy

$\frac{TP}{FP + TP}$: Precision.

귀무가설(H_0)이 '참'일때 = Ha을 기각(0) 해야하는데...

1. HuggingFace : 싱 audwns 이*

2. OpenAi (MS_login)

- Text Generation Model

1. Completion Model (prompt)
2. chat Completion Model % (Message)

→ CompletionAPI는 2024.1 종료됨

→ "chat CompletionAPI" 형식으로 사용해야 함.

→ Application을 위한 Langchain

→ OpenAI : assistants API

3. Llama2

- Codellama → Interprit Code?
- Llama - sumarise → PDf, HTML... → 요약.

- fine-tuning 필요할까?
- 한글로 fine-tuning 된걸 쓸까?

1. 먼저 Local 에서 순수 Llama를 써보기

→ 한글 tuning Model 써보기

2. 어떻게 App을 구성할까?

(Langchain ?) $\boxed{\text{gradio}}$

→ 뭘로? gitHub Actions?

코드요약 Review.

→ 요약.

↳ Web, PDF 등.

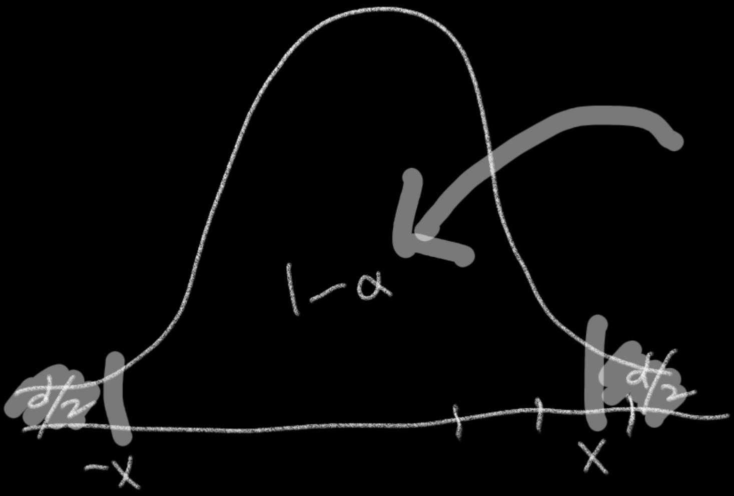
예) PDGGPT 같은거?

⇒ 성능 check?

Computing은?

로컬머신, Cloud (Azure)

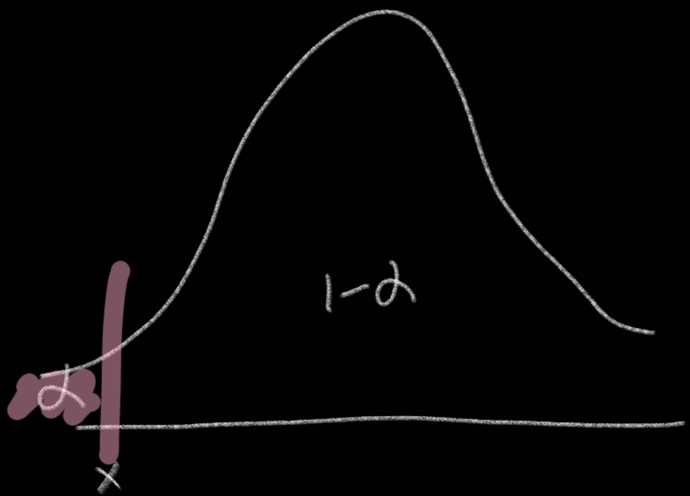
O X



interval (1-α)	cdf(x)-cdf(-x)	
90%	1.65	0.95
95%	1.96	0.975
99%	2.58	0.995
99.7	3σ	ppf

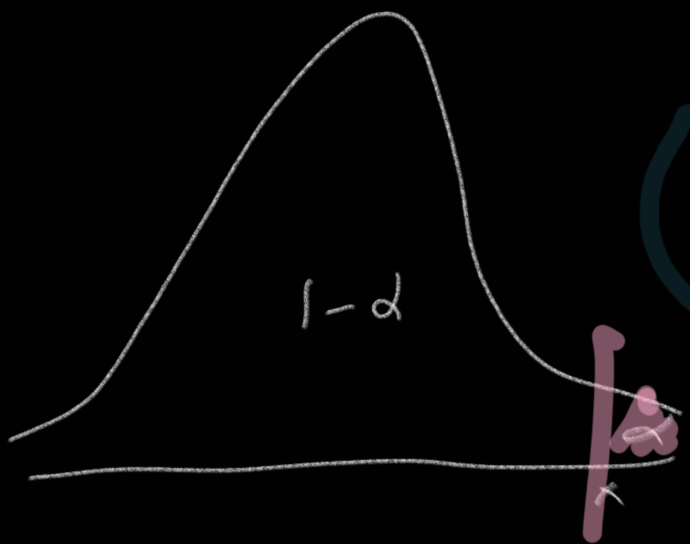
$-x = ppf(\alpha/2)$

$x = ppf(1-\alpha/2)$



$cdf(x) = 1 - sf(x)$

$ppf(\alpha) = isf(1-\alpha)$



$sf(x) = 1 - cdf(x)$

$isf(\alpha) = ppf(1-\alpha)$

※ 6-sigm : 하한(LCL)~상한(UCL) 사이에 6sigma 분포를 갖도록 하자.

허용 범위, 정상제품 / 생산제품의 분포가

H_a 기준		H_0 선택 = H_a 기각	H_0 기각 = H_a 선택
		0	1
H_0 참 = H_a 거짓	0	TN	FP
H_0 거짓 = H_a 참	1	FN	TP