Two-stage designs

for experiments with a large number of hypotheses

2022-08-19 lab seminar

Hwang Seo-hwa

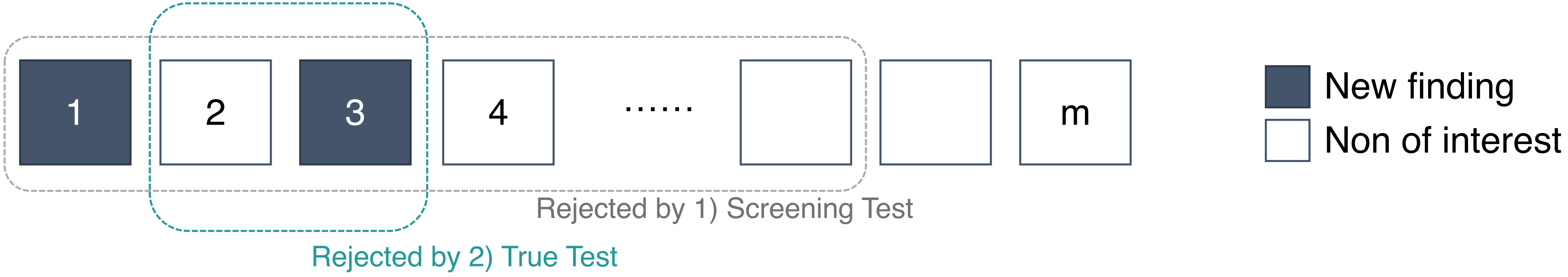
contents

- Review of two-stage designs
  - Concept of two-stage design
  - Two-stage design with iid assumption
- Research progress
  - Summary of last seminar
  - Ongoing progress
  - Plan

Concept of two-stage design

$$H_0 : z_{1i} = z_{10} \quad or \; z_{2i} = z_{20} \qquad vs. \qquad H_1 : z_{1i} > z_{10} \;\&\; z_{2i} > z_{20}$$

1) $H_{01i}: z_{1i} = z_{10} \quad vs. \quad H_{11i}: z_{1i} > z_{10}$
2) $H_{02i}: z_{2i} = z_{20} \quad vs. \quad H_{21i}: z_{2i} > z_{20}$



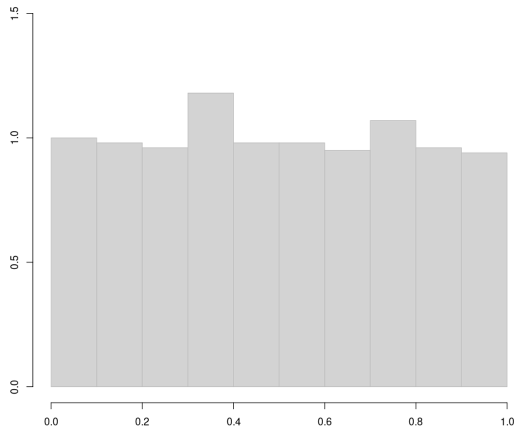
What if Statistics for 1) and 2) are Not independent?

Concept of two-stage design

- $Assume\ all\ statistics\ are\ generated\ by\ null$
- $z_{1i} = z_{2i}\ ;\ cor(z_1, z_2) = 1$

1) H_{01i}: $z_{1i} = z_{10}$ $vs.$ H_{11i}: $z_{1i} > z_{10}$

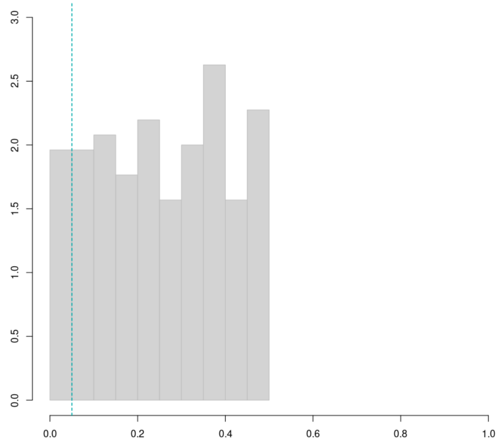
Compute p val & reject hypotheses s.t. p val<0.5



P val ~ U(0,1), under H_{01}

2) H_{02i}: $z_{2i} = z_{20}$ $vs.$ H_{12i}: $z_{2i} > z_{20}$

Compute p val & reject hypotheses s.t. passing H_{01i} & p val<0.05



P val $\not\sim$ U(0,1), under H_{01}

Prob) cannot control FDR ➡ Sol) { Make 1) and 2) indep. / Consider cor. Between z_1 & z_2 }

Two-stage design with i.i.d assumption

Idea) split samples for testing H_{01} and H_{02}

Total sample size = n

n·r = n_1	n·(1-r) = n_2

1) $H_{01i}\colon z_{1i} = z_{10} \quad vs. \quad H_{11i}\colon z_{1i} > z_{10}$

- Compute p val using n_1 samples & reject hypotheses s.t. p val<0.5

2) $H_{02i}\colon z_{2i} = z_{20} \quad vs. \quad H_{12i}\colon z_{2i} > z_{20}$

- Redefine p val ~ U(0,1) under H_0
- Compute p val using n_2 samples & reject hypotheses s.t. passing H_{01i} & p val<0.05

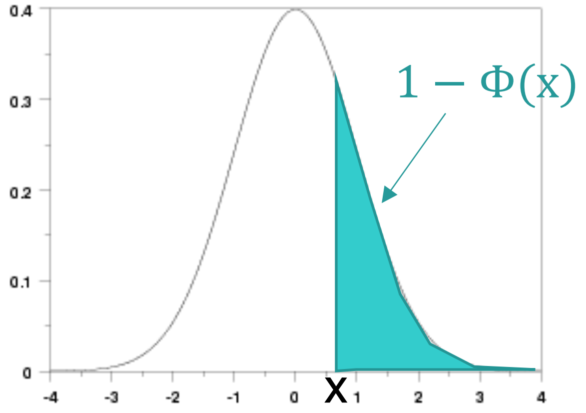
Two-stage design with i.i.d assumption : Notation & setting

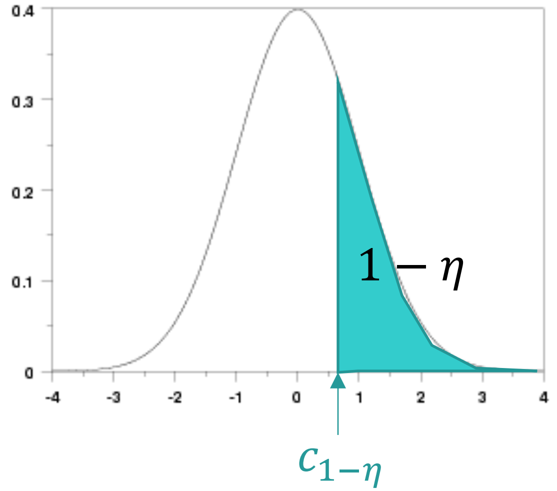
- Notation
  - n: total number of samples
    - $n_1 : n \times r$
    - $n_2 : n \times (1-r)$
  - m : # of hypothesis
  - z_{1i}, z_{2i} : test statistics
  - p_{1i} : p $value$ $computed$ at $stage$ $1, using$ n_1 $samples$
  - p_i : p $value$ $computed$ at $stage$ 2
  - γ_1, γ_2 : pre-defined cutoff value for rejection of H_0, H_1

- Setting:
  - $z_{1i}, z_{2i} \sim iid\ N(0,1)$; or σ^2 is known
  - $n \gg m$: sample size is much larger than # of hypotheses
  - One-sided test

Two-stage design with i.i.d assumption : Notation & setting

- ϕ: pdf of N(0,1)
- Φ: cdf of N(0,1)





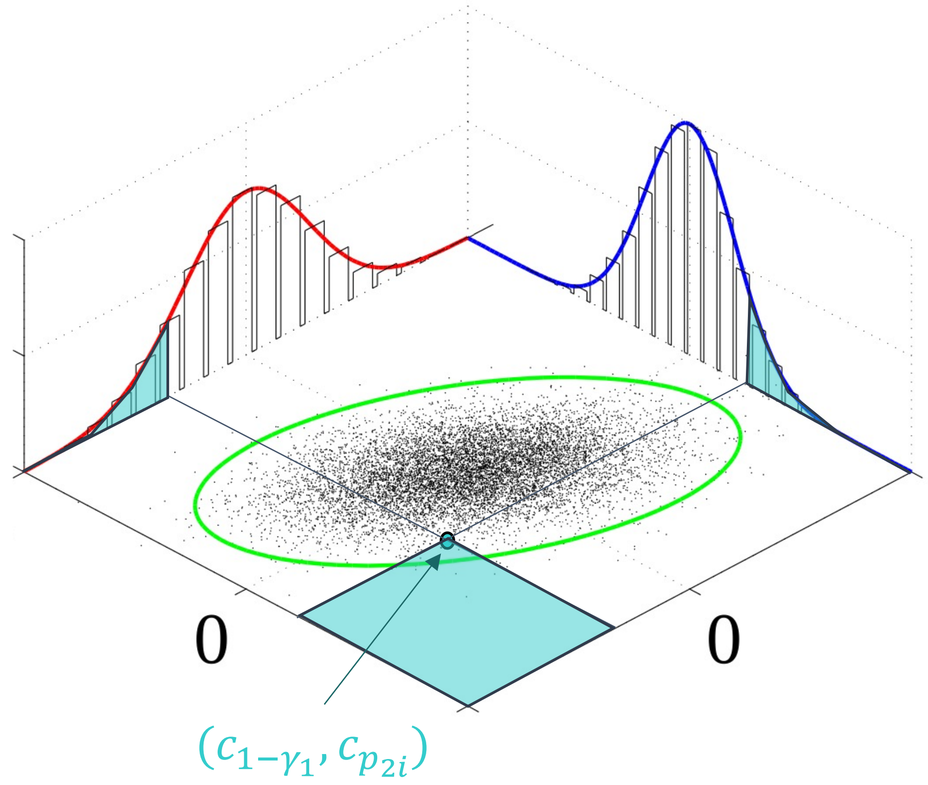
- $\Phi(\mathrm{z}_1), \Phi(z_2) \sim U(0,1), under\ H_0$

  $\Rightarrow 1-\Phi(\mathrm{z}_1), 1-\Phi(z_2) \sim U(0,1), under\ H_0$
- $(\mathrm{z}_1, z_2) \sim N((0,0), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})$
- $\varphi_{z_1}: pdf\ of\ z_2|z_1$

Two-stage design with i.i.d assumption : Redefine $\mathrm{p_i}$

- p_{1i} : p $value$ $computed$ at $stage$ $1, using$ n_1 $samples$
- p_i : p $value$ $computed$ at $stage$ 2
- γ_1, γ_2 : pre-defined cutoff value for rejection of H_0, H_1

- $\mathrm{p_{1i}} = 1 - \Phi(\mathrm{z_{1i}})$
- $\mathrm{p_{2i}} = 1 - \Phi(\mathrm{z_{2i}})$
- $\gamma = \mathrm{P}(\mathrm{p_{1i}} \leq \gamma_1, \mathrm{p_{2i}} \leq \gamma_2)$
- $\mathrm{p_i} = \begin{cases} \mathrm{p_{1i}}, \ \text{if} \ \mathrm{p_{1i}} > \gamma_1 \\ \int_{c_{1-\gamma_1}}^{\infty} \int_{c_{p_{2i}}}^{\infty} \varphi_{\mathrm{z_1}}(\mathrm{z_2})\, \phi_1(\mathrm{z_1}) \mathrm{d}\mathrm{z_2} \mathrm{d}\mathrm{z_1} \end{cases}$



Two-stage design with i.i.d assumption : Redefine $\mathrm{p_i}$

1) $P_{H_o}(p_i < \gamma) = \gamma \quad (i.e.\ p_i \sim U(0,1)\ under\ H_0)$

$$\mathrm{P}(\mathrm{p_i} \le \gamma) = \text{①}\mathrm{P}(\mathrm{p_i} \le \gamma, \mathrm{p_{1i}} > \gamma_1) + \text{②}\mathrm{P}(\mathrm{p_i} \le \gamma, \mathrm{p_{1i}} \le \gamma_1)$$

Case1: $\gamma \le \gamma_1$

- ①$\mathrm{P}(\mathrm{p_i} \le \gamma,\ \mathrm{p_{1i}} > \gamma_1) = \mathrm{P}(\mathrm{p_{1i}} \le \gamma,\ \mathrm{p_{1i}} > \gamma_1) = 0$
- ②$\mathrm{P}(\mathrm{p_i} \le \gamma,\ \mathrm{p_{1i}} \le \gamma_1) = \mathrm{P}(\mathrm{p_{2i}} \le \gamma_2,\ \mathrm{p_{1i}} \le \gamma_1) = \gamma$ (by def)

∴ ① + ② $= \gamma$

Case2: $\gamma > \gamma_1$

- ①$\mathrm{P}(\mathrm{p_i} \le \gamma,\ \mathrm{p_{1i}} > \gamma_1) = \mathrm{P}(\gamma_1 < \mathrm{p_{1i}} \le \gamma) = \gamma - \gamma_1$
- ② when $\mathrm{p_{1i}} \le \gamma_1 < \gamma$,
  - $\mathrm{p_i} = \int_{\mathrm{W}_{1-\gamma_1}}^{\infty} \left\{\int_{-\infty}^{\infty} \varphi_{\mathrm{z_1}}(\mathrm{z_2}) \left(\mathrm{I}_{\left(-\infty, \mathrm{V}_{\frac{\mathrm{z_{2i}}}{2}}\right)} + \mathrm{I}_{\left(\mathrm{V}_{1-\frac{\mathrm{z_{2i}}}{2}}, \infty\right)}\right)\right\} \phi_1(\mathrm{z_1})\mathrm{dz_2dz_1}$

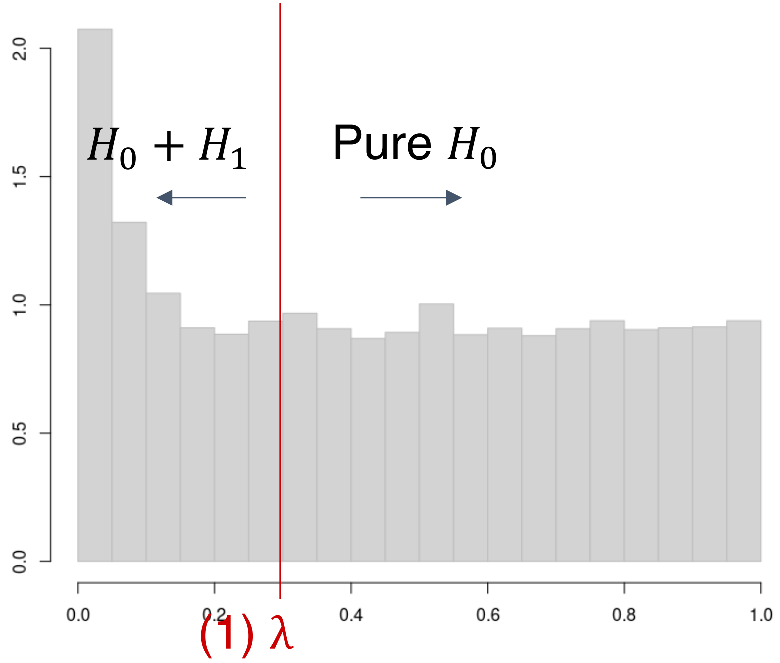
    $\le \int_{\mathrm{W}_{1-\gamma_1}}^{\infty} \{1\} \phi_1(\mathrm{z_1})\mathrm{dz_2dz_1} = \gamma_1 < \gamma$ (always true)

  ∴ ②$\mathrm{P}(\mathrm{p_i} \le \gamma,\ \mathrm{p_{1i}} \le \gamma_1) = \mathrm{P}(\mathrm{p_{1i}} \le \gamma_1) = \gamma_1$

∴ ① + ② $= \gamma$

Two-stage design with i.i.d assumption : Redefine $\mathrm{p_i}$

2) $\mathrm{p_i}$ controls FDR



(2) $\widehat{\pi_0} = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)}$

(3) $FDP = \frac{\#\{rejected\ null\}}{\#\{rejected\ \ hypotheses\}}$

$$= \frac{\widehat{\pi_0} \times m \times \gamma}{\#\{p_i < \gamma\}} \leq \alpha$$

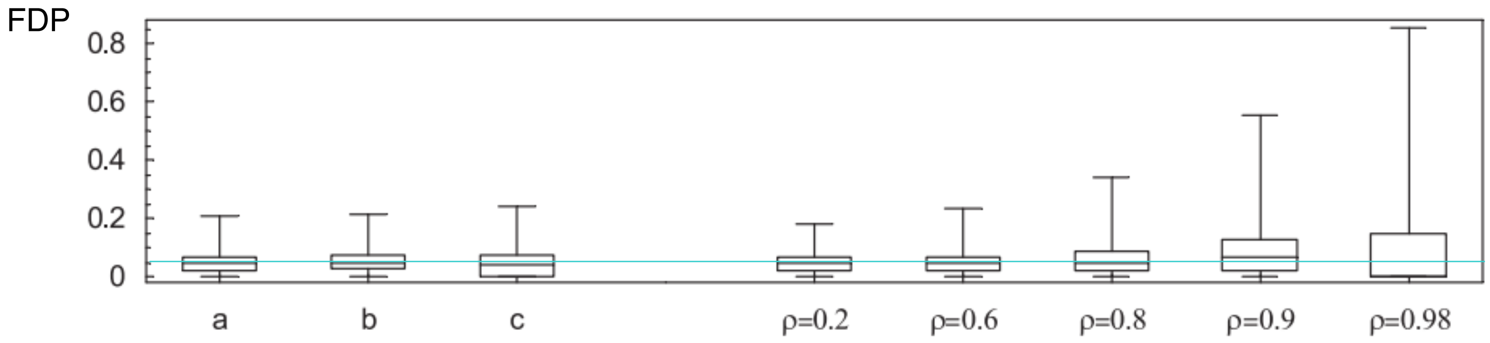
$\rightarrow Find\ largest\ \gamma\ satisfying\ above\ inequality$

Two-stage design with i.i.d assumption : Simulation

Setting (Optimal setting)

- n = 40,000 / m=5000
- $\pi_0 = 0.99,\ \alpha = 0.05,\ \lambda = 0.5,\ r = 0.625,\ \gamma_1 = 0.123,$ mean of alternative dist. = 1 sd of null dist.

		γ	FDP	power
a	Known variance	0.00045 (0.000032)	0.0488 (0.033)	0.848 (0.051)
b	Unknown variance	0.00041 (0.000037)	0.0493 (0.034)	0.774 (0.061)
c	Distributed mean	0.00028 (0.000042)	0.0497 (0.042)	0.523 (0.074)
	$\rho = 0.20$	0.00045 (0.000032)	0.0487 (0.033)	0.848 (0.051)
	$\rho = 0.60$	0.00045 (0.000034)	0.0491 (0.036)	0.848 (0.051)
	$\rho = 0.80$	0.00046 (0.000041)	0.0581 (0.050)	0.849 (0.052)
	$\rho = 0.90$	0.00048 (0.000063)	0.0860 (0.086)	0.851 (0.052)
	$\rho = 0.98$	0.00052 (0.00018)	0.0968 (0.154)	0.851 (0.061)



High correlation
->fail to control FDR

contents

- Review of two-stage designs
  - Concept of two-stage design
  - Two-stage design with iid assumption
- Research progress
  - Summary of last seminar
  - Ongoing progress
  - Plan

Review of previous presentation

- Research question
  - which gene is regulated by SET4(gene regulator)?

- Experimental design:
  - Compare the gene expression rate between WT (control group) and KO (case group; SET4 gene is removed)

1) $H_{01}: \quad \mathrm{X}_1 = \min\left(\frac{\mu_{KO}}{\sigma_{KO}}, \frac{\mu_{WT}}{\sigma_{WT}}\right) \sim f_0 \qquad vs. \qquad H_{11}: X_1 \sim f_1$

2) $H_{02}: \quad X_2 = \ log\frac{\mu_{KO}}{\mu_{WT}} \sim g_0 \qquad vs. \qquad H_{11}: \quad \mathrm{X}_2 \sim g_1$

Review of previous presentation

1) H_{01} : $X_1 = \min\left(\frac{\mu_{KO}}{\sigma_{KO}}, \frac{\mu_{WT}}{\sigma_{WT}}\right) \sim f_0$ vs. $H_{11} : X_1 \sim f_1$

2) H_{02} : $X_2 = \log\frac{\mu_{KO}}{\mu_{WT}} \sim g_0$ vs. H_{11} : $X_2 \sim g_1$

- Plan

Model: $f(X_1) = \pi_0 f_0 + (1 - \pi_0) f_1$

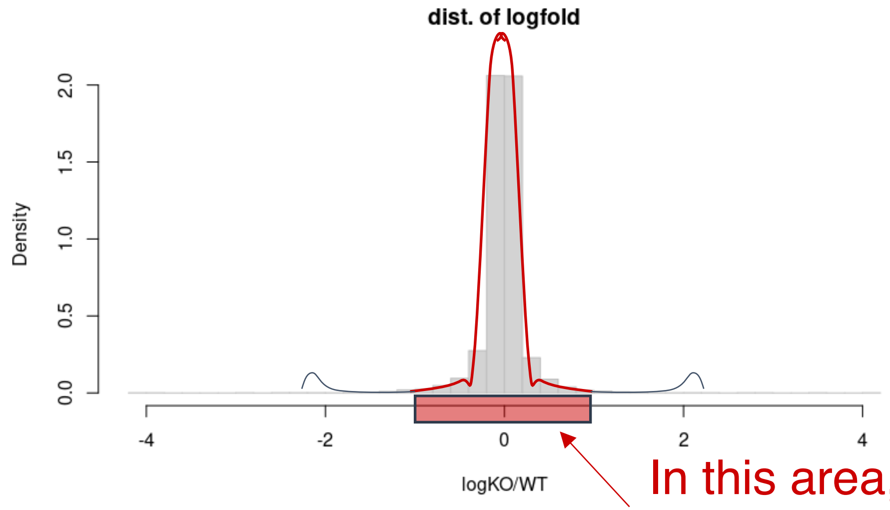
Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

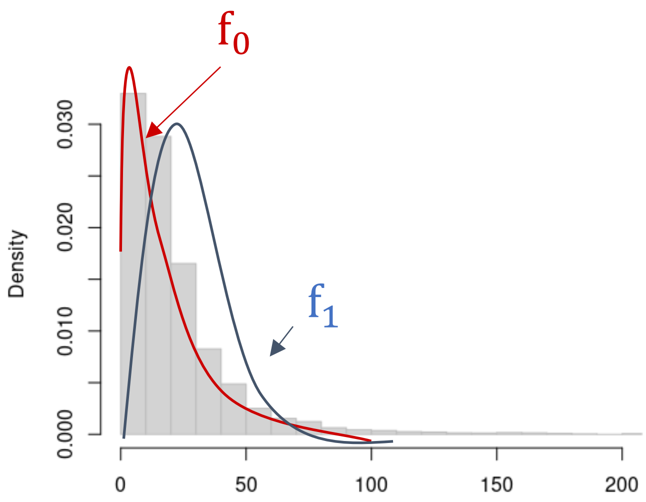
Problem)

In general



In this area, $f(x) \cong \pi_0 f_0(x)$

In this case,

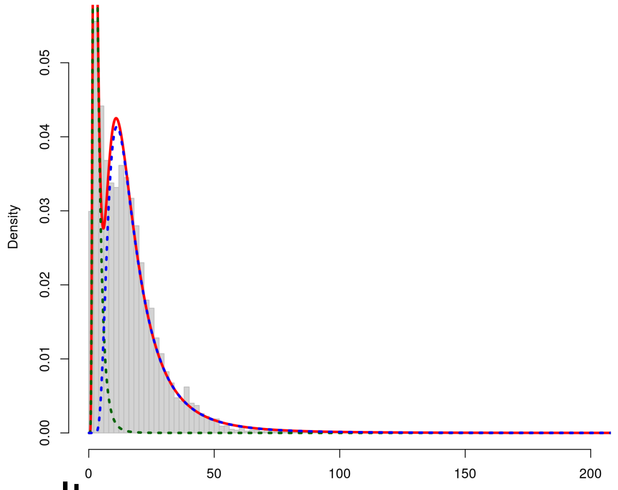


Cannot apply zero assumption

Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

Sol) add assumption that f_1 ~inv.gamma(α, β)



$x_i = data\ of\ i^{th}\ \min(snr_{KO}, snr_{WT})$

$\tau_i = P(ith\ person \in f_0)$

- $\hat{\beta}^{t+1} = \frac{\hat{\alpha}^t \sum_i \tau_i^t}{\sum_i \frac{\tau_i^t}{x_i}}$
- $\hat{\alpha}^{t+1} = \hat{\alpha}^t - \frac{g(\hat{\alpha}^t)}{g'(\hat{\alpha}^t)}$
- Where $g(a) = \psi(a) \sum_i \tau_i^t + \sum_i \tau_i^t \log x_i - \sum_i \tau_i^t * \log \frac{\alpha \sum_i \tau_i^t}{\sum_i \frac{\tau_i^t}{x_i}}$
- $\tau^{t+1} = \frac{\sum_i \tau_i^t}{n}$
- $\tau_i^{t+1} = \frac{\tau^{t+1} f_0(\hat{\alpha}^{t+1}, \hat{\beta}^{t+1})}{\tau^{t+1} f_0 + (1-\tau^{t+1}) f_1(\hat{\alpha}^{t+1}, \hat{\beta}^{t+1})}$

Result:

alpha	beta	1-pi
3.42	49.69	0.73

Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

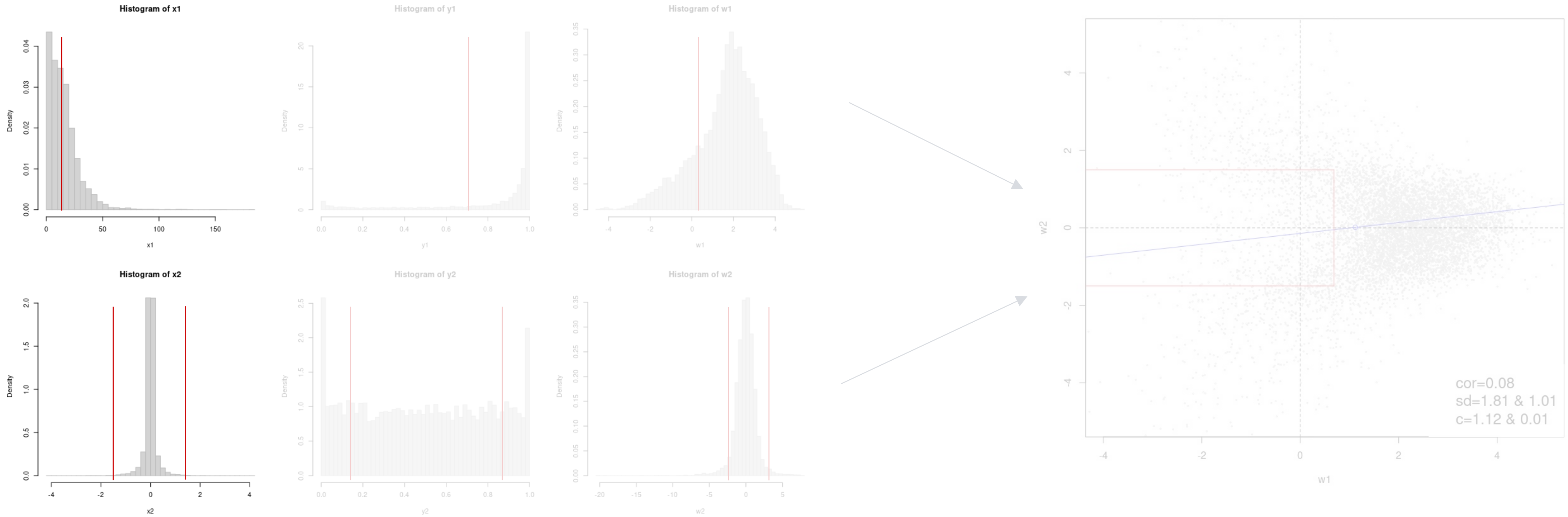
Original data

Cdf(X) ~U(0,1)

$\Phi^{-1}(F_0(X_1)), \Phi^{-1}(G_0(X_2)) \sim N(0,1)$

X_1

X_2



Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

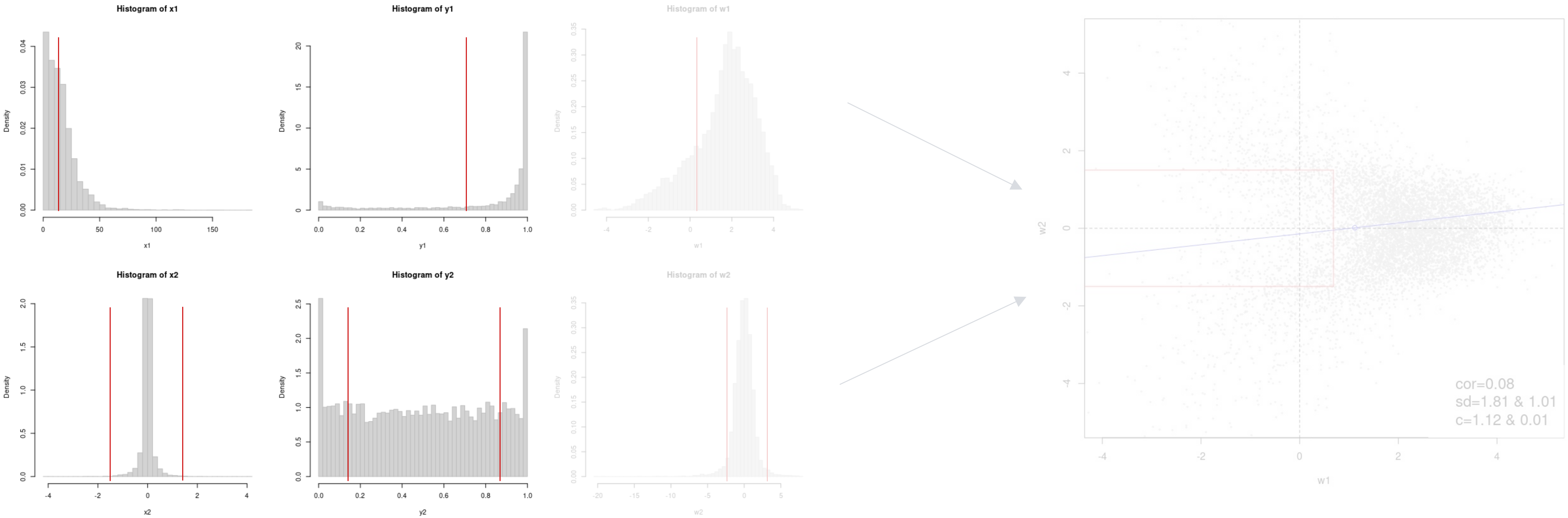
Original data

Cdf(X) ~U(0,1)

$\Phi^{-1}(F_0(X_1)), \Phi^{-1}(G_0(X_2)) \sim N(0,1)$

X_1

X_2



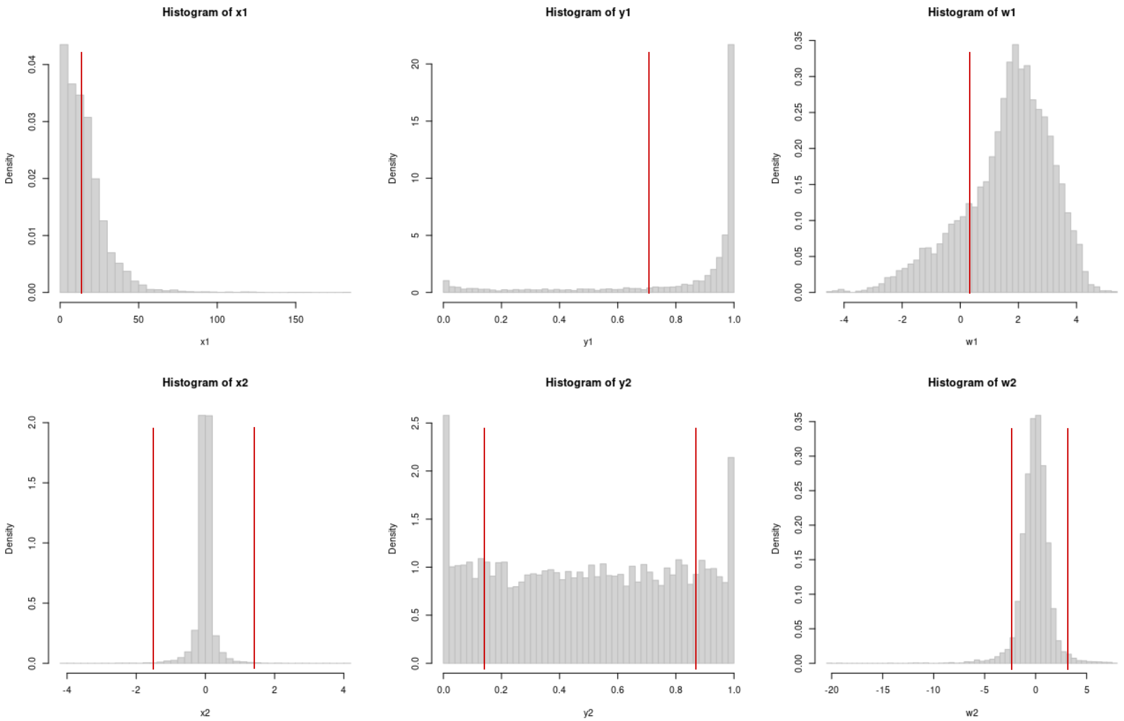
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

Original data | Cdf(X) ~U(0,1) | $\Phi^{-1}(F_0(X_1)), \Phi^{-1}(G_0(X_2)) \sim N(0,1)$

X_1

X_2



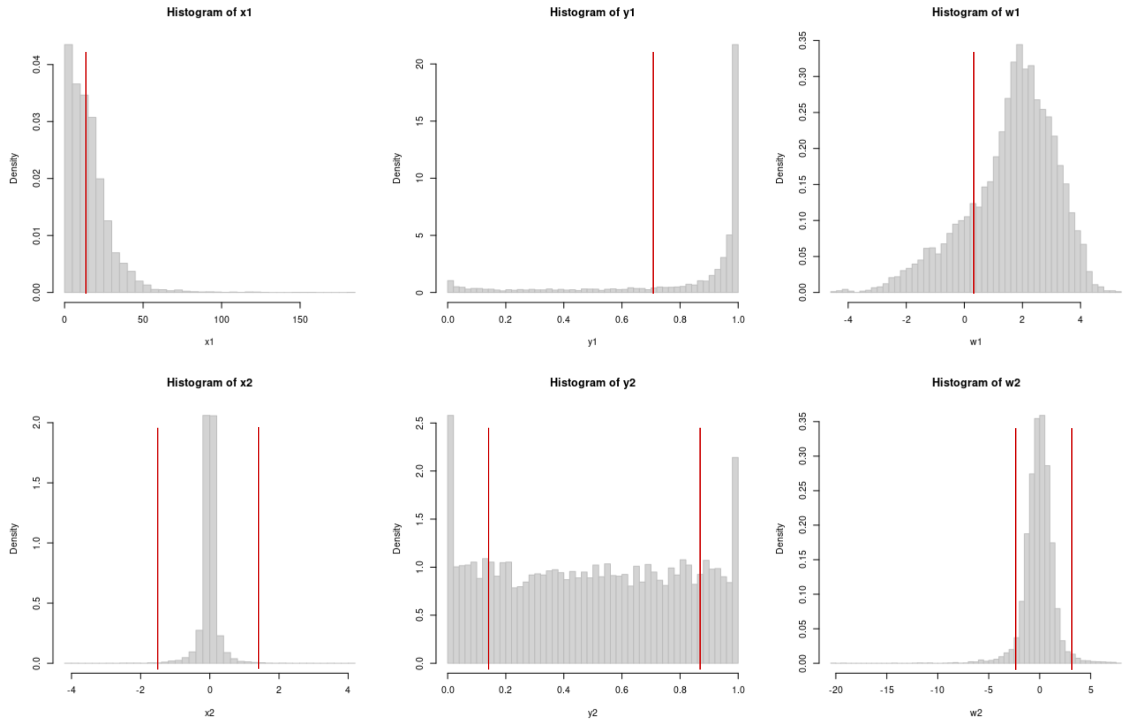
Ongoing Progress

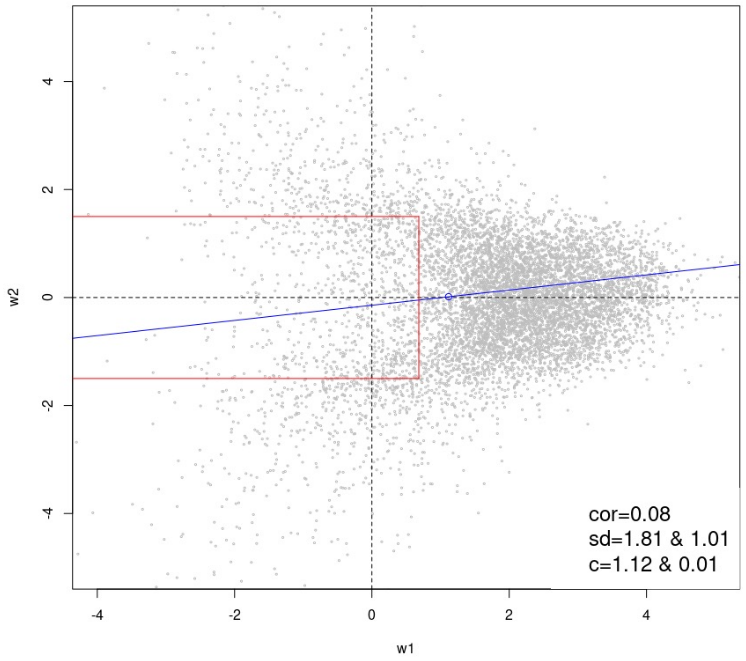
Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

Original data | Cdf(X) ~U(0,1) | $\Phi^{-1}(F_0(X_1)), \Phi^{-1}(G_0(X_2)) \sim N(0,1)$

X_1

X_2



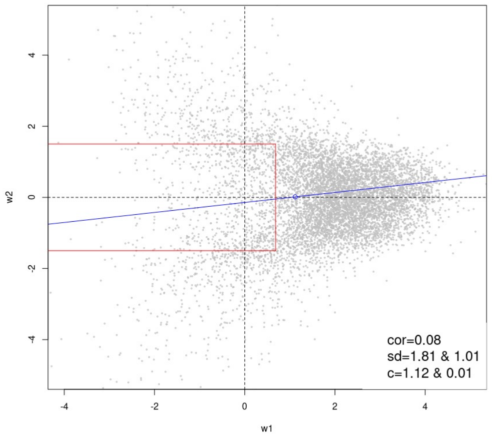


Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

(3) Sample data

(1) Select area



(2) Remove data out of the area

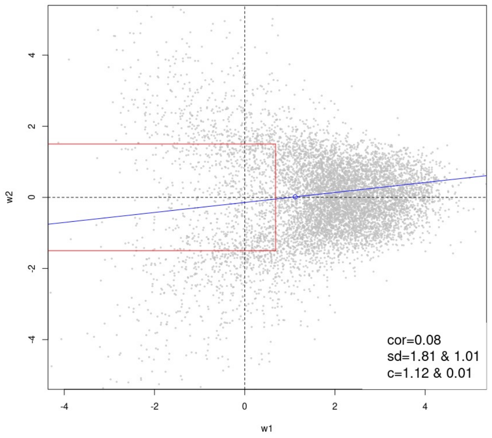
(4) Set the center and variance
->repeat (2)~(4) until convergence

(5) Get the parameters from (4)

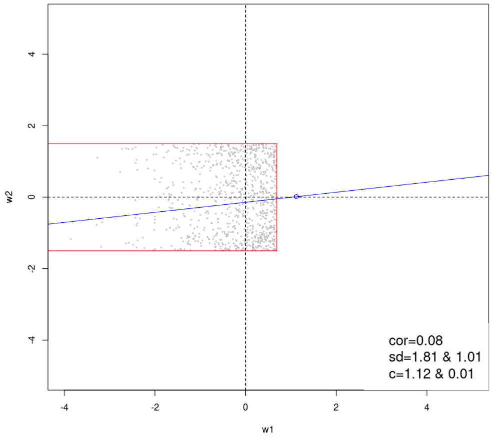
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

(1) Select area



(2) Remove data out of the area



(3) Sample data

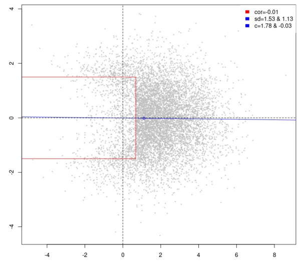
(4) Set the center and variance
->repeat (2)~(4) until convergence

(5) Get the parameters from (4)

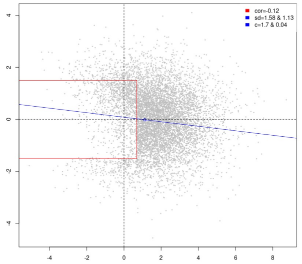
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

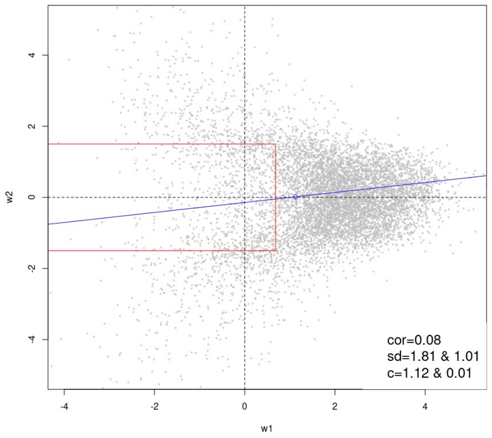
(3) Sample data



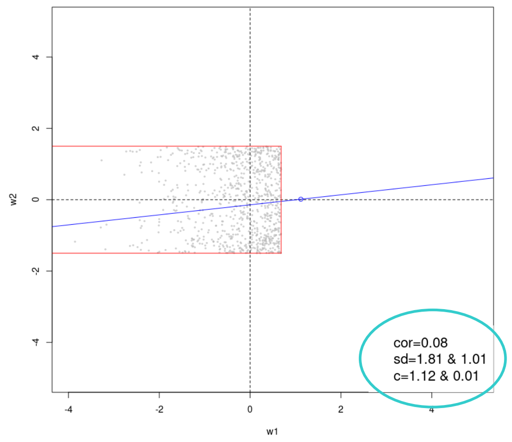
⋮



(1) Select area



(2) Remove data out of the area



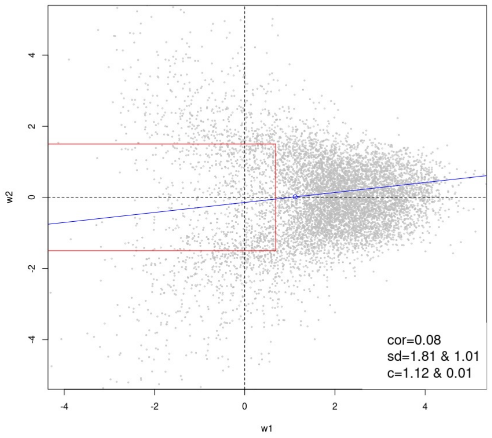
(4) Set the center and variance
->repeat (2)~(4) until convergence

(5) Get the parameters from (4)

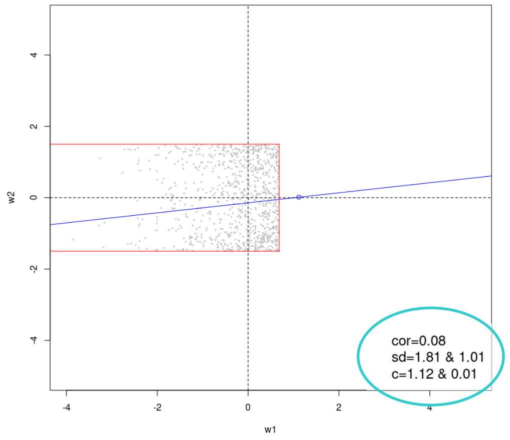
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

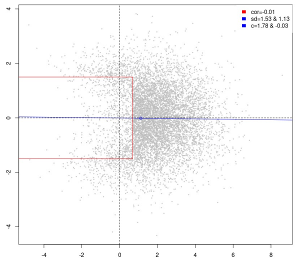
(1) Select area



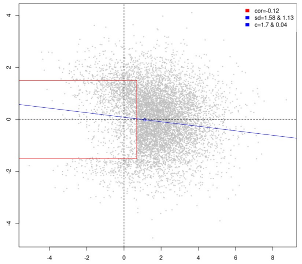
(2) Remove data out of the area



(3) Sample data



⋮



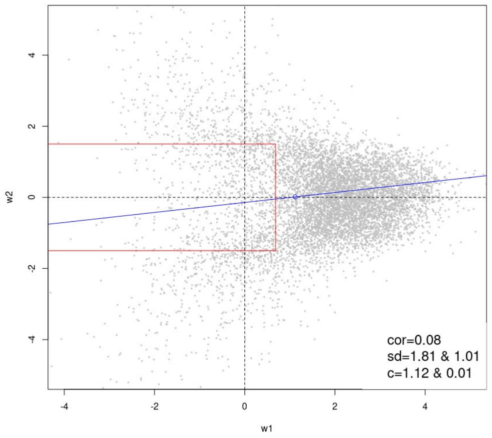
(4) Set the center and variance
->repeat (2)~(4) until convergence

(5) Get the parameters from (4)

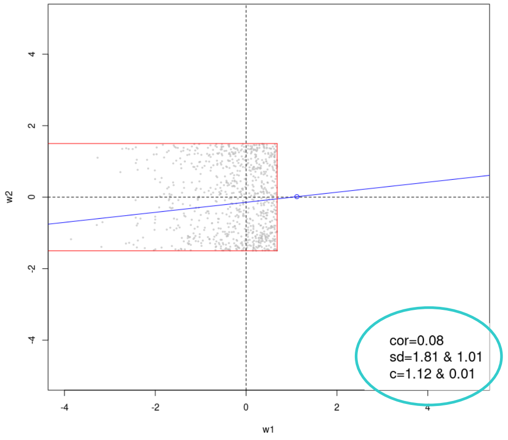
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

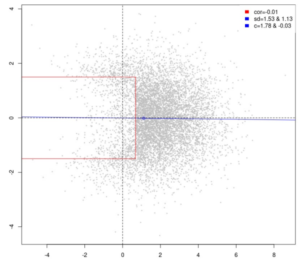
(1) Select area



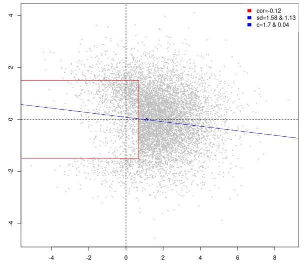
(2) Remove data out of the area



(3) Sample data

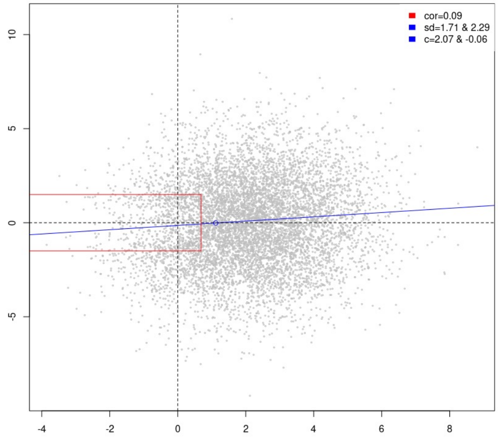


⋮



(4) Set the center and variance
->repeat (2)~(4) until convergence

(5) Get the parameters from (4)



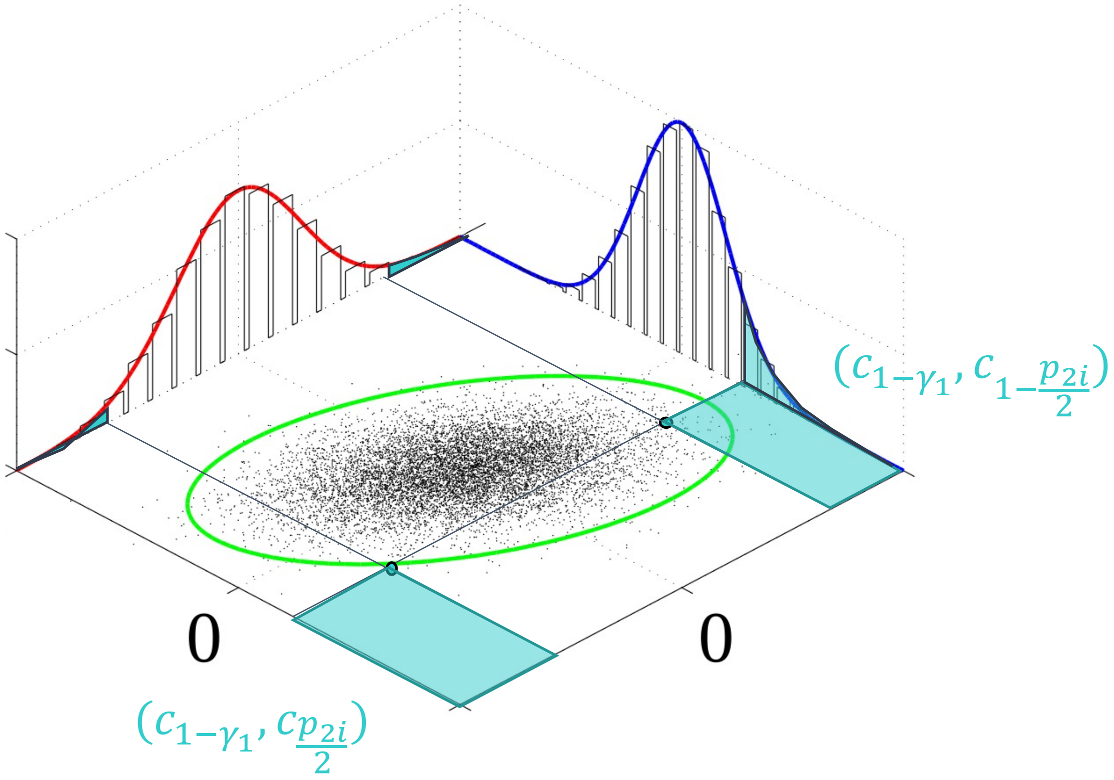
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → **Define p_i and reject null**

- $p_{1i} = 1 - \Phi_1(z_{1i})$
- $p_{2i} = 2\min(\Phi(z_{2i}), 1 - \Phi_2(z_{2i}))$
- $\gamma = P(p_{1i} \le \gamma_1, p_{2i} \le \gamma_2)$

$$p_i = \begin{cases} p_{1i}, \quad \text{if } p_{1i} > \gamma_1 \\ \int_{W_{1-\gamma_1}}^{\infty} \left\{ \int_{-\infty}^{\infty} \varphi_{z_1}(z_2) \left(I_{\left(-\infty, V_{\frac{p_{2i}}{2}}\right)} + I_{\left(V_{1-\frac{p_{2i}}{2}}, \infty\right)} \right) \right\} \phi_1(z_1) dz_2 dz_1 \end{cases}$$

- $W_c : \int_{W_c}^{\infty} \phi_1(z)\, dz = c$
- $V_c : \int_{V_c}^{\infty} \phi_2(z)\, dz = \int_{-\infty}^{V_c} \phi_2(z)\, dz = c$

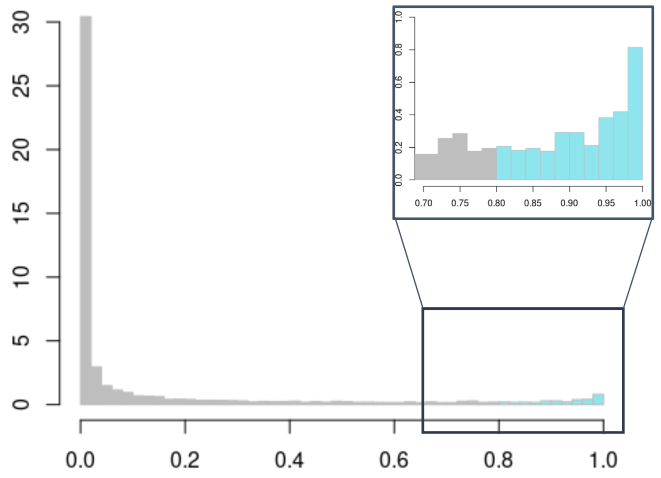


Ongoing Progress

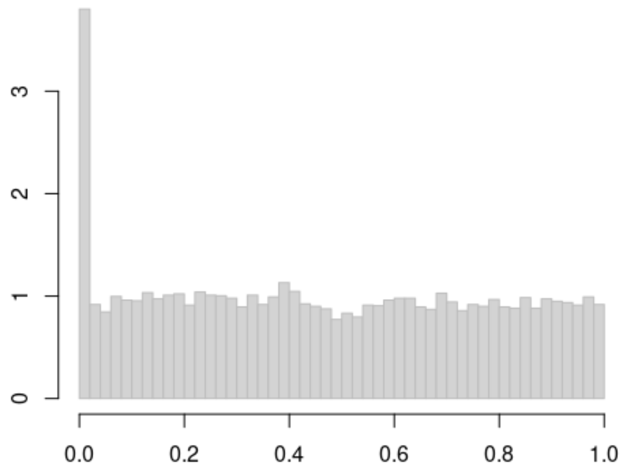
Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

For example) γ_1=0.8

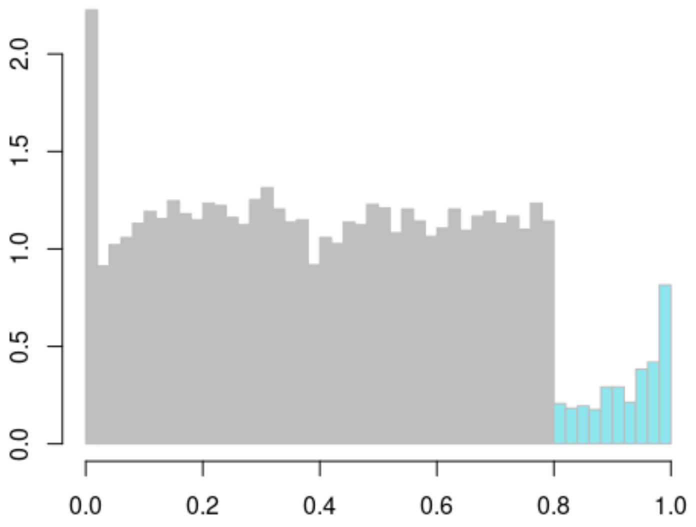
p_1



p_2

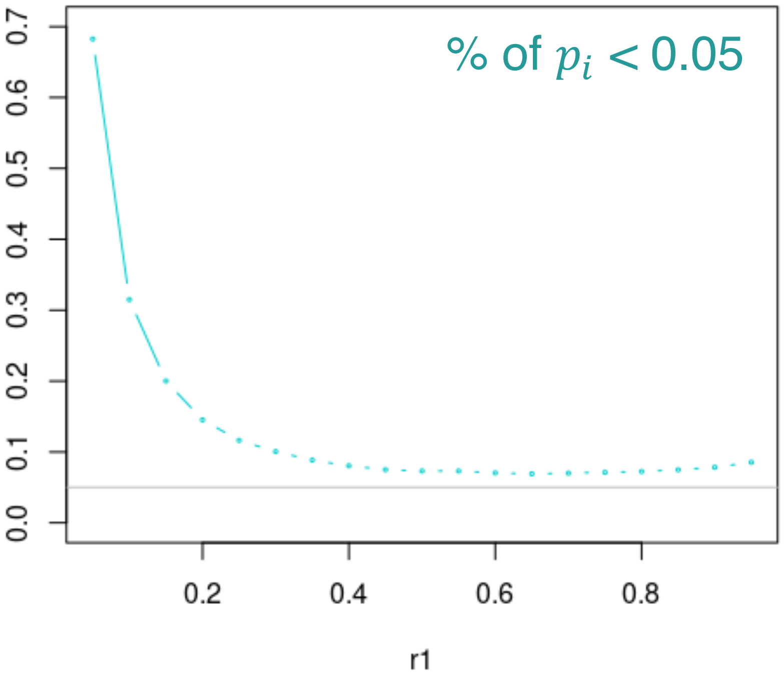


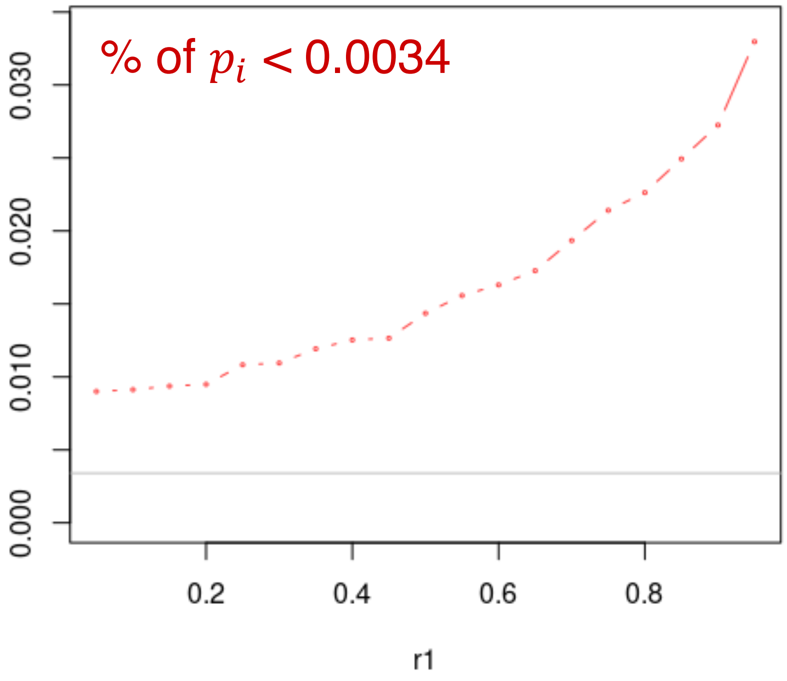
p_i



Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null

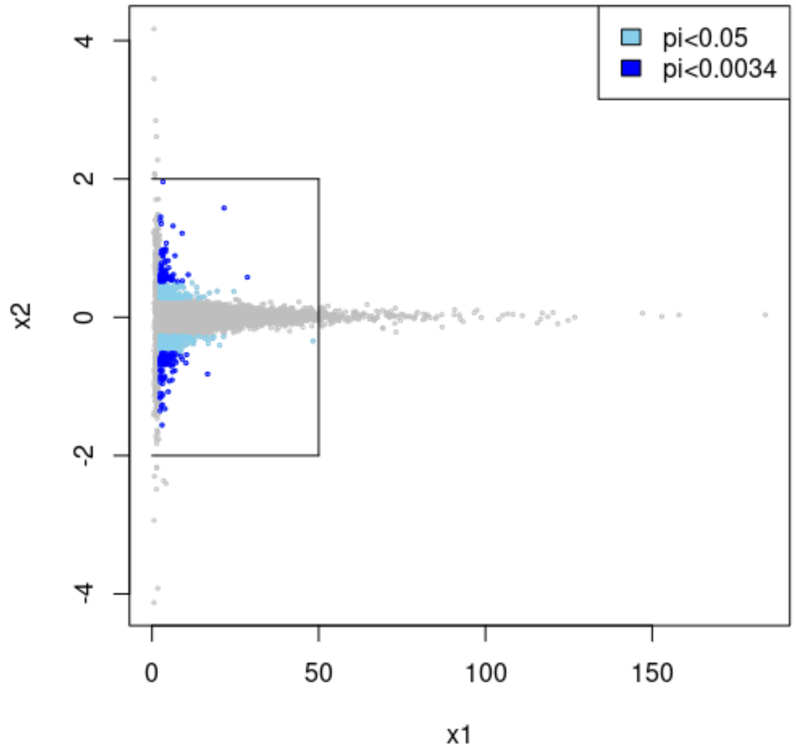


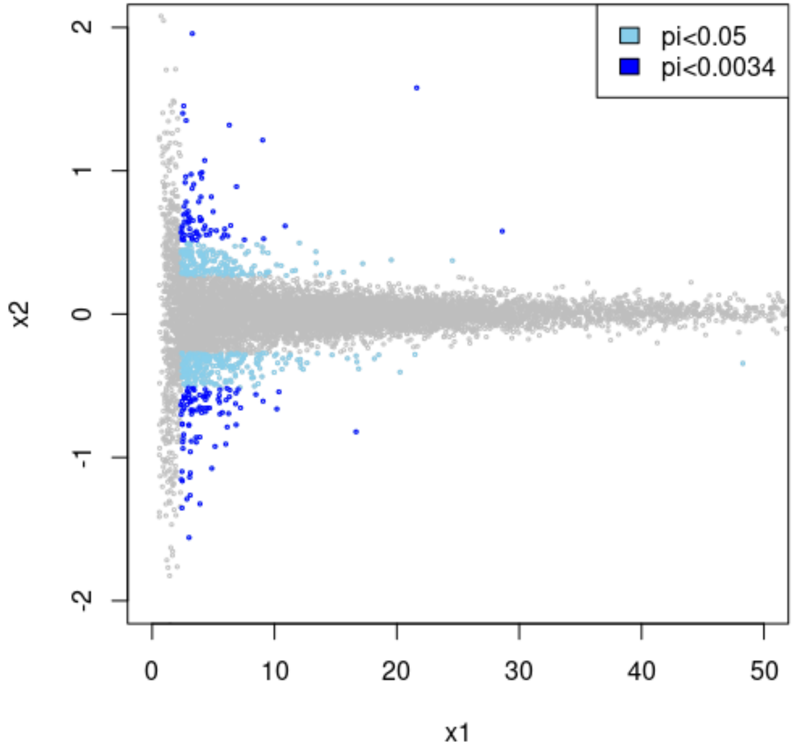


*0.0034 satisfies FDP<0.05

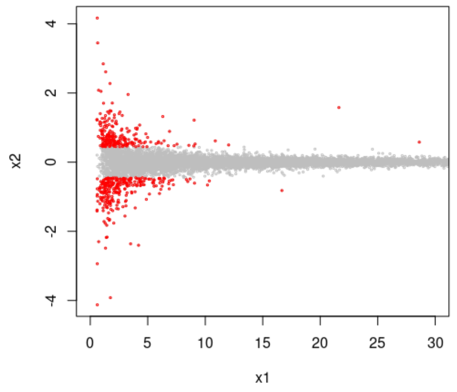
Ongoing Progress

Estimate f_0, π_0 → Using Gaussian Copula, fit the dist. of (X_1, X_2) under null → Define p_i and reject null





Cf) rejected only by locfdr



*example image

Plan

- Compare the result with local fdr
- Simulation

Thank you!