Mathematical Statistics II

Ch.7 Interval Estimation

Jungsoon Choi

jungsoonchoi@hanyang.ac.kr

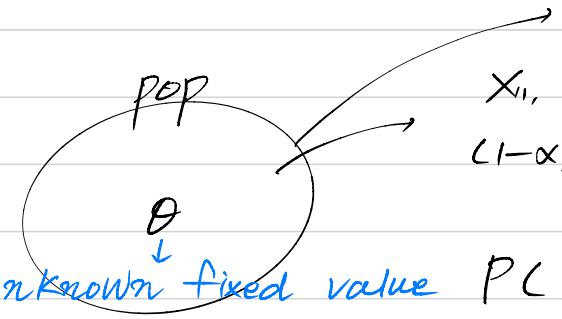
Table of Contents

Introduction - Interval Estimation

- Confidence interval: an interval of values that is likely to contain the true value of the parameter.
- A confidence interval for θ is an interval $[L, U]$ computed from the sample observations $X_1, \ldots, X_n$.
- A $(1-\alpha) \times 100\%$ confidence interval for θ is given by

$$P(L \leq \theta \leq U) = 1 - \alpha$$

  - L: lower confidence limit
  - U: upper confidence limit
  - $1-\alpha$: confidence coefficient



$X_{21}, \cdots, X_{2n}$

$X_{11}, \cdots, X_{1n}$

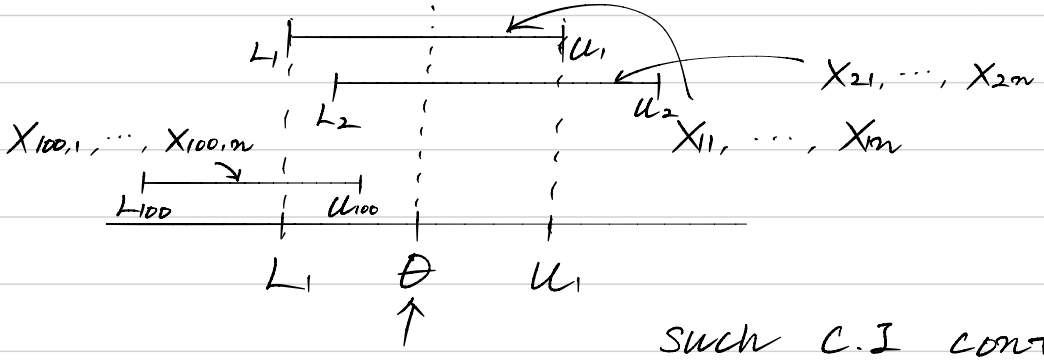
$(1-\alpha) \times 100\%$ C.I for θ

$P(L \leq \theta \leq U) = 1-\alpha$

↑ fixed

$[L, U]$

95% C.I for θ



such C.I contain parameter θ or not.

Prob. of C.I contain θ ≠ 95%.

포함하거나 안하거나. 포함할 확률을 가지는 게 아님

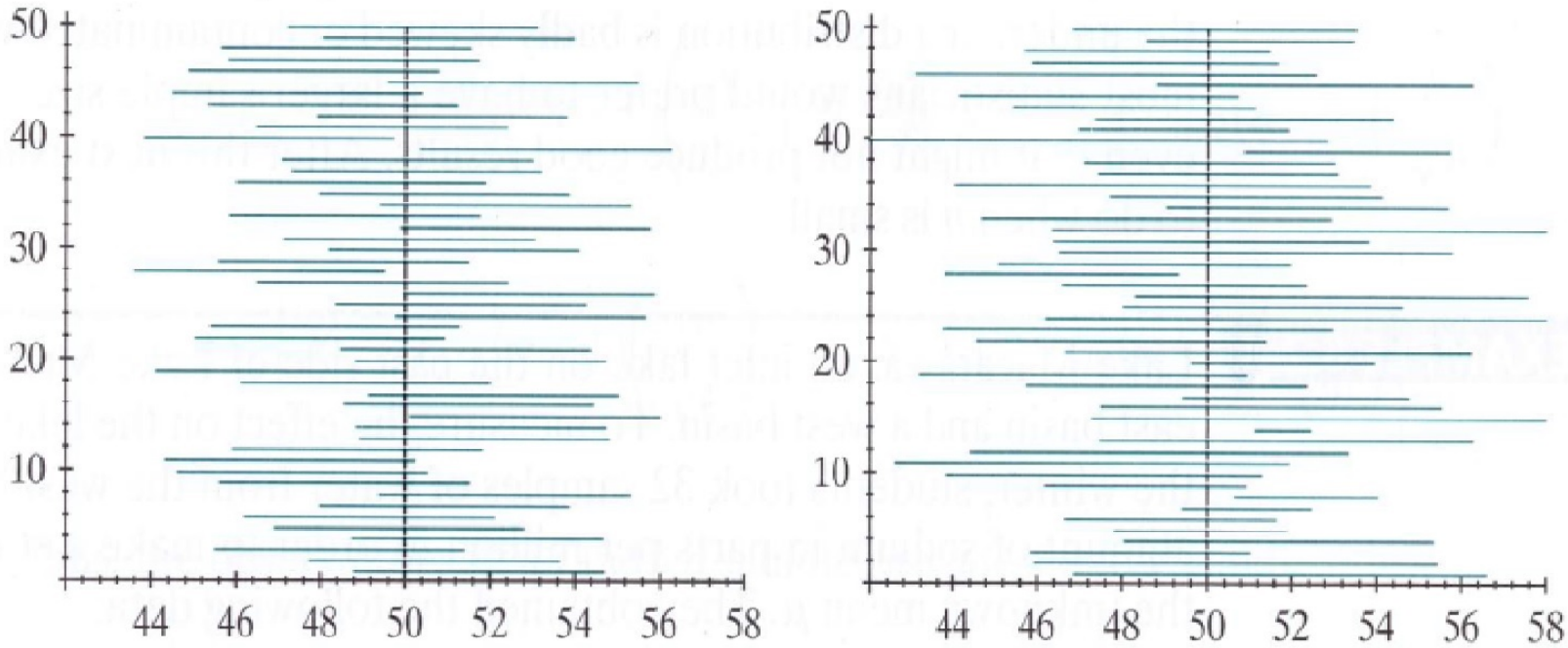
95% of C.I contain θ

5% " not contain θ.

In beisian statistic,

θ not fixed, can move. C.I

thus 95% C.I means that contain θ with 95% P.



(a) 90% confidence intervals, σ known

(b) 90% confidence intervals, σ unknown

Confidence intervals using z and t

Ch7.1 Confidence Intervals for Means

Confidence Intervals for the population mean μ

Case 1: Normal with the population variance σ^2 known

Suppose that $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. The sample mean $\overline{X}$ has a normal distribution, $\mathrm{N}(\mu, \frac{\sigma^2}{n})$. Then, the standardized random variable $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$.

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

pop → $X_i \sim N(\mu, \sigma^2)$ σ^2: known $i = 1, \cdots, n$, iid

$N(\mu, \sigma^2)$ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

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

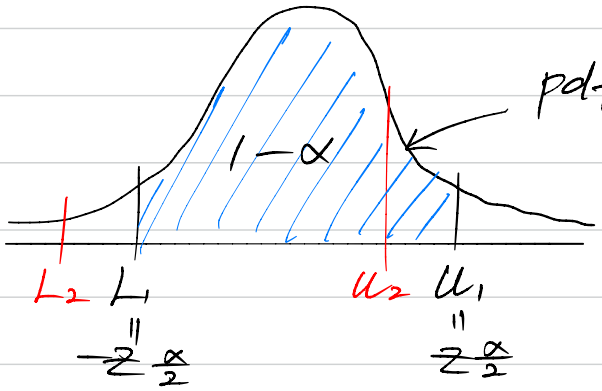
$(1-\alpha)\times 100\%$ C.I for μ

$P(\quad \le Z \le \quad) = 1-\alpha$

cf). $(1-\alpha)\times 100\%$ C.I for θ

: $P(L \le \theta \le U) = 1-\alpha$

$P(\quad \le \mu \le \quad) = 1-\alpha$



양 끝의 percentile이 같을 때 C.I의 길이가 가장 짧아짐.
가장 짧은 C.I이가 더 better.

$P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$

$\Leftrightarrow P(-z_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}) = 1-\alpha$

$\Leftrightarrow P(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

$\Leftrightarrow P(-\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

$\Leftrightarrow P(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

시험 때 따로 만들어서 사용

$$
\begin{aligned}
P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) &= P(-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) \\
&= P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
&= 1 - \alpha
\end{aligned}
$$

$(1-\alpha) \times 100\%$ confidence interval for μ with known σ^2

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Case 2: Normal with the population variance σ^2 unknown

Suppose that $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ from the observations $X_1, \ldots, X_n$.

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

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim \mathrm{t}_{(n-1)}$$

A $(1-\alpha)\times 100\%$ confidence interval for μ with unknown σ^2 is given by

$$\left[\overline{X} - t_{\alpha/2,(n-1)}\frac{S}{\sqrt{n}},\ \overline{X} + t_{\alpha/2,(n-1)}\frac{S}{\sqrt{n}}\right]$$

pop

$N(\mu, \sigma^2)$

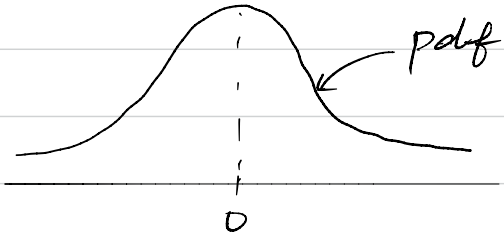
$X_i \sim N(\mu, \sigma^2)$ σ^2: unknown

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

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

unknown

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



pdf of $T \sim t(n-1)$

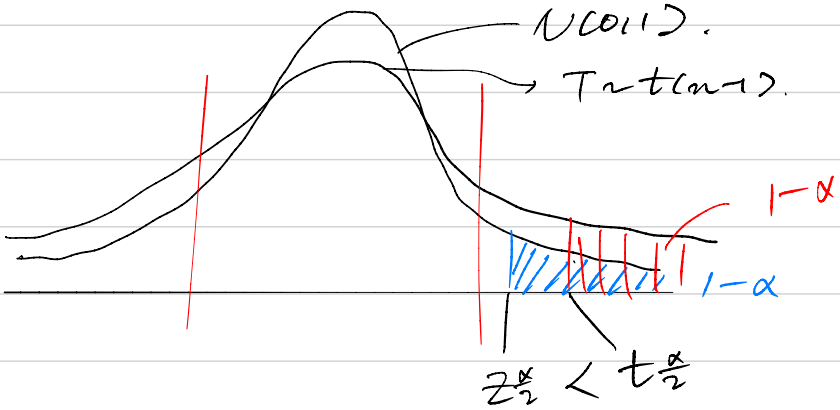
$(1-\alpha) \times 100\%$ C.I for μ

$P(-t_{\frac{\alpha}{2}} \le T \le t_{\frac{\alpha}{2}}) = 1-\alpha$.

$\Leftrightarrow P(-t_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\frac{\alpha}{2}}) = 1-\alpha$.

$\Leftrightarrow P(\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) = 1-\alpha$

2. $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ ~~vs~~ 2. $t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$



Note

We know that $t_{\alpha/2,(n-1)} > z_{\alpha/2}$ in the tail parts. If $\sigma = s$, then the length of the confidence interval $\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$ is shorter than the length of the confidence interval $\overline{x} \pm t_{\alpha/2,(n-1)}s/\sqrt{n}$.

Case 3: the population variance σ^2 known and large n

Suppose that $X_1, \ldots, X_n \sim (\mu, \sigma^2)$, where σ^2 is known and n is large enough ($n > 30$). Then, the standardized random variable $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is approximate to $\mathrm{N}(0, 1)$.

A $(1 - \alpha) \times 100\%$ confidence interval for μ with known σ^2 is given by

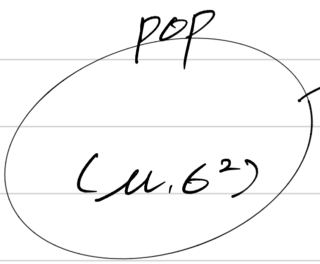
$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Case 4: the population variance σ^2 unknown and large n

Suppose that $X_1, \ldots, X_n \sim (\mu, \sigma^2)$, where n is large enough. Then, the standardized random variable $Z = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ (S^2: sample variance) is approximate to $\mathrm{N}(0, 1)$.

A $(1 - \alpha) \times 100\%$ confidence interval for μ with unknown σ^2 is given by

$$\left[\overline{X} - t_{\alpha/2,(n-1)}\frac{S}{\sqrt{n}},\ \overline{X} + t_{\alpha/2,(n-1)}\frac{S}{\sqrt{n}}\right]$$
$$\approx \left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$



$X_i \sim (\mu, \sigma^2)$. < σ^2: known, n: large.

By CLT

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow[n \to \infty]{} N(0,1).$$

$(1-\alpha) \times 100\%$ approximate for μ

$$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

$X_i \sim (\mu, \sigma^2)$. σ^2: unknown

n: large.

By CLT, approximation & replacement

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \longrightarrow N(0,1).$$

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\bar{X} - \mu}{S / \sqrt{n}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

One-sided confidence interval

Suppose that $\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$ where σ^2 is known.

A $(1-\alpha) \times 100\%$ one-sided lower confidence interval for μ is

$$P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1-\alpha$$

$$P\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu\right) = 1-\alpha$$

$$\left[\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \infty\right)$$

Similarly, a $(1-\alpha) \times 100\%$ one-sided upper confidence interval for μ is

$$\left(-\infty,\ \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

pop μ unknown fixed value → $X_1, \cdots, X_n$

$(1-\alpha)\times 100\%$ C.I for the mean μ.

1) $X_i \sim N(\mu, \sigma^2)$ σ^2: known.

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

$P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$

$\Leftrightarrow P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

2) $X_i \sim N(\mu, \sigma^2)$ σ^2: unknown.

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

$P(-t_{\frac{\alpha}{2},(n-1)} \le T \le t_{\frac{\alpha}{2},(n-1)}) = 1-\alpha$

$\Leftrightarrow P(\bar{X} - t_{\frac{\alpha}{2},(n-1)} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},(n-1)} \frac{S}{\sqrt{n}}) = 1-\alpha$

3) $X_i \sim (\mu, \sigma^2)$ σ^2: known. large n

By CLT. $\bar{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$ $(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$

4) $X_i \sim (\mu, \sigma^2)$., σ^2: unknown. large n

$[\bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$ $(\bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}})$

$(\bar{X} - t_{\frac{\alpha}{2}(n-1)} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},(n-1)} \frac{S}{\sqrt{n}})$

Ch7.2 Confidence Intervals for the Difference of Two Means

C.I for difference of two population means $\mu_X,\ \mu_Y$

Case 1: Normal with the population variances $\sigma_X^2,\ \sigma_Y^2$ known

Suppose that $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are known. Then, the difference of the sample means $(\overline{X} - \overline{Y})$ has a normal distribution,

$$\mathrm{N}\left(\mu_X - \mu_Y,\ \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

The standardized random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim \mathrm{N}(0,1).$$

$$
\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} \le \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}} \le z_{\alpha/2}\right) \\
&= P\left((\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}} \le \mu_X-\mu_Y \le (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)
\end{aligned}
$$

$(1-\alpha)\times 100\%$ confidence interval for the difference $(\mu_X-\mu_Y)$ with known $\sigma_X^2,\ \sigma_Y^2$

$$
\left[(\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},\ (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right]
$$

pop1 grade for 1st year $N(\mu_X, \sigma_X^2)$

pop2 grade for 4th year $N(\mu_Y, \sigma_Y^2)$

$\mu_X - \mu_Y$

$X_1, \cdots, X_m$ $\quad$ $Y_1, \cdots, Y_m$

eg). $\bar{X} = 3.5$ $\quad$ diff $= 0.1$ $\quad$ $\bar{Y} = 3.6$

$(1-\alpha) \cdot 100\%$ C.I for $\mu_X - \mu_Y$

pop1 — indep. — pop2

μ_X $\quad$ μ_Y

unknown fixed value

$\mu_X - \mu_Y \gtreqless 0$

$X_1, \cdots, X_m$ $\quad$ $Y_1, \cdots, Y_m$

$\bar{X}$ $\quad$ $-$ $\quad$ $\bar{Y}$

3.5 $\quad$ $-$ $\quad$ 3.7 $= -0.2$

2.5 $= 1$

1) $X_i \sim N(\mu_X, \sigma_X^2)$ $\quad$ $Y_i \sim N(\mu_Y, \sigma_Y^2)$ $\quad$ σ_X^2, σ_Y^2 : known.

$\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$ $\quad$ $\bar{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$

$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

$P(-Z_{\frac{\alpha}{2}} \le Z \le Z_{\frac{\alpha}{2}}) = 1-\alpha$

length of C.I depends on sample size as well as variance info.

Case 2: Normal with unknown σ_X^2, σ_Y^2 but $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Suppose that $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are unknown but $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Then, the difference of the sample means $(\overline{X} - \overline{Y})$ has a normal distribution, $\mathrm{N}\left(\mu_X - \mu_Y,\ \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$.

The standardized random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{1/n + 1/m}} \sim \mathrm{N}(0, 1).$$

Also, since both random variables are independent,

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{(n+m-2)}.$$

2). $X_i \sim N(\mu_X, \sigma_X^2)$ $\quad Y_i \sim N(\mu_Y, \sigma_Y^2)$

$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ unknown.

Testing. $\sigma_X^2 = \sigma_Y^2$ vs not.

$\Rightarrow \sigma_X^2 = \sigma_Y^2$

at this moment, just assume.

$\bar{X} \sim N(\mu_X, \frac{\sigma^2}{n})$, $\quad \bar{Y} \sim N(\mu_Y, \frac{\sigma^2}{m})$

$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$.

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$$

indep.

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(m-1)$$

$$U = \frac{1}{\sigma^2}\left[(n-1)S_X^2 + (m-1)S_Y^2\right] \sim \chi^2(n+m-2)$$

$$T = \frac{Z}{\sqrt{U/(n+m-2)}}$$

$$= \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \cdot \left(\frac{1}{n}+\frac{1}{m}\right)}} \sim t(n+m-2).$$

$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = S_p^2$

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2}$$

point estimator for σ^2

cf). $S_X^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$

Since the sample means and sample variances are independent,

$$T = \frac{Z}{\sqrt{U/(n+m-2)}} \sim t(n+m-2)$$

$$T = \frac{\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{1/n+1/m}}}{\sqrt{\left[\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}\right]/(n+m-2)}}$$

$$= \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\left[\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\right]\left[\frac{1}{n}+\frac{1}{m}\right]}} \sim t(n+m-2)$$

where S_p^2 is called the pooled estimator of σ^2, defined by

$$S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}.$$

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n + m - 2)$$

$$\begin{aligned} 1 - \alpha &= P(-t_0 \le T \le t_0) \\ &= P\left((\overline{X} - \overline{Y}) - t_0 S_p\sqrt{\frac{1}{n} + \frac{1}{m}} \le \mu_X - \mu_Y \le (\overline{X} - \overline{Y}) + t_0 S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right) \end{aligned}$$

where $t_0 = t_{\alpha/2}(n + m - 2)$.

$(1 - \alpha) \times 100\%$ confidence interval for the difference $(\mu_X - \mu_Y)$ with known $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

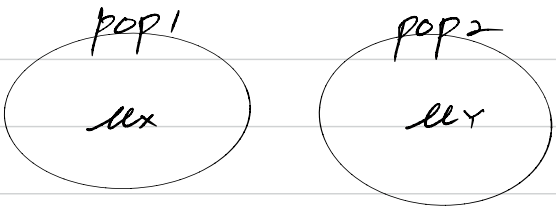
$$\left[(\overline{X} - \overline{Y}) - t_0 S_p\sqrt{\frac{1}{n} + \frac{1}{m}},\ (\overline{X} - \overline{Y}) + t_0 S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right]$$

Case 3: Unknown σ_X^2, σ_Y^2 but large $n,\ m$

Suppose that $X_1, \ldots, X_n \sim (\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim (\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are unknown but $n,\ m$ are large. Then, by CLT, the difference of the sample means $(\overline{X} - \overline{Y})$ is approximate to a normal distribution, $\mathrm{N}\left(\mu_X - \mu_Y,\ \dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}\right)$.

$(1-\alpha) \times 100\%$ confidence interval for the difference $(\mu_X - \mu_Y)$ with unknown $\sigma_X^2,\ \sigma_Y^2$ but large $n,\ m$

$$\left[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}},\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \right]$$



$X_i \sim (\mu_X, \sigma_X^2)$ $Y_j \sim (\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 : unknown.

$i = 1, \cdots, n$ $j = 1, \cdots, m$

n & m : large.

$(1-\alpha) \times 100\%$ C.I for $\mu_X - \mu_Y$

By CLT,

$\bar{X} \doteq N(\mu_X, \frac{\sigma_X^2}{n})$.
$\bar{Y} \doteq N(\mu_Y, \frac{\sigma_X^2}{m})$.
$\Rightarrow \bar{X} - \bar{Y} \doteq N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$.

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \doteq N(0,1).$$

$P(\ \leq Z \leq\ ) = 1-\alpha$

$$\Leftrightarrow P\left((\bar{X}-\bar{Y}) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}} \leq \mu_X - \mu_Y \leq (\bar{X}-\bar{Y}) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)$$

$$\Rightarrow P\left((\bar{X}-\bar{Y}) - z_{\frac{\alpha}{2}}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}} \leq \mu_X - \mu_Y \leq (\bar{X}-\bar{Y}) + z_{\frac{\alpha}{2}}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}\right).$$

Case 4 (Welch's test): Normal with unknown σ_X^2, σ_Y^2 and small $n,\ m$

Suppose that $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are unknown and $n,\ m$ are small. A $(1-\alpha)\times 100\%$ confidence interval for the difference $(\mu_X - \mu_Y)$ is given by

$$\left[(\overline{X} - \overline{Y}) - t_{\alpha/2}([r])\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}},\ (\overline{X} - \overline{Y}) + t_{\alpha/2}([r])\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \right]$$

where $[r]$ is the largest integer not greater than r defined by

$$r = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2}$$

Case 5: Paired datasets

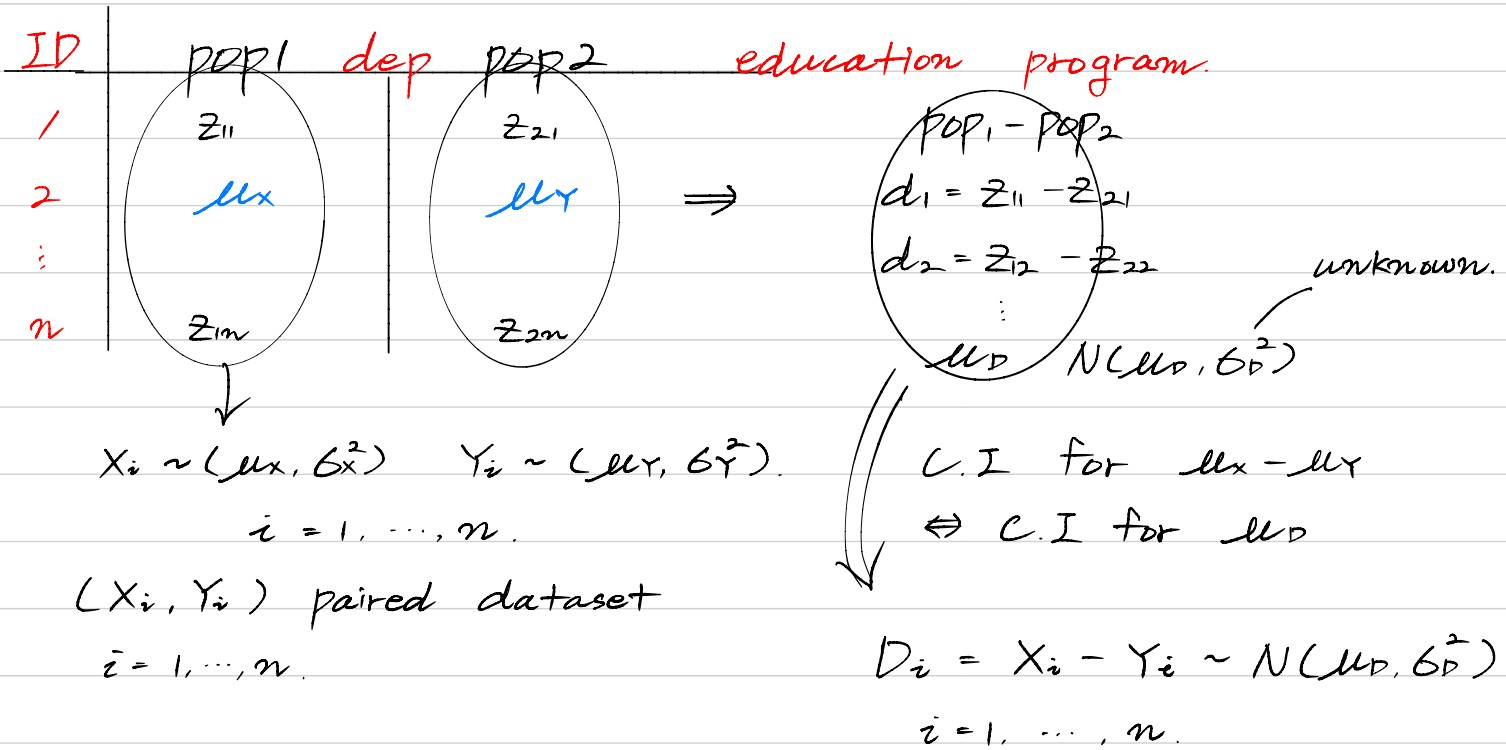
Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are paired samples and $D_i = X_i - Y_i$ are from $\mathrm{N}(\mu_D, \sigma_D^2)$, $i = 1, \ldots, n$. The confidence interval of $\mu_D = \mu_X - \mu_Y$ is

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t(n-1),$$

where $\overline{D}$ and S_D^2 are the sample mean and variane of the differences.

A $(1-\alpha) \times 100\%$ confidence interval for μ_D is given by

$$\left[\overline{D} - t_{\alpha/2}(n-1)\frac{S_d}{\sqrt{n}},\ \overline{D} + t_{\alpha/2}(n-1)\frac{S_d}{\sqrt{n}} \right]$$



1). $\bar{D} \sim N(\mu_D, \frac{\sigma_D^2}{n}) \Rightarrow Z = \frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{n}} \sim N(0,1)$.

$$U = \frac{(n-1)S_D^2}{\sigma_D^2} \sim \chi^2(n-1),$$

$$T = \frac{Z}{\sqrt{U/(n-1)}} = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t(n-1)$$

$$P\left(-t_{\frac{\alpha}{2}}(n-1) \le T \le t_{\frac{\alpha}{2}}(n-1)\right) = 1 - \alpha.$$

Ch7.3 Confidence Intervals for Proportions

Confidence Intervals for Proportions

Case 1: Single population

Let X be the number of successes with n trials, $X \sim \text{Bin}(n, p)$. By CLT,

$$\frac{X/n - p}{\sqrt{p(1-p)/n}} = \frac{X - np}{\sqrt{np(1-p)}} \to N(0,1), \quad \text{as } n \to \infty$$

$X \sim (np, np(1-p))$

$\frac{X}{n} \sim (p, \frac{p(1-p)}{n})$

$\sum X_i$

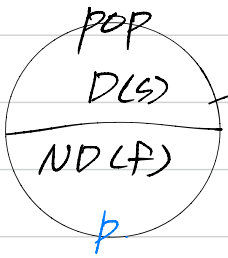
$$1 - \alpha \approx P\left(-z_{\alpha/2} \le \frac{X/n - p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\alpha/2} \right)$$

$$= P\left(\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right)$$

$\hat{P}$

$$= P\left(\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}} \le p \le \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}} \right)$$

$\hat{P}$



$X_1, \cdots, X_n \sim Ber(p)$.

p : probability of S.

$X_i = \begin{cases} 1 & \text{if } D(S) \\ 0 & \text{if } ND(f) \end{cases}$

$\hat{P} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ $\quad n = 100$ $\quad \begin{bmatrix} 40 : D \\ 60 : ND \end{bmatrix}$ $\quad \hat{P}_{MLE} = 0.4$

* $(1-\alpha) \times 100\%$ CI for p.

n large

$\sum_{i=1}^{n} X_i \sim Bin(n, p)$.

By CLT

$$Z = \frac{\sum X_i - np}{\sqrt{np(1-p)}} \dot{\sim} N(0,1)$$

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

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\frac{\sum X_i}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\frac{\sum X_i}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{\sum X_i}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

($\hat{P}$: $\frac{\sum X_i}{n}$)

이것과 approximately probability. plug in.

$$\Rightarrow P\left(\frac{\sum X_i}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \frac{\sum X_i}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\left|\frac{\frac{\sum X_i}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P(\quad \le p \le \quad) = 1-\alpha$$

approximate $(1-\alpha)\times 100\%$ confidence interval for p

$$\left[\frac{X}{n}-z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}},\ \frac{X}{n}+z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}}\right]$$

$X = \sum x_i$

Alternative way is

$$\frac{|X/n - p|}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}$$

$$\left(\frac{X}{n} - p\right)^2 - \frac{z_{\alpha/2}^2 p(1-p)}{n} \leq 0$$

$$\left(1 + \frac{z_{\alpha/2}^2}{n}\right) p^2 - \left(2\hat{p} + \frac{z_{\alpha/2}^2}{n}\right) p + \hat{p}^2 \leq 0$$

where $\hat{p} = X/n$.

An approximate $(1-\alpha) \times 100\%$ confidence interval for p is

$$\frac{\left[\hat{p} + z_{\alpha/2}^2/(2n)\right] \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}$$

One-sided confidence interval

The one-sided upper confidence interval for p is

$$\left[0,\ \frac{X}{n} + z_\alpha \sqrt{\frac{X/n(1-X/n)}{n}}\right]$$

The one-sided lower confidence interval for p is

$$\left[\frac{X}{n} - z_\alpha \sqrt{\frac{X/n(1-X/n)}{n}}, 1\right]$$

Case 2: $p_1 - p_2$ in two populations

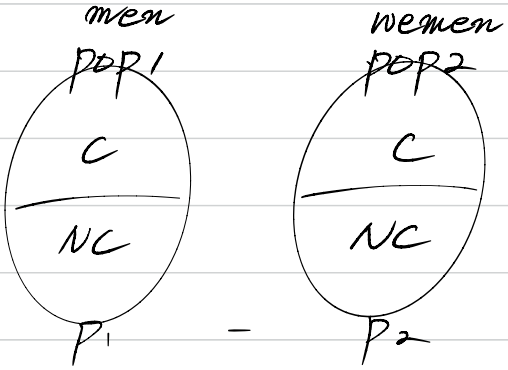
Let X be the number of successes with n_1 trials for the first experiment. Let Y be the number of successes with n_2 trials for the second experiment. They are independent. Thus, $X \sim \text{Bin}(n_1, p_1)$, $Y \sim \text{Bin}(n_2, p_2)$. As $n_1,\ n_2 \to \infty$,

$$\frac{X/n_1 - p_1}{\sqrt{p_1(1-p_1)/n_1}} \to N(0,1), \quad \frac{Y/n_2 - p_2}{\sqrt{p_2(1-p_2)/n_2}} \to N(0,1)$$

Thus the difference of both proportions is

$$\frac{X}{n_1} - \frac{Y}{n_2} \sim \text{N}\left(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

$$Z = \frac{(X/n_1 - Y/n_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim \text{N}(0,1)$$



$X_i \sim Ber(p)\ i = 1, \cdots, n$ $\quad$ $Y_j \sim Ber(p_2)\ j = 1, \cdots, m$

$\hat{p}_1 = \frac{\sum X_i}{n}$ $\quad$ $\hat{p}_2 = \frac{\sum Y_j}{m}$

$\hat{p}_1 = 0.01$ $\quad - \quad$ $\hat{p}_2 = 0.008$ $\quad = 0.002$

* $(1-\alpha) \times 100\%$ CI for $p_1 - p_2$

n & m large.

$$Z_1 = \frac{\frac{\sum X_i}{n} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \approx N(0,1) \qquad Z_2 = \frac{\frac{\sum Y_j}{m} - p_2}{\sqrt{\frac{p_2(1-p_2)}{m}}} \approx N(0,1)$$

$$\Leftrightarrow \frac{\sum X_i}{n} = \hat{p}_1 \approx N\left(p_1, \frac{p_1(1-p_1)}{n}\right) \qquad \frac{\sum Y_j}{m} = \hat{p}_2 \approx N\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right)$$

$$
\begin{aligned}
1-\alpha &\approx P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} \le \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \le z_{\alpha/2} \right) \\
&= P\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \right)
\end{aligned}
$$

where $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, $\hat{p}_1 = X/n_1$, and $\hat{p}_2 = Y/n_2$.

approximate $(1-\alpha) \times 100\%$ confidence interval for $p_1 - p_2$

$$
\left(\frac{X}{n_1} - \frac{Y}{n_2} \right) \pm z_{\alpha/2}\sqrt{\frac{(X/n_1)(1 - X/n_1)}{n_1} + \frac{(Y/n_2)(1 - Y/n_2)}{n_2}}
$$