Probability and Bayes' Review

Discretely...

Marginal Distributions $p(A), p(B)$

Joint Distributions $p(A,B) = p(A)p(B)$ if A, B are independent events.

Conditional Distributions $p(A|B), p(B|A)$

$$p(A) = \sum_B p(A,B), \quad p(B) = \sum_A p(A)$$

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A,B)}{\sum_A p(A,B)}, \quad p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)}$$

Given $p(A|B)$, $p(B)$, we can also find $p(B|A)$.

$$p(A|B) = \frac{p(A,B)}{p(B)} \Rightarrow p(A,B) = p(A|B)p(B)$$

Bayes' Rule

$$p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)} = \frac{p(A|B)p(B)}{\sum_B p(A|B)p(B)}$$

Continuously...

Joint Probability Density $p(x,y)$

Marginal Probability Density $p(x) = \int p(x,y)\,dy$, $p(y) = \int p(x,y)\,dx$

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\int p(x,y)\,dx} \qquad p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\int p(x,y)\,dy}$$

$$= \frac{p(y|x)p(x)}{\int p(y|x)p(x)\,dx} \qquad = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\,dy}$$

Bayes' Rule $\longrightarrow$ we can find $p(x|y)$ given that we know $p(y|x)$ and $p(x)$!

Maximum Likelihood Estimation

Let X be a discrete r.v. with pmf p (or pdf for continuous case) depending on parameter θ. ← parameter that we are trying to solve for our distribution.

$L(\theta|x) = p_\theta(x) = P_\theta(X=x)$ is the likelihood function, given the outcome x of the r.v. X.

For Bernoulli r.v. X,

$L(\theta|x) = p_\theta(x) = \theta^x(1-\theta)^{1-x}$ s.t. $P_\theta(X=1)=\theta$; $P_\theta(X=0) = (1-\theta)$
$= 1 - P_\theta(X=1)$

Suppose we have collected some data ... (Bernoulli coin tosses) i.i.d.

$data = \{x_1, x_2, x_3, \dots, x_N\}$

Likelihood is: $L(\theta|data) = p_\theta(data) = \prod_{i=1}^{N} p_\theta(x_i) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$

↳ Probability of observing the data that we observe assuming each coin toss was i.i.d

e.g. $L(\theta \mid x_1=1, x_2=0, x_3=1) = \theta^1(1-\theta)^1\theta^1$

We want to find the value of θ that maximizes $L(\theta|data)$!

i.e. the value of θ that makes the data that we collect most probable.

To Maximize $L(\theta|data)$, we want to take its derivative wrt. θ. s.t. $\frac{dL}{d\theta} = 0$

$\hat{\theta} = \arg\max_\theta L(\theta)$.

Note that argmax is the operation that finds the argument that gives the max value from function $L(\theta)$.

Most of the time it is better to take the log of the likelihood before differentiating,

- and it usually leads to a simpler expression for the derivative that is easier to set to 0 and solve.
- since log is a monotonous function, i.e. whatever maximizes L also maximizes $\log L$.

$l(\theta) = L(\theta) = \log \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \sum_{i=1}^{N} \{x_i \log\theta + (1-x_i)\log(1-\theta)\}$.

Set $\frac{dl}{d\theta} = \frac{1}{\theta}\sum_{i=1}^{N} x_i - \frac{1}{1-\theta}\sum_{i=1}^{N}(1-x_i) = 0$

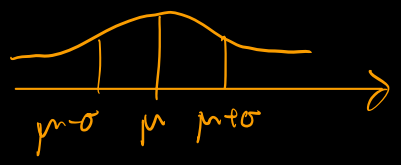
Solve for θ $(1-\theta)\sum_{i=1}^{N} x_i = \theta\sum_{i=1}^{N}(1-x_i)$.

$\sum_{i=1}^{N} x_i - \theta\sum_{i=1}^{N} x_i = N\theta - \theta\sum_{i=1}^{N} x_i$ $\Rightarrow$ $\theta = \frac{1}{N}\sum_{i=1}^{N} x_i$

Note that for Binomial dist., $P(x=k) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$

For Gaussian (Normal) r.v.'s

$\text{data} = \{x_1, x_2, \dots, x_N\}$ i.i.d.



$$L(\theta \mid \text{data}) = \prod_{i=1}^{N} p_\theta(x_i), \quad \text{where } \theta = \{\mu, \sigma^2\}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

Goal: Find parameters that best describe data collected.
These values will maximizes the likelihood.

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} L(\mu, \sigma^2 \mid \text{data})$$

$$l(\mu, \sigma^2) = \log L(\mu, \sigma^2) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \sum_{i=1}^{N}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]$$

$$= \sum_{i=1}^{N}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] = \underbrace{-\frac{1}{2}\log(2\pi\sigma^2)\sum_{i=1}^{N}(1)}_{\text{constant}} - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Set $\frac{\partial L}{\partial \mu} = 0$, $\quad \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0.$

$$\sum_{i=1}^{N} x_i - N\mu = 0.$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i = f(x_1, x_2, \dots, x_N).$$

$$l(\mu,\sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$l(\mu, v) = -\frac{N}{2}\log(2\pi v) - \frac{1}{2}\frac{1}{v}\sum_{i=1}^{N}(x_i-\mu)^2, \quad \text{where } v=\sigma^2$$

$$\frac{\partial l}{\partial v} = -\frac{N}{2}\frac{1}{v} - \frac{1}{2}(-1)\frac{1}{v^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

Set $\frac{\partial l}{\partial v} = 0$, $N = \frac{1}{v}\sum_{i=1}^{N}(x_i-\mu)^2$

$$v = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 = g(x_1, x_2, \dots, x_N).$$

Our MLE estimates are also random variables!

If they are random variables, then we can ask:

- What is their distribution and expectation?

$\hat{\mu} \sim ?$ $\quad \mathbb{E}(\hat{\mu}) = ?$

$\hat{\sigma}^2 \sim ?$ $\quad \mathbb{E}(\hat{\sigma}^2) = ?$

Can use characteristics functions to prove:

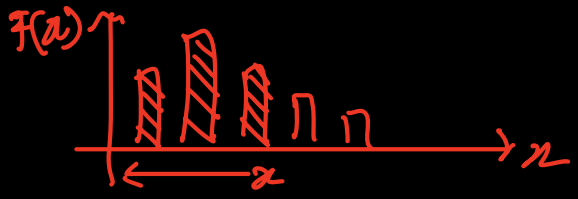
$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\mathbb{E}(\hat{\mu}) = \mu$$

$\mathbb{E}(\hat{\sigma}^2) \neq \sigma^2$; $\mathbb{E}(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 \to \sigma^2$ as $N \to \infty$.

See 'unbiased estimate of the covariance matrix' for full derivation

CDF and Percentiles



for Discrete rvs:

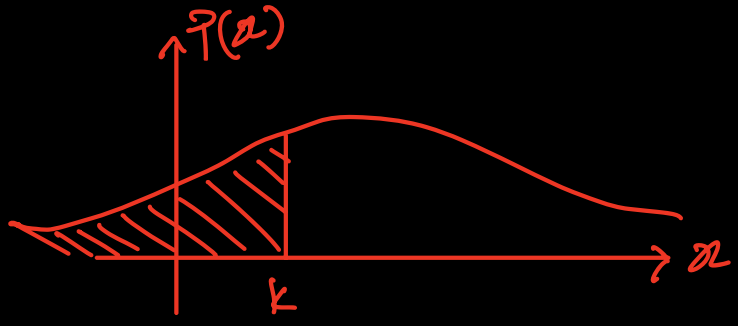
CDF: $F(x) = P(X \leq x) = \sum_{k=-\infty}^{x} p(k)$, where $p(k) = \text{Prob}(X = k)$

for Continuous rvs:

CDF: $F(k) = P(X \leq k) = \int_{-\infty}^{k} f(t)\,dt$

t is a dummy variable, and disappears after integration.

PDF: $f(x) = \frac{dF(x)}{dx}$



Inverse CDF (percentile function)

$F^{-1}(p)$ — What value of x would yield a CDF value of p?

Eg. Heights — $\mu = 170$ cm, 7cm

$F^{-1}(0.95, \mu = 170, \sigma = 7) = 181.5$

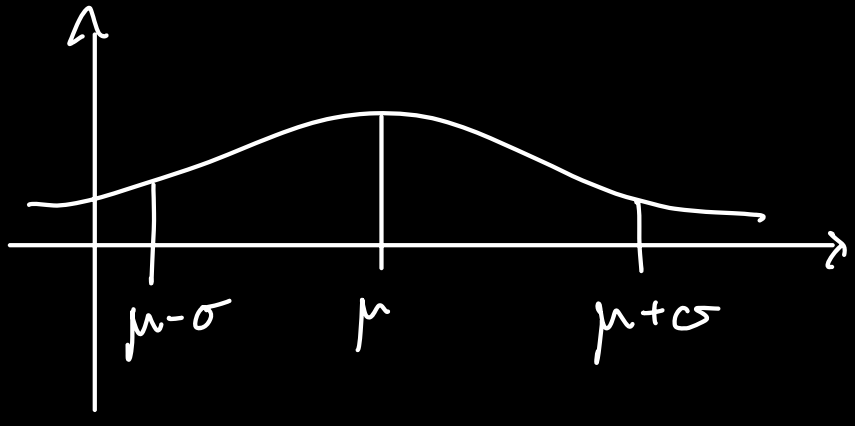
$F(160, \mu = 170, \sigma = 7) = 0.08$

Traditional A/B Testing.

all about using point estimates to model a population's distribution

Using height as an example, suppose that height has a normal distribution,

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



Note that:

In probability theory, CLT establishes that in many situations for i.i.d samples, the standardized sample mean tends towards the standard normal distribution even if the original variables are not normally distributed.

We can say that we are more confident that the mean of our sample is representative of the entire population if:

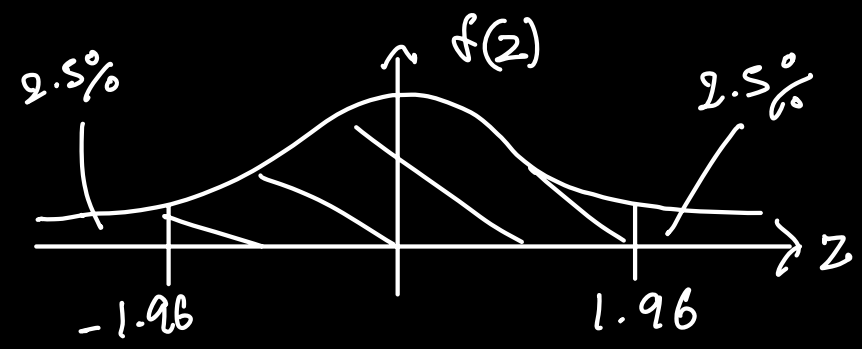
$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$
$$\left(E(\hat{\mu}),\ Var(\hat{\mu})\right)$$

1. we have a large sample
2. the sample variance is small

In statistics, we typically want to standardize our data so we can easily calculate the probability of certain values occurring in our distribution; or to compare data sets with different means and standard deviations.

i.e. $x \sim N(\mu, \sigma^2) \rightarrow z = \frac{x-\mu}{\sigma} \sim N(0,1)$

For a 95% confidence interval, the lower and upper end points are -1.96 and 1.96.



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

standardizing a variable.

$$\frac{x-\mu}{\sigma} \sim N(0,1).$$

If $X \sim N(\mu, \sigma^2)$, then $\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$

$$Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{N}}$$

$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$

$$E(\hat{\mu}) = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\left[\sum_{i=1}^{N} E(x_i)\right] = \mu$$

$$Var(\hat{\mu}) = E\left[(\hat{\mu} - E(\hat{\mu}))^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right]$$

$$= \frac{1}{N^2} E\left[\left(\sum_{i=1}^{N} x_i - N\mu\right)^2\right] = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N}$$

Note that $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ is the unbiased estimator for sample variance (which we should use to replace σ^2 that is unknown)

for 95% C.I.,

$$-1.96 \leq Z \leq 1.96$$

$$-1.96 \leq \frac{\hat{\mu}-\mu}{\sigma/\sqrt{N}} \leq 1.96$$

$$-1.96\left(\frac{\sigma}{\sqrt{N}}\right) - \hat{\mu} \leq -\mu \leq 1.96\left(\frac{\sigma}{\sqrt{N}}\right) - \hat{\mu}$$

$$\hat{\mu} - 1.96\left(\frac{\sigma}{\sqrt{N}}\right) \leq \mu \leq \hat{\mu} + 1.96\left(\frac{\sigma}{\sqrt{N}}\right)$$

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{N}} \leq \mu \leq \hat{\mu} + 1.96\frac{\sigma}{\sqrt{N}}$$

replace with $\hat{\sigma}$.

Essentially,

0.95 is the no. of times μ is contained in the 95% CI out of the total no. of experiments

For a γ% CI, $\left[\hat{\mu} + \Phi^{-1}\left(\frac{1-\gamma}{2}\right)\frac{\sigma}{\sqrt{N}},\ \hat{\mu} + \Phi^{-1}\left(1 - \frac{1-\gamma}{2}\right)\frac{\sigma}{\sqrt{N}}\right]$

However, we don't know the true variance σ^2.

Consider a new standardized r.v.
(which is explicitly divided by the estimated sd)

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{N}} = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{N}} \times \frac{1/\sigma}{1/\sigma} = \left(\frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}}\right) \Big/ \left(\frac{\hat{\sigma}}{\sigma}\right)$$

Consider denominator:

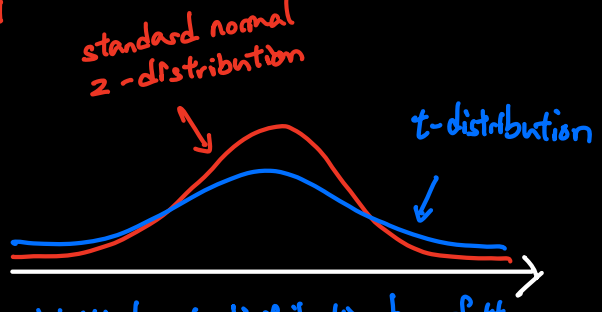
$$\left(\frac{\hat{\sigma}}{\sigma}\right)^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma}\right)^2$$

$$(N-1)\left(\frac{\hat{\sigma}}{\sigma}\right)^2 \sim \chi^2_{N-1}$$ chi-square distribution with $N-1$ df

Note: $z_i \sim N(0,1) \longrightarrow \sum_{i=1}^{N} (z_i - \bar{z})^2 \sim \chi^2_{N-1}$

$$t = \frac{z}{\sqrt{V/v}} \sim t_v$$

$$\frac{\text{standard normal}}{\sqrt{\text{chi-square} / \text{deg of freedom}}} \sim t_{\text{deg of freedom}}$$



intuitively, t-distribution has fatter tails since we don't know σ, so we expect the confidence interval to be fatter

t-distribution is parameterized by $(N-1)$ degrees of freedom!

$t_{left} = F^{-1}(0.025;\ df = N-1)$; $t_{right} = F^{-1}(0.975;\ df = N-1)$.

$t_{left} \le t \le t_{right}$

$$\hat{\mu} + t_{left}\frac{\hat{\sigma}}{\sqrt{N}} \le \mu \le \hat{\mu} + t_{right}\frac{\hat{\sigma}}{\sqrt{N}}$$

For large values of N, t-values $\approx$ z-values

As $N \to \infty$, $\hat{\sigma} \to \sigma \Rightarrow t \to$ Normal

Z-statistic

$$Z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}}, \text{ where } \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \; ; \; \hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

and σ is replaced with estimator of σ, $\hat{\sigma}$

$$Z_{left} = \Phi^{-1}(0.025) \; ; \; Z_{right} = \Phi^{-1}(0.975)$$

$$lower = \hat{\mu} + Z_{left} \cdot \frac{\hat{\sigma}}{\sqrt{N}} \; ; \; upper = \hat{\mu} + Z_{right} \cdot \frac{\hat{\sigma}}{\sqrt{N}}$$

t-statistic

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{N}} = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{N}} \times \left(\frac{1/\sigma}{1/\sigma}\right) = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \left(\frac{\hat{\sigma}}{\sigma}\right)$$

$$\Longrightarrow t = \frac{Z}{\sqrt{V/r}} \sim t_r \text{, where } r = N-1 \text{ degrees of freedom}$$

$$t_{left} = \Phi^{-1}(0.025, r = N-1) \; ; \; t_{right} = \Phi^{-1}(0.975, r = N-1)$$

$$lower = \hat{\mu} + t_{left} \cdot \frac{\hat{\sigma}}{\sqrt{N}} \; ; \; upper = \hat{\mu} + t_{right} \cdot \frac{\hat{\sigma}}{\sqrt{N}}$$

Hypothesis Testing.

- Groups of Data:

  1 → 1-Sample Test e.g. Avg. daily stock return

  2 → 2-Sample Test Control vs Treatment Group e.g. Drug efficacy

- 1-sided vs 2-sided test

Consider 2-sample test for new drug.

(2-sided test)

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

$\leftrightarrow$

$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

(1-sided test)

$H_0: \mu_1 \leq \mu_2$
$H_1: \mu_1 > \mu_2$

→ Output:

Test Statistic (does it fall in the rejection region?)

✡ P-value

tells us whether the difference is statistically significant (given a significance threshold / level of significance)

Probability to reject the null when the null is true.

How likely that the data observed is to have occurred under the Null Hypothesis.

→ Either reject the null hypothesis,
or fail to reject the null hypothesis

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

Consider 1-Sample Test:

For 2-sided test: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$

Rewrite the null hypothesis

$$H_0: \hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{N}\right) \longleftrightarrow H_0: z = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}} \sim N(0,1)$$

To calculate p-value,

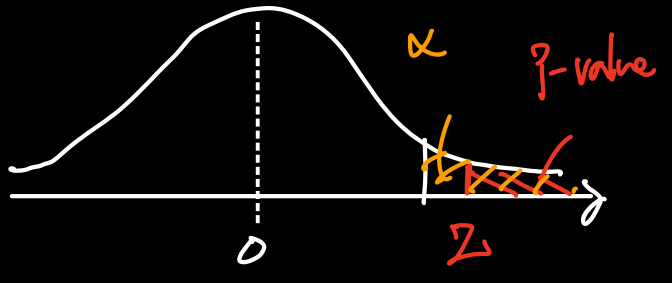
$$p_{left} = \Phi(-|z|) \; ; \; p_{right} = 1 - \Phi(|z|)$$

$$p = p_{left} + p_{right} = p_{left} \times 2 \quad \text{(since distribution is symmetric)}$$

For 1-sided test: $\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$

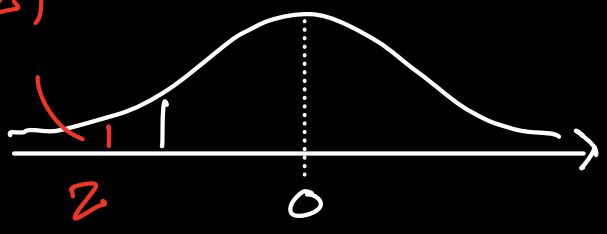
(for the greater than hypothesis)

$$p = 1 - \Phi(z)$$



Similarly, for the lesser than hypothesis,

$$p = \Phi(z)$$



For 2-sample test,

For 2-sided test : $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$

Let $y = \mu_1 - \mu_2$, $\begin{cases} H_0: y = 0 \\ H_1: y \neq 0 \end{cases}$

$$\hat{y} = \hat{\mu}_1 - \hat{\mu}_2 = \frac{\sum_{i=1}^{N_1} x_i}{N_1} - \frac{\sum_{i=1}^{N_2} x_i'}{N_2}$$

$$\begin{aligned} Var(\hat{y}) &= Var\left(\frac{x_1}{N_1}\right) + \dots + Var\left(\frac{x_{N_1}}{N_1}\right) + Var\left(\frac{x_1'}{N_2}\right) + Var\left(\frac{x_{N_2}'}{N_2}\right) \\ &= \frac{\sigma_1^2}{N_1^2} + \dots + \frac{\sigma_1^2}{N_1^2} + \frac{\sigma_2^2}{N_2^2} + \dots + \frac{\sigma_2^2}{N_2^2} \\ &= \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \end{aligned}$$

$$\sigma_{\hat{Y}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

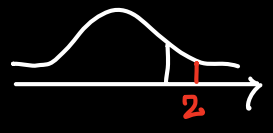
$$Z = \frac{\hat{y} - \mu_0}{\sigma_{\hat{Y}}} = \frac{\hat{y} - \mu_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

For P-value, 2-sided test:

$P_{right} = 1 - \Phi(|Z|)$; $P_{left} = \Phi(-|Z|)$; $P = P_{left} + P_{right}$

1-sided test:

$(>)$: $p = 1 - \Phi(Z)$



$(<)$: $p = \Phi(Z)$