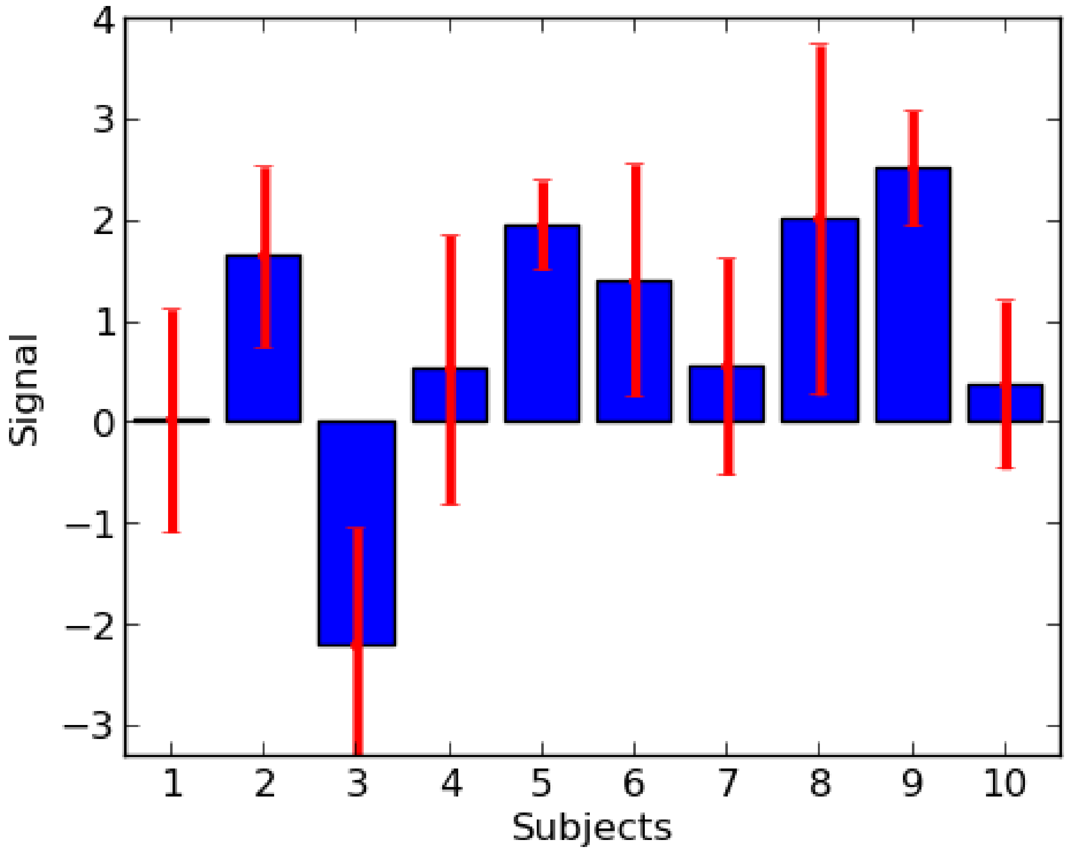
Fixed effects

We make observations of random variables $(\beta_i)_{i=1..n}$, that are assumed to be instances of unique random variable β. For each $i \in [1..n]$, let $\hat{\beta}_i$ and σ_i^2 be the empirical estimate and variance of β_i. What is the linear combination of these variables $\hat{\beta} = \sum_{i=1}^{n} w_i \beta_i \; s.t. \sum_{i=1}^{n} w_i = 1 \; and \; w_i > 0 \; \forall i \in [1..n]$
such that $\hat{\beta}$ has the minimal variance ? What are the corresponding variance and effect estimates ?

Illustration



Fixed effects

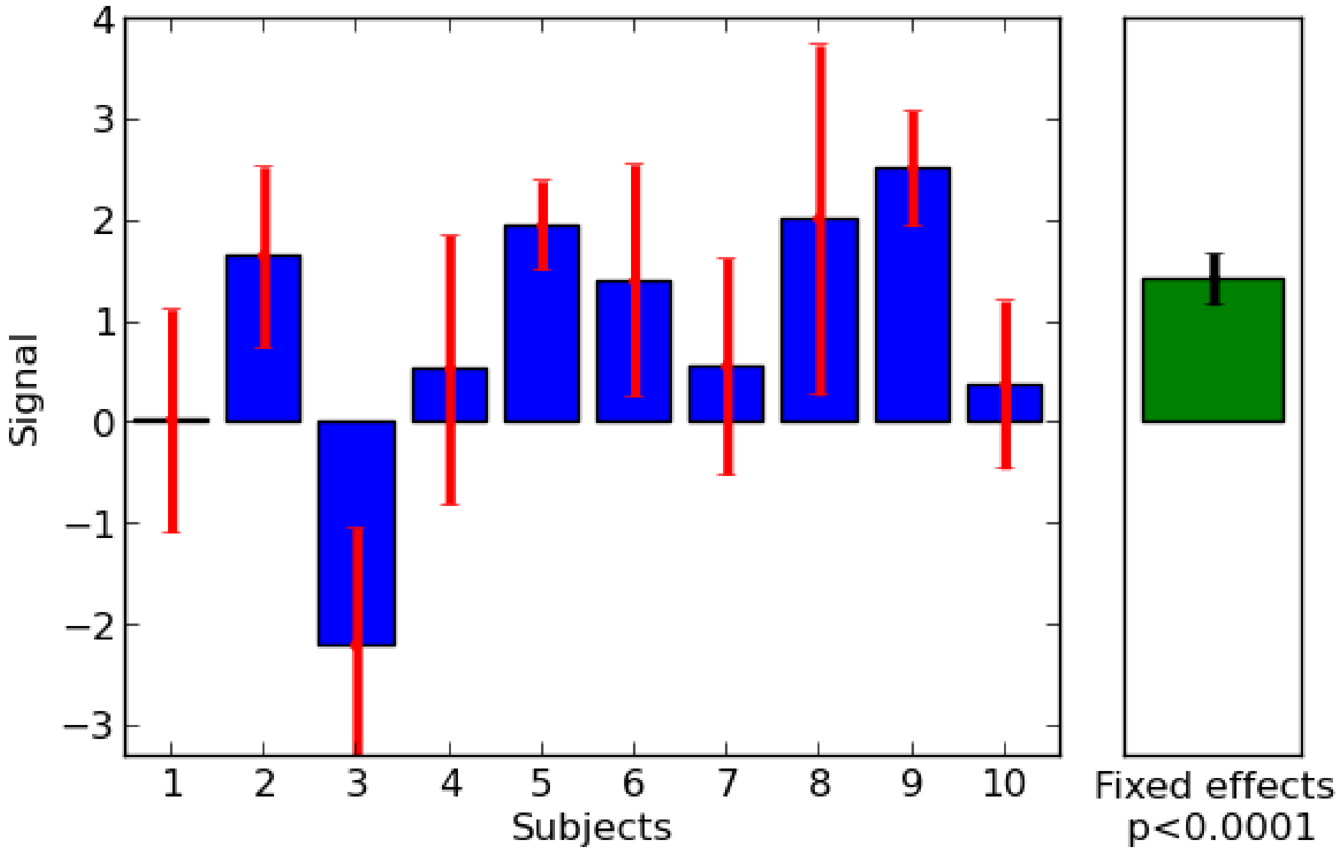
$$\min_{(w_i)} \sum_{i=1}^{n} w_i^2 \sigma_i^2 \text{ s.t. } \sum_{i=1}^{n} w_i = 1$$

$$\text{Constrained opt. } w_i \propto \frac{1}{\sigma_i^2} \implies w_i = \frac{1}{\sigma_i^2} \left(\sum_{j=1}^{n} \frac{1}{\sigma_j^2} \right)^{-1}$$

$$\text{Effect estimate: } \bar{\beta} = \sum_{i=1}^{n} \frac{\hat{\beta}_i}{\sigma_i^2} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1}$$

$$\text{Effect variance: } \bar{\sigma}^2 = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1}$$

Illustration



On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

Definition: $\text{p-value}(\mathbf{X}) = \alpha \Leftrightarrow \mathbb{P}(\tau > \hat{\tau}(\mathbf{X}) | H_0(\mathbf{X})) \leq \alpha$

On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

Definition: $\text{p-value}(\mathbf{X}) = \alpha \Leftrightarrow \mathbb{P}(\tau > \hat{\tau}(\mathbf{X}) | H_0(\mathbf{X})) \leq \alpha$

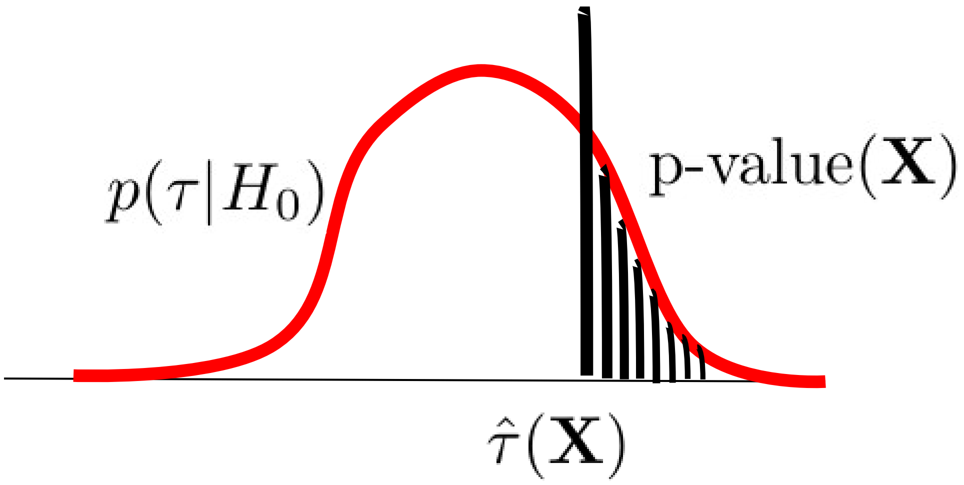
$\text{p-value}(\mathbf{X}) = 1 - \Phi(\hat{\tau}(\mathbf{X}))$ where $\Phi(t) = \int_{-\infty}^{t} p(\tau | H_0) d\tau$ is the null density

On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

Definition: $\text{p-value}(\mathbf{X}) = \alpha \Leftrightarrow \mathbb{P}(\tau > \hat{\tau}(\mathbf{X})|H_0(\mathbf{X})) \leq \alpha$

$\text{p-value}(\mathbf{X}) = 1 - \Phi(\hat{\tau}(\mathbf{X}))$ where $\Phi(t) = \int_{-\infty}^{t} p(\tau|H_0)d\tau$ is the null density



On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

Definition: $\text{p-value}(\mathbf{X}) = \alpha \Leftrightarrow \mathbb{P}(\tau > \hat{\tau}(\mathbf{X})|H_0(\mathbf{X})) \leq \alpha$

$\text{p-value}(\mathbf{X}) = 1 - \Phi(\hat{\tau}(\mathbf{X}))$ where $\Phi(t) = \int_{-\infty}^{t} p(\tau|H_0)d\tau$ is the null density

$$\boxed{\textbf{Property}: \mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha|H_0) = \alpha}$$

On the meaning of p-values

$$\boxed{\text{Data } \mathbf{X} \Longrightarrow \text{statistics } \hat{\tau}(\mathbf{X}) \Longrightarrow \text{p-value}(\mathbf{X})}$$

Definition: $\text{p-value}(\mathbf{X}) = \alpha \Leftrightarrow \mathbb{P}(\tau > \hat{\tau}(\mathbf{X}) | H_0(\mathbf{X})) \leq \alpha$

$\text{p-value}(\mathbf{X}) = 1 - \Phi(\hat{\tau}(\mathbf{X}))$ where $\Phi(t) = \int_{-\infty}^{t} p(\tau|H_0) d\tau$ is the null density

$\boxed{\textbf{Property}: \mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha | H_0) = \alpha}$

$$\begin{aligned}
\mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha | H_0) &= \mathbb{P}(1 - \Phi(\hat{\tau}(\mathbf{X})) \leq \alpha | H_0) \\
&= \mathbb{P}(\hat{\tau}(\mathbf{X}) \geq \Phi^{-1}(1-\alpha) | H_0) \\
&= \int_{\Phi^{-1}(1-\alpha)}^{\infty} p(\tau|H_0) d\tau \\
&= 1 - \Phi(\Phi^{-1}(1-\alpha)) = \alpha
\end{aligned}$$

On the meaning of p-values

$$\boxed{\textbf{Property: } \mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha | H_0) = \alpha}$$

On the meaning of p-values

$\boxed{\textbf{Property}: \mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha | H_0) = \alpha}$

Or equivalently: Under H_0, $\text{p-value}(\mathbf{X}) \leq \mathcal{U}(0,1)$

On the meaning of p-values

$\boxed{\textbf{Property}: \mathbb{P}(\text{p-value}(\mathbf{X}) \leq \alpha | H_0) = \alpha}$

Or equivalently: Under H_0, p-value$(\mathbf{X}) \leq \mathcal{U}(0,1)$

Not to be confused with Bayesian posterior:

$$p(H_0|\mathbf{X}) = \frac{p(\mathbf{X}|H_0)p(H_0)}{p(\mathbf{X}|H_0)p(H_0) + p(\mathbf{X}|H_1)p(H_1)}$$

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, ..., \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \text{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, ..., \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \text{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Goal **Non-asymptotic control of FWER**

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \mathrm{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Goal **Non-asymptotic control of FWER**

Critical region $W_\alpha(\mathbf{X}) = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha(\mathbf{X})\}$

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, ..., \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \text{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Goal **Non-asymptotic control of FWER**

Critical region $W_\alpha(\mathbf{X}) = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha(\mathbf{X})\}$

Null region $H_0 = \{i \in [p] : \mu_i = 0\}$

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, ..., \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \text{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Goal **Non-asymptotic control of FWER**

Critical region $W_\alpha(\mathbf{X}) = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha(\mathbf{X})\}$

Null region $H_0 = \{i \in [p] : \mu_i = 0\}$

$\text{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0)$; $\text{FWER}(W_\alpha) \leq \alpha$ wanted

Resampling in high dimension: multiple tests

[Arlot et al. Annals of Stats 2010]

Setting

Sample $\mathbf{X} = (\mathbf{X}^1, ..., \mathbf{X}^n) \in \mathbb{R}^{n \times p}$, $\mu_i = \text{mean}(\mathbf{X}_i)$

test $H_0 : \mu_i \leq 0$ vs $H_1 : \mu_i > 0$

$\mathbf{X}$ Gaussian, unknown Covariance

Goal **Non-asymptotic control of FWER**

Critical region $W_\alpha(\mathbf{X}) = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha(\mathbf{X})\}$

Null region $H_0 = \{i \in [p] : \mu_i = 0\}$

$\text{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0)$; $\text{FWER}(W_\alpha) \leq \alpha$ wanted

For instance: if covariance$= \sigma \mathbb{I}_p$; Bonferroni $t_\alpha^{bonf} = \sigma \Phi^{-1}(1 - \frac{\alpha}{p})$

Resampling in high dimension: multiple tests

$$\boxed{\mathrm{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0);\ \mathrm{FWER}(W_\alpha) \leq \alpha \text{ wanted}}$$

Resampling in high dimension: multiple tests

$$\boxed{\mathrm{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0); \ \mathrm{FWER}(W_\alpha) \leq \alpha \text{ wanted}}$$

Family of thresholds indexed by region C: $\tau(\alpha, C), C \subset [p]$

Resampling in high dimension: multiple tests

$$\boxed{\mathrm{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0);\ \mathrm{FWER}(W_\alpha) \leq \alpha \text{ wanted}}$$

Family of thresholds indexed by region C: $\tau(\alpha, C), C \subset [p]$

2 hypotheses:

- $\mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > \tau(\alpha, H_0)) \leq \alpha$
- $\tau(\alpha, C) \nearrow C$: $C \subset C' \Rightarrow \tau(\alpha, C) \leq \tau(\alpha, C')$

Resampling in high dimension: multiple tests

$$\boxed{\mathrm{FWER}(W_\alpha) = \mathbb{P}(W_\alpha \bigcap H_0) \neq 0);\ \mathrm{FWER}(W_\alpha) \leq \alpha \text{ wanted}}$$

Family of thresholds indexed by region C: $\tau(\alpha, C), C \subset [p]$

2 hypotheses:

- $\mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > \tau(\alpha, H_0)) \leq \alpha$
- $\tau(\alpha, C) \nearrow C$: $C \subset C' \Rightarrow \tau(\alpha, C) \leq \tau(\alpha, C')$

Proposition: Taking $W_\alpha = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha([p])\}$ yields $\alpha-$level control of the FWER.

Resampling for multiple tests

Proposition: Taking $W_\alpha = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha([p])\}$ yields $\alpha-$level control of the FWER.

Resampling for multiple tests

Proposition: Taking $W_\alpha = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha([p])\}$ yields $\alpha-$level control of the FWER.

$$\begin{aligned} \mathbb{P}(\exists i : \bar{\mathbf{X}}_i > t_\alpha([p]) \text{ and } i \in H_0) &= \mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > t_\alpha([p])) \\ &\leq \mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > t_\alpha(H_0)) < \alpha \end{aligned}$$

Resampling for multiple tests

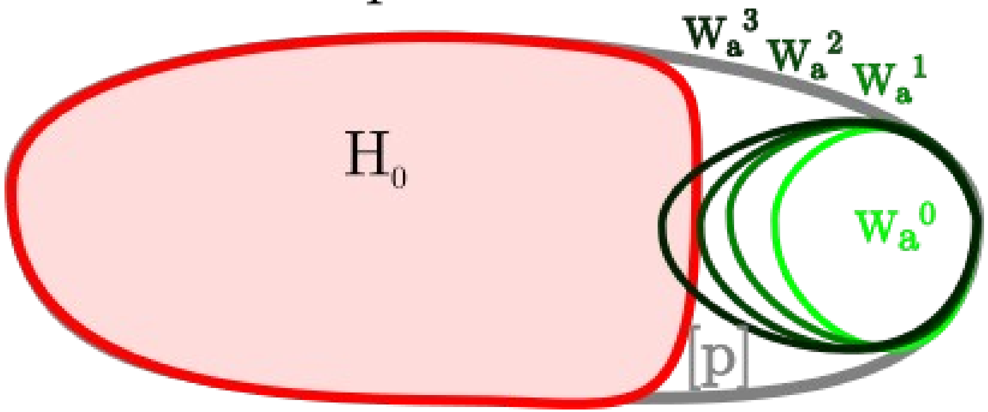
Proposition: Taking $W_\alpha = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha([p])\}$ yields $\alpha-$level control of the FWER.

$$\begin{aligned} \mathbb{P}(\exists i : \bar{\mathbf{X}}_i > t_\alpha([p]) \text{ and } i \in H_0) &= \mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > t_\alpha([p])) \\ &\leq \mathbb{P}(\sup_{i \in H_0} \bar{\mathbf{X}}_i > t_\alpha(H_0)) < \alpha \end{aligned}$$

Note: Generalization to multi-step procedure: "step-down"

$$W_\alpha^{j+1} = \{i \in [p] : \bar{\mathbf{X}}_i > t_\alpha([p] - W_\alpha^j)\}$$

$\bigcup_j W_\alpha^j$ controls the FWER



Resampling for multiple tests

Rademacher variables $Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$

Resampling for multiple tests

Rademacher variables $Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$

$$\boxed{t_\alpha(\mathbf{X}, C) = \inf\{t \in \mathbb{R} : \mathbb{P}_Z(\sup_{i \in C} \left(\frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}_i^j > t \right) \leq \alpha\}}$$

Resampling for multiple tests

$$\text{Rademacher variables } Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$$

$$\boxed{t_\alpha(\mathbf{X}, C) = \inf\{t \in \mathbb{R} : \mathbb{P}_Z(\sup_{i \in C} \left(\frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}_i^j > t \right) \leq \alpha\}}$$

$$Z\bar{\mathbf{X}} = \frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}^j$$

Resampling for multiple tests

Rademacher variables $Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$

$$\boxed{t_\alpha(\mathbf{X}, C) = \inf\{t \in \mathbb{R} : \mathbb{P}_Z(\sup_{i \in C} \left(\frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}_i^j > t \right) \leq \alpha\}}$$

$$\bar{Z\mathbf{X}} = \frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}^j$$

$\boxed{\text{Proposition: } t_\alpha(\mathbf{X}, [p]) \text{ controls the FWER at level } \alpha}$

Resampling for multiple tests

Rademacher variables $Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$

$$\boxed{t_\alpha(\mathbf{X}, C) = \inf\{t \in \mathbb{R} : \mathbb{P}_Z(\sup_{i \in C} \left(\frac{1}{n}\sum_{j=1}^n Z^j \mathbf{X}_i^j > t\right) \le \alpha\}}$$

$$Z\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^n Z^j \mathbf{X}^j$$

$\boxed{\text{Proposition: } t_\alpha(\mathbf{X}, [p]) \text{ controls the FWER at level } \alpha}$

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}}\left(\sup_{i \in H_0} \bar{\mathbf{X}}_i > t_\alpha([p])\right) &= \mathbb{P}_{\mathbf{X}}\left(\sup_{i \in H_0} Z\bar{\mathbf{X}}_i > t_\alpha(Z\mathbf{X}, [p])\right) \text{ for any } Z \\
&= \mathbb{E}_Z \mathbb{P}_{\mathbf{X}}\left(\sup_{i \in H_0} Z\bar{\mathbf{X}}_i > t_\alpha(Z\mathbf{X}, [p])\right) \\
&= \mathbb{E}_{\mathbf{X}} \mathbb{P}_Z\left(\sup_{i \in H_0} Z\bar{\mathbf{X}}_i > t_\alpha(\bar{\mathbf{X}}, [p])\right) \\
&= \mathbb{E}_{\mathbf{X}} \mathbb{P}_Z\left(\sup_{i \in H_0} Z\bar{\mathbf{X}}_i > t_\alpha([p])\right) \\
&\le \alpha
\end{aligned}$$

Resampling for multiple tests

Rademacher variables $Z = (Z^j, j = 1..n) : \mathbb{P}(Z_j = 1) = \mathbb{P}(Z_j = -1) = \frac{1}{2}$

$$\boxed{t_\alpha(\mathbf{X}, C) = \inf\{t \in \mathbb{R} : \mathbb{P}_Z(\sup_{i \in C} \left(\frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}_i^j > t\right) \leq \alpha\}}$$

$$\bar{Z\mathbf{X}} = \frac{1}{n} \sum_{j=1}^n Z^j \mathbf{X}^j$$

$\boxed{\text{Proposition: } t_\alpha(\mathbf{X}, [p]) \text{ controls the FWER at level } \alpha}$

Notes:

- this holds for any non-Gaussian symmetric distribution
- Can be combined with step-down procedure