

Logistic Regression Equivalence

A Framework for Comparing Logistic
Regression Models Across Populations

Guy Ashiri-Prossner

February 4th, 2020

- **Introduction**

- Equivalence Testing
- Background
- Methods
- Simulation Results
- MATAI Results
- Discussion

- Psychometric Introduction

- Motivation
- Research Aim

Psychometrics

- An applied research domain, dealing with both theory and methodology
- Main issue: measuring and evaluating human abilities, knowledge, etc.
- Usually we measure latent variables
- Two main tools:
 - Factor Analysis (understanding hidden structures)
 - Logistic Regression (prediction of abilities or properties)
- Obtained models are judged for their:
 - Accuracy
 - Interpretation

Psychometric Models Comparison

- Populations differ by culture, socio-economic status, physiology, etc.
- The comparison of similar tools across populations is a common technique
- Equivalence across populations is very important (scientifically):
 - We want to know whether our explanation fits other populations
- More important than possible statistical or computational considerations
- Holding to an existing model could be a priority:
 - It has provided us a good explanation of a phenomenon
 - Experts were trained to use this model
 - Further researches were based on this model

Psychometric Models Comparison (cont.)

- Introducing a new model over an existing one might imply:
 - Loss of explanatory power and generality (scientifically expensive)
 - Reduced fit to the previous data
 - Weakening the existing model and the researches it was used
 - Possible changes to the relations between covariates
- Examples:
 - Equivalence of the 4-factor personality model Across Gender (Byrne, 1988)
 - Equivalence of the 5-factor occupational preferences model across race (Collins & Gleaves, 1998)
 - Nonequivalence of the factorial structure of the family values questionnaire across cultures (Byrne & van de Vijver, 2010)

Motivation: Comparing Logistic Regression Models

We would like to provide different viewpoints for the comparison

- How does a model describe a phenomenon?
- How do the model outputs fit the observations?
- On aggregate, how accurate are the model predictions?

Motivation (Cont.)

- There is an existing framework for comparing factor analysis models
- It is named “Measurement Invariance” (Mellenbergh, 1989)
- It checks whether “the measurement properties of X [...] are the same across populations” (Millsap, 2012)
- Measurement Invariance consists of several comparison steps
- Each step relating to a different aspect of the factor analysis models
- No such framework exists for comparing logistic regression models
- Existing logistic regression comparison tools are insufficient for our needs

Research Aim

Developing methods for comparing logistic regression models across populations, relating to different properties of logistic regression.

- Introduction
- **Equivalence Testing**
- Background
- Methods
- Simulation Results
- MATAL Results
- Discussion

- The Problem With Significance Testing
- Indifference/Acceptance Regions
- Equivalence Testing

The Problem With Significance Testing

- Significance tests are sensitive to sample size, and tend to reject the null as n grows
- Expects no effect ($H_0: |\mu_1 - \mu_2| = 0$)
- Example: $x_1, \dots, x_n \underset{iid}{\sim} \mathcal{N}(0,1)$

- Tests on the sample mean are in the form of $\left\{ |\bar{x}| > \frac{1}{\sqrt{n}} z_{1-\alpha/2} \right\}$
- The region of not rejecting the null shrinks as n grows:

n	10	25	50	100	250	500
$s_{\bar{x}}$	0.316	0.2	0.141	0.1	0.063	0.045

- This does not encourage good research practices

Setting an Indifference Region

- Consider two populations $x^A \sim \mathcal{N}(\mu_A, \sigma^2)$, $x^B \sim \mathcal{N}(\mu_B, \sigma^2)$
- We would like to test for their mean difference
- Define a minimal observable effect size ϵ
- Hypotheses are now $H_0: |\mu_A - \mu_B| < \epsilon$, $H_1: |\mu_A - \mu_B| \geq \epsilon$
- Tests on the sample means are in the form of $\left\{ \frac{\bar{x}_A - \bar{x}_B}{\sigma/\sqrt{n}} \geq z_{1-\alpha/2} + \frac{\epsilon}{\sigma/\sqrt{n}} \right\}$
- Assume given significance level α and power $1 - \beta$
- A large enough sample would lead to reject H_0 (Bickel & Doksum, 2015)

$$n \geq \frac{\sigma^2}{\epsilon^2} \left(\Phi^{-1}(1 - \alpha) - \Phi^{-1}(\beta) \right)^2$$

- How could we overcome the sample size sensitivity?

Equivalence Testing

- Define a minimal observable effect size ϵ :

$$H_0: |\mu_A - \mu_B| < \epsilon, \quad H_1: |\mu_A - \mu_B| \geq \epsilon$$

- Flip hypotheses direction:

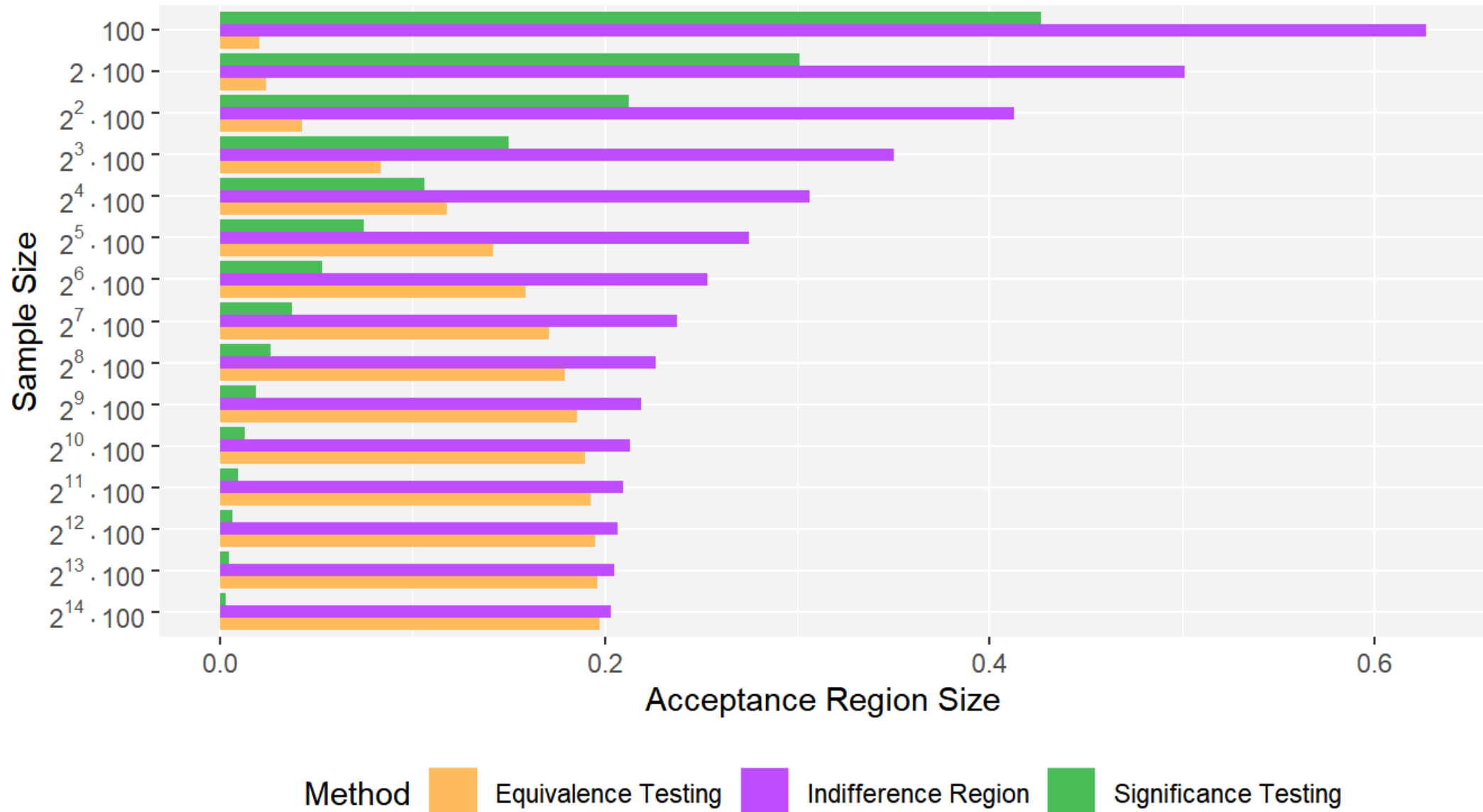
$$H_0: |\mu_A - \mu_B| \geq \epsilon, \quad H_1: |\mu_A - \mu_B| < \epsilon$$

- The burden of proof is now on the 'equivalent' side
- There are formulae for univariate and multivariate cases (Wellek, 2010)
- Requires defining both α and ϵ

Acceptance Region Example

- Let $y_1, \dots, y_{100} \sim \mathcal{N}(0,1)$
- Let X be n repetitions of Y : $(y_1, \dots, y_{100}, y_1, \dots, y_{100}, \dots, y_1, \dots, y_{100}) = X$
- Define $\epsilon = 0.1$ as the indifference region and $\alpha = 0.05$
- Significance test: $\{\bar{x} \geq s_x \cdot z_{1-\alpha/2}\}$, using ϵ : $\{\bar{x} - \epsilon \geq s_x \cdot z_{1-\alpha/2}\}$
- For both methods, the acceptance region shrinks as n grows

Comparison of Acceptance Region Sizes



- Introduction
- Equivalence Testing
- **Background**
- Methods
- Simulation Results
- MATAL Results
- Discussion

- Logistic Regression Recap
- Comparable Logistic Regression Model Properties
- Existing Logistic Regression Comparison Tools


Logistic Regression Recap

- $X \in \mathbb{R}^{n \times p}, y \in \{0,1\}^n$
- $\theta_i = x_i^T \beta$
- $E[y_i | x_i] = P(y_i = 1 | x_i) = \pi_i$


x_i is the sample
 y_i is the predicted variable
 π_i is the probability
 θ_i is the linear predictor
 β is the coefficients vector
 V is the covariance of $\hat{\beta}$

- $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$
- $L(\beta) = \prod_{i=1}^n \frac{(e^{\theta_i})^{y_i}}{1+e^{\theta_i}} \rightarrow l(\beta) = \sum_{i=1}^n \left(y_i \beta^T x_i - \log \left(1 + e^{\beta^T x_i} \right) \right)$
- $\hat{\beta} = \arg \max l(\beta) \rightarrow \hat{\beta} \sim \mathcal{N}(\beta, V), V = \text{Cov}(\beta)$
- $\hat{\theta}_i \sim \mathcal{N}(\theta_i, x_i^T V x_i)$

Data & Models


 $\rightarrow \underbrace{\begin{pmatrix} 0.1 & 1.3 & 0 & \dots & 4.7 \\ \vdots & \vdots & \vdots & \dots & \vdots \end{pmatrix}}_{X_A, n_A \times p}, \underbrace{\begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}}_{y_A, n_A \times 1} \rightarrow M^A: \underbrace{\begin{pmatrix} \hat{\beta}_0^A \\ \vdots \\ \hat{\beta}_p^A \end{pmatrix}}_{\hat{\beta}^A}, \hat{V}^A = \text{Cov}(\hat{\beta}^A), \hat{\theta}^A$

Population A


 $\rightarrow \underbrace{\begin{pmatrix} 0.6 & 2.5 & 8 & \dots & 9.3 \\ \vdots & \vdots & \vdots & \dots & \vdots \end{pmatrix}}_{X_B, n_B \times p}, \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}}_{y_B, n_B \times 1} \rightarrow M^B: \underbrace{\begin{pmatrix} \hat{\beta}_0^B \\ \vdots \\ \hat{\beta}_p^B \end{pmatrix}}_{\hat{\beta}^B}, \hat{V}^B = \text{Cov}(\hat{\beta}^B), \hat{\theta}^B$

Population B

Comparable Logistic Regression Model Properties

We would like to provide different viewpoints for the comparison

- How does a model describe a phenomenon?
 - We use the regression coefficients vector $\hat{\beta}$
 - It provides an insight to the model's explanation of the phenomenon
- How do the model outputs fit the observations?
 - We use the log-odds (linear predictors) vector $\hat{\theta}$
 - $\hat{\theta}_i \in \mathbb{R}$ is a better choice than $\hat{\pi}_i \in [0,1]$ or $\hat{y}_i \in \{0,1\}$
- On aggregate, how accurate are the model predictions?
 - We use Brier score

Existing Logistic Regression Comparison Tools

- Information criteria (AIC/BIC)
 - χ^2 statistics, derived from the maximal value of the likelihood function
 - Describing a model's fit to its training data
 - Compared directly or normalized
 - Becomes degenerate as sample size grows (Tanaka, 1987)
- Deviance Test (Pregibon and others, 1981)
 - Using two datasets, which differ by variable(s) x_g
 - Building two logistic regression model
 - Hypothesis testing H_0 : "Introducing x_g to the model does not improve its fit"

Existing Logistic Regression Comparison Tools

- Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1989)
 - Data is split to G groups according to the fitted probabilities
 - A classical goodness of fit statistic ($\sum_{g=1}^G \frac{(O_g - E_g)^2}{E_g}$) is computed
 - Is a significance test and thus suffers from large-sample setbacks
- Other interesting equivalence testing methods relate to:
 - Collapsibility ($y = X\beta + Z\gamma$ against $y = X\beta$), which isn't our case (von Davier 2003)
 - Univariate ($p = 1$) regression models (Counsell & Cribbie 2014)
 - Ad-hoc techniques (e.g. $\bigwedge_{i=1}^p \epsilon_L < \frac{\beta_i^1}{\beta_i^2} < \epsilon_U$) (Jonkman 2009)

Existing Logistic Regression Comparison Tools

- Accuracy: $\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$
 - Describing the model predictions fit to the observed dependent variable
 - Isn't a proper scoring function (Gnieting & Raftery, 2007)
- Brier Score: $\frac{1}{N} \sum_{i=1}^N (y_i - \hat{\pi}_i)^2$ (Brier, 1950; Benedetti, 2010)
 - Describing the model output probabilities fit to the observed dependent variable
 - Is a proper scoring function
 - Full sampling distribution is still unknown

Proper Scoring Functions

- A proper scoring function gives the highest reward (lowest penalty) for forecasts closest to the true distribution of the data (Gneiting & Raftery, 2007)
- Example:

y_i	Forecasts	Distance	Accuracy penalty	Brier Penalty
$y_1 = 0$	$\hat{\pi}_1 = 0.1, \hat{y}_1 = 0$	0.1	$(0 - 0)^2 = 0$	$(0 - 0.1)^2 = 0.01$
$y_2 = 1$	$\hat{\pi}_2 = 0.3, \hat{y}_2 = 0$	0.7	$(1 - 0)^2 = 1$	$(1 - 0.3)^2 = 0.49$
$y_3 = 1$	$\hat{\pi}_3 = 0.6, \hat{y}_3 = 1$	0.4	$(1 - 1)^2 = 0$	$(1 - 0.6)^2 = 0.16$
$y_4 = 1$	$\hat{\pi}_4 = 0.8, \hat{y}_4 = 1$	0.2	$(1 - 1)^2 = 0$	$(1 - 0.8)^2 = 0.04$

- Introduction
- Equivalence Testing
- Background
- **Methods**
- Simulation Results
- MATAL Results
- Discussion

- General Recipe
- Regression Coefficients
Equivalence Testing
- Log-Odds Equivalence Testing
- Brier Score Equivalence Testing

General Recipe for Equivalence Tests

1. Pick a comparable logistic regression property
2. Find its approximate distribution
3. Construct a significance test
4. Assume sensitivity level δ
5. Invert hypotheses
6. Obtain equivalence test
7. Refine sensitivity level δ (if needed)

Method 1: Coefficient Vectors Equivalence Testing

- Let models M^A, M^B with coefficient vectors $\hat{\beta}^A, \hat{\beta}^B$
- Define $\hat{\beta}^A - \hat{\beta}^B = \hat{q} \sim \mathcal{N}(\beta^A - \beta^B, S_q)$
- Covariance matrix:
$$S_q = \text{Var}(\hat{q}) = (\hat{V}^A + \hat{V}^B), \quad \hat{V}^A = \text{Cov}(\hat{\beta}^A), \hat{V}^B = \text{Cov}(\hat{\beta}^B)$$

- Wald Test statistic:

$$W_\beta = \hat{q}^T S_q^{-1} \hat{q}$$

- Null hypothesis:

$$H_0: \|\beta^A - \beta^B\|_\Sigma = 0$$

- Exact level- α significance test:

$$\{W_\beta < \chi_{p,\alpha}^2\}$$

Method 1: Coefficient Vectors Equivalence Testing

- Denote δ_β as an allowed difference per coefficient ($|\hat{\beta}_j^A - \hat{\beta}_j^B| \leq \delta_\beta$)
- δ_β^p is a $p \times 1$ vector of δ_β
- Updated null hypothesis

$$H_0: \|\beta^A - \beta^B\|_\Sigma \leq \lambda_\beta, \quad \lambda_\beta^2 = \delta_\beta^{pT} S_q^{-1} \delta_\beta^p$$

- Updated test statistic distribution:

$$W_\beta \sim \chi_{p,\alpha}^2(\lambda_\beta^2)$$

- Inverted hypotheses for equivalence testing:

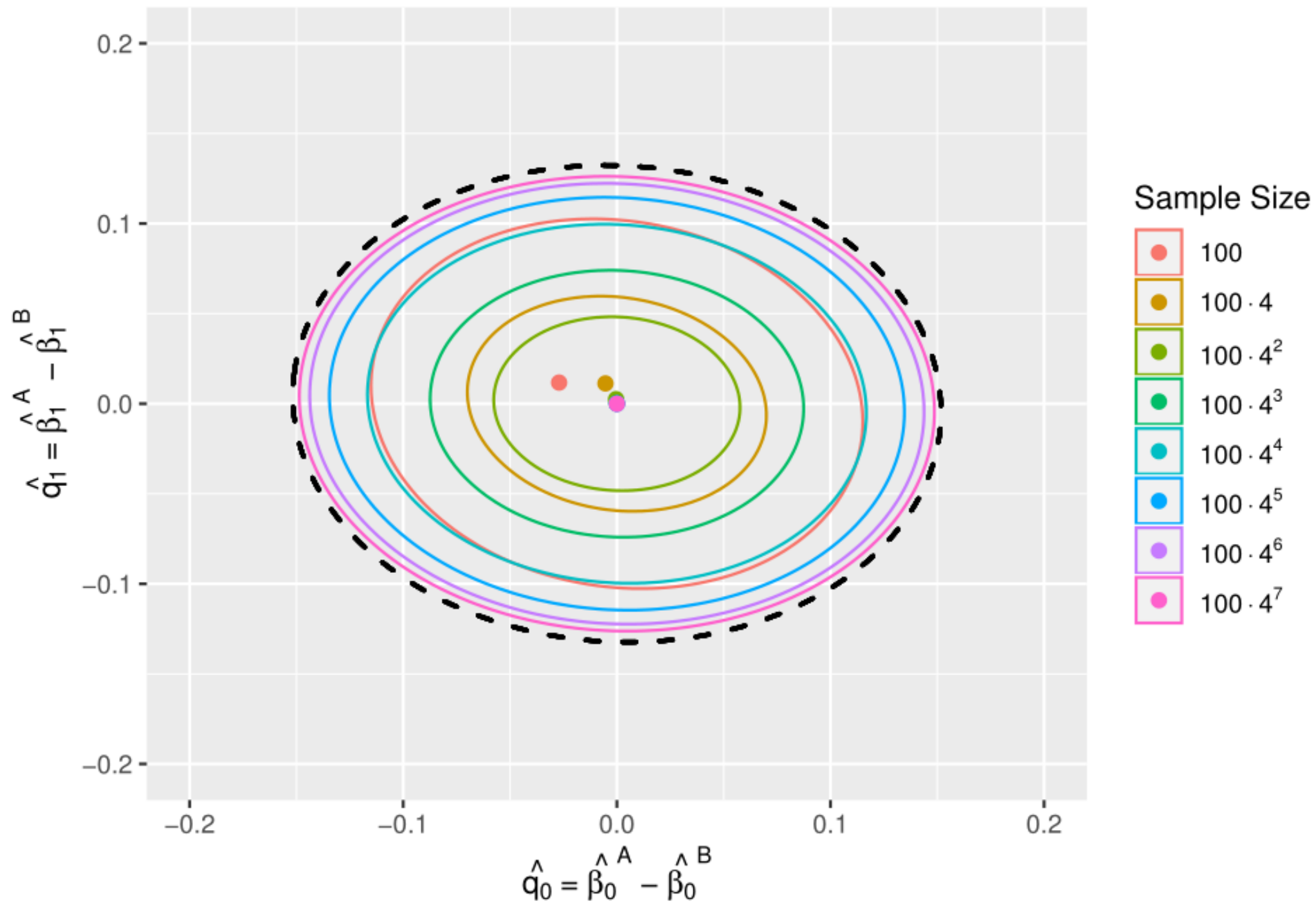
$$H_0: \|\beta^A - \beta^B\|_\Sigma \geq \lambda_\beta, \quad H_1: \|\beta^A - \beta^B\|_\Sigma < \lambda_\beta$$

- Exact level- α equivalence test:

$$\left\{ \hat{q}^T S_q^{-1} \hat{q} < \chi_{p,\alpha}^2 \left(\delta_\beta^{pT} S_q^{-1} \delta_\beta^p \right) \right\}$$

Equivalence Regions Example

- Consider models M^A, M^B where $x_i^T \hat{\beta}^A = x_i$, $x_i^T \hat{\beta}^B = a + (1 - a)x_i$
- $\hat{\beta}^A = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\hat{\beta}^B = \begin{pmatrix} a \\ 1 - a \end{pmatrix}$
- Let $\hat{q} = \hat{\beta}^A - \hat{\beta}^B$
- We can plot the equivalence regions using \hat{q}_0, \hat{q}_1 coordinates



Reminder: Log-Odds Ratio

- Consider population A with model M^A as our *source*
- New population B with model M^B is our *target* of size k
- Does the source model fit the target population?
- Let $x_i \in X_B$ and $\hat{\theta}_i^A, \hat{\theta}_i^B$ the linear predictors obtained by the models
- $\hat{\theta}_i$ is also the log-odds:

$$\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = \frac{e^{\theta_i}}{1 + e^{\theta_i}} / \frac{1}{1 + e^{\theta_i}} = e^{\theta_i}$$

- Log-odds ratio:

$$\log \left(\frac{e^{\hat{\theta}_i^A}}{e^{\hat{\theta}_i^B}} \right) = (x_i^T \hat{q} | \hat{\beta}^A, \hat{\beta}^B) = \hat{\theta}_i^A - \hat{\theta}_i^B = \hat{\xi}_i$$

Method 2: Log-Odds Equivalence Testing

- Let $\tilde{\xi}_i = |\hat{\xi}_i|$.
- $\tilde{\xi} \sim G$, the conditional distribution of the change in the log-odds, with mean $\mu_{\tilde{\xi}}$
- Significance testing with indifference region:
$$H_0: \mu_{\tilde{\xi}} < \delta_\theta, \quad H_1: \mu_{\tilde{\xi}} \geq \delta_\theta$$
- Equivalence hypotheses:
$$H_0: \mu_{\tilde{\xi}} \geq \delta_\theta, \quad H_1: \mu_{\tilde{\xi}} < \delta_\theta$$
- Exact level- α equivalence test:

$$\left\{ \frac{\sqrt{k}(\bar{\tilde{\xi}} - \delta_\theta)}{\sqrt{\widehat{Var}(\tilde{\xi})}} < t_{1-\frac{\alpha}{2}, k-1} \right\}$$

Choosing δ for Log-Odds Equivalence

- \hat{y}_i takes its value from the sign of $\hat{\theta}_i$: $\hat{y}_i = I_{\{\hat{\theta}_i > 0\}}$
- We would like δ_θ to be a bound on the changes to \hat{y}
- Label x_i as *flip* if a change of δ_θ to its log-odds might change its \hat{y}_i value:

$$\text{sign}(\hat{\theta}_i) \neq \text{sign}(\hat{\theta}_i + \delta_\theta) \vee \text{sign}(\hat{\theta}_i) \neq \text{sign}(\hat{\theta}_i - \delta_\theta)$$

- Given target population X_B of size k , denote r as the allowed *flips* ratio in the data
- δ_θ is then the r^{th} quantile of $|\hat{\theta}^B|$

Brier Score

- Let models M^A, M^B and test set X_C of size k
- For $x_i \in X_C$, denote $b_i^A = (\hat{\pi}_i^A - y_i)^2$
- Brier Score for model M^A on dataset X_C is then

$$BS_{AC} = \frac{1}{k} \sum_{i=1}^k (\hat{\pi}_i^A - y_i)^2 = \frac{1}{k} \sum_{i=1}^k b_i^A$$

- As $b_i^A \in [0,1]$, we can approximate $Var(BS_{AC}) \leq \frac{1}{4k}$
- Similarly, BS_{BC} is the Brier score for model M^B on the same data

$$BS_{BC} = \frac{1}{k} \sum_{i=1}^k (\hat{\pi}_i^B - y_i)^2 = \frac{1}{k} \sum_{i=1}^k b_i^B$$

Method 3: Brier Score Equivalence Testing

- BS_{AC}, BS_{BC} are means for paired samples
- $D_{AB} = BS_{AC} - BS_{BC}$
- Using indifference region $[-\delta_B, \delta_B]$ requires two significance tests:
 - $t_L = \frac{\sqrt{k}(D_{AB} + \delta_B)}{\sqrt{s_D^2}}, t_U = \frac{\sqrt{k}(D_{AB} - \delta_B)}{\sqrt{s_D^2}}$
$$\{t_L \leq t_{1-\alpha, k-1} \vee t_U \geq -t_{1-\alpha, k-1}\}$$
- Exact level- α equivalence test:
$$\{t_L > t_{1-\alpha, k-1} \wedge t_U < -t_{1-\alpha, k-1}\}$$

Choosing δ for Brier Scores Equivalence

- As $BS \in [0,1]$, differences smaller than 0.01 might be meaningless
- The BS variance is bounded $Var(BS) \leq \frac{1}{4k}$, so $SD(BS) \leq \frac{1}{2\sqrt{k}}$
- The allowed difference should be manually selected:
 $\delta_B \in [0.01, 0.5]$

- Introduction
- Equivalence Testing
- Background
- Methods
- **Simulation Results**
- MATA Results
- Discussion

- Sim 1: Comparison Against Common Methods
- Sim 2: Controlling α

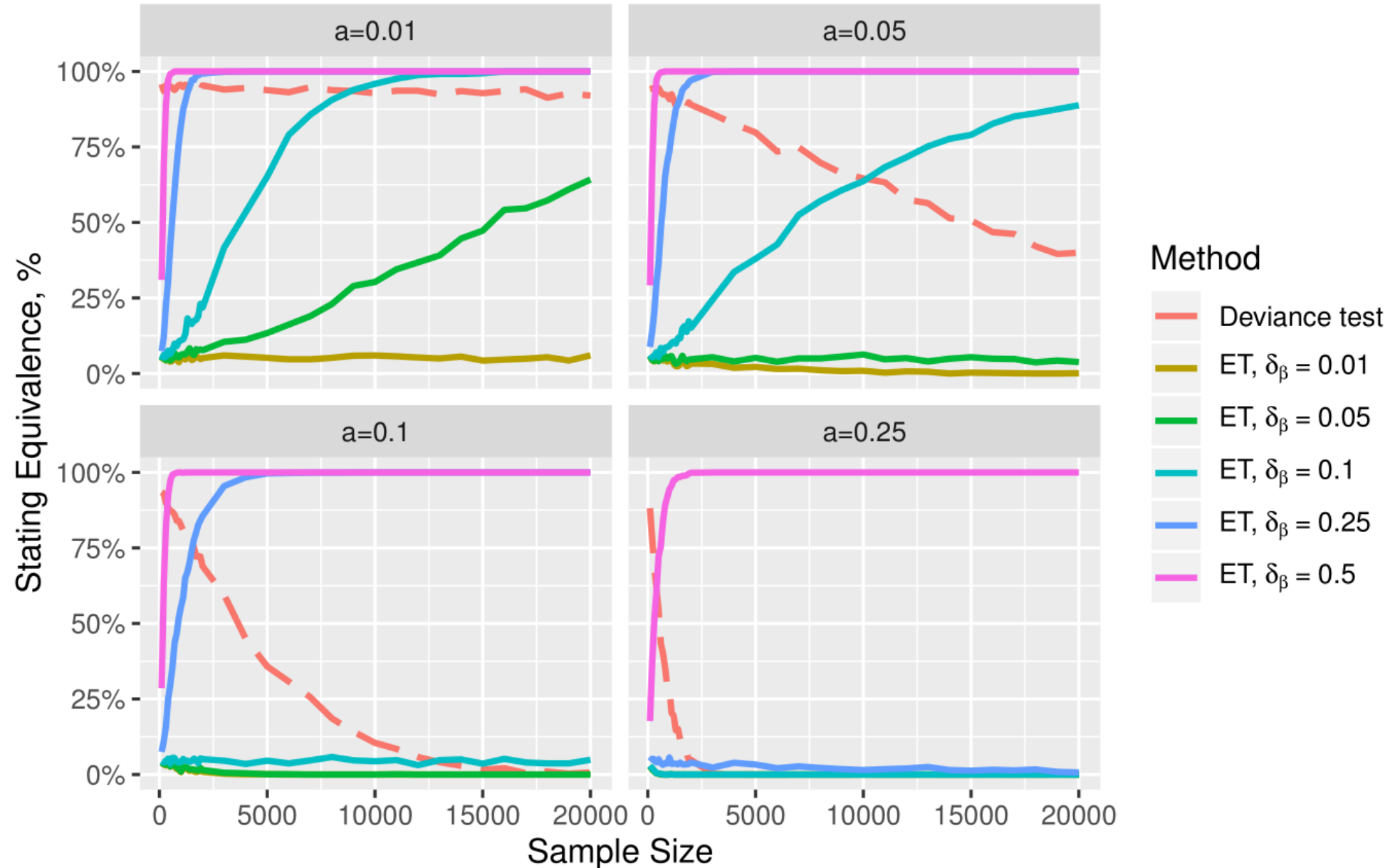
Sim 1: Comparison Against Common Methods

- We construct datasets with effect size a and test the performance of the proposed methods and common methods
- We expect the equivalence methods to perform better than the common methods for large sample size
- General setting:
 - Sample x^A, x^B of size n , where $x_i^A, x_i^B \sim \mathcal{N}(0,1)$.
 - Set $\theta_i^A = x_i^A$, $\theta_i^B = a + (1 - a)x_i^B$
 - Set $\pi_i^A = \text{sigmoid}(\theta_i^A)$ and then sample $y_i^A \sim \text{Ber}(\pi_i^A)$, similarly for y_i^B
 - Build logistic regression models M^A, M^B using $(x^A, y^A), (x^B, y^B)$
- We use $n = 100, 200, \dots, 2000, 3000, \dots, 20000$, $a = 0.01, 0.05, 0.1, 0.25$ and $\alpha = 0.05$.
- Each combination of n, a and method is repeated 1000 times.

Sim 1.1: Coefficients Equivalence Method

- We would like to test the performance of the coefficients equivalence method against the deviance test
- Deviance test:
 - Construct $X_{G_r} = \begin{pmatrix} x^A & 0 \\ 0 & x^B \end{pmatrix}$, $X_G = \begin{pmatrix} (1 & x^A) & 0 \\ 0 & (1 & x^B) \end{pmatrix}$
 - Use X_{G_r} for the *reduced* model, X_G for the *full* model
 - Test statistic: $D = 2 \left(l(\hat{\beta}_{reduced}) - l(\hat{\beta}_{full}) \right) \sim \chi^2$
 - Null hypothesis: “Introducing indicators to the model does not improve its fit”
 - For a given n , how many times (out of 1000) was H_0 not rejected?
- Coefficients Equivalence Method:
 - Hypotheses $H_0: \|\hat{\beta}^A - \hat{\beta}^B\|_{\Sigma} \geq \lambda_{\beta}$, $H_1: \|\hat{\beta}^A - \hat{\beta}^B\|_{\Sigma} < \lambda_{\beta}$
 - For a given n , how many times (out of 1000) was H_0 rejected?

Sim 1.1: Coefficients Equivalence Method



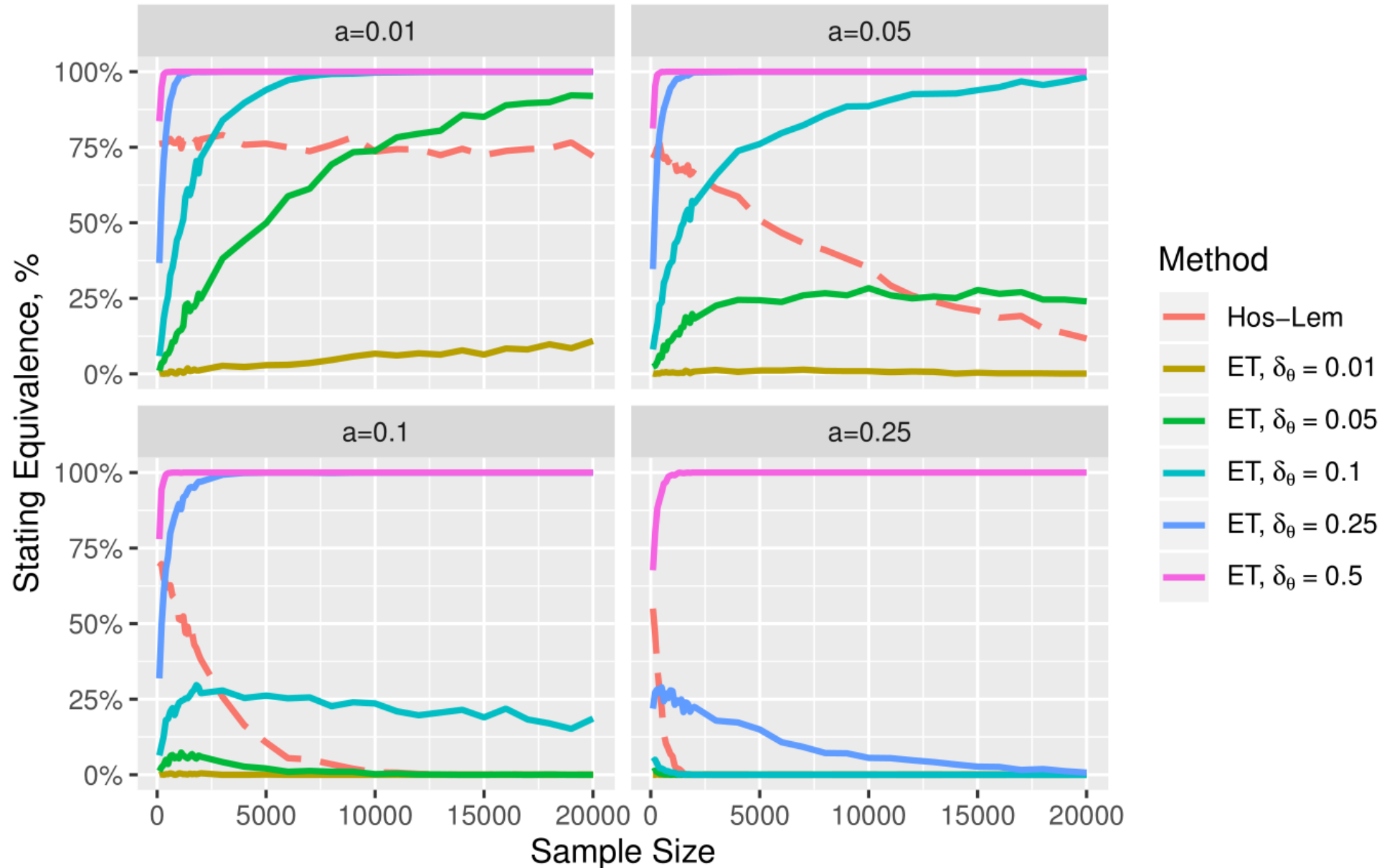
Sim 1.1: Coefficients Equivalence Method

- Choosing sensitivity level smaller than the effect size ($\delta_\beta < a$) yields constant failure in identifying equivalence.
- Choosing $\delta_\beta > a$ succeeds, as we tolerate effect sizes larger than the actual effect.
- The coefficients equivalence test improves its performance as n grows. It also outperforms the deviance test.

Sim 1.2: Log-Odds Equivalence Method

- Same $(x^A, y^A), (x^B, y^B), M^A, M^B$
- Conduct the Hosmer-Lemeshow test
 - Classify the samples (x_i, y_i) to G distinct groups, according to $\hat{\pi}_i$
 - For each group g denote $\bar{\pi}_g$ the average probability for $y = 1$, E_{1_g} the expected number of 1's and O_{1_g} the observed number of 1's
 - Test statistic: $\sum_{g=1}^G \left\{ \left(O_{1_g} - E_{1_g} \right)^2 / \left(n_g \bar{\pi}_g (1 - \bar{\pi}_g) \right) \right\} = H \sim \chi_{G-2}^2$
 - Hypotheses $H_0: M^A$ fits X_B , $H_1: M^A$ doesn't fit X_B
 - For a given n , how many times (out of 1000) was H_0 not rejected?
- Conduct the log-odds equivalence test
 - Hypotheses $H_0: E[|\hat{\theta}_i^A - \hat{\theta}_i^B|] \geq \delta_\theta$, $H_1: E[|\hat{\theta}_i^A - \hat{\theta}_i^B|] < \delta_\theta$
 - For a given n , how many times (out of 1000) was H_0 rejected?

Sim 1.2: Log-Odds Equivalence Method



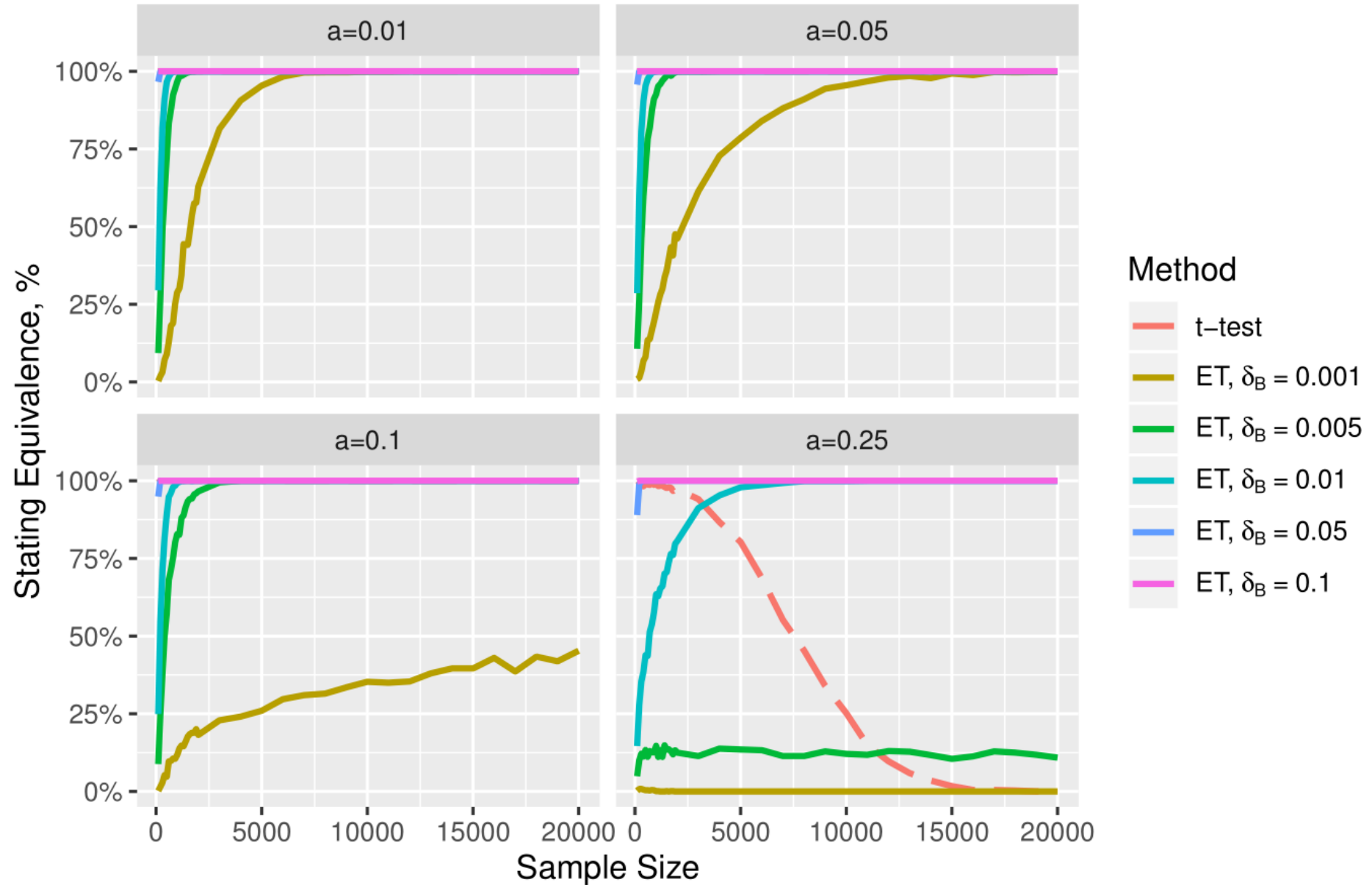
Sim 1.2: Log-Odds Equivalence Method

- Again, choosing sensitivity $\delta_\theta < a$ yields constant failure in identifying equivalence.
- Choosing $\delta_\theta > a$ succeeds, as we tolerate effect sizes larger than the actual effect.
- The log-odds equivalence test improves its performance as n grows. It also outperforms the Hosmer-Lemeshow test.

Sim 1.3: Brier Score Equivalence Method

- Same $(x^A, y^A), (x^B, y^B), M^A, M^B$
- Brier score of model M^A on data (x^C, y^C) : $BS_{AC} = \frac{1}{n_C} \sum_k (\hat{\pi}_k^A - y_k)^2$
- Calculate $BS_{AA}, BS_{AB}, BS_{BA}, BS_{BB}$
- For each dataset C , conduct a t-test on the Brier scores difference
 - Test statistic: $T = \sqrt{n_C}(BS_{BC} - BS_{AC}) / \sqrt{Var(BS_{BC}) + Var(BS_{AC})}$
 - Hypotheses $H_0: BS_{BC} = BS_{AC}, H_1: BS_{BC} \neq BS_{AC}$
 - For a given n , how many times (out of 1000) was H_0 not rejected?
- Conduct the Brier score equivalence test
 - Hypotheses $H_0: |BS_{BC} - BS_{AC}| \geq \delta_B, H_1: |BS_{BC} - BS_{AC}| < \delta_B$
 - For a given n , how many times (out of 1000) was H_0 rejected?

Sim 1.3: Brier Score Equivalence Method



Sim 1.3: Brier Score Equivalence Method

- It seems that choosing $\delta_B \geq 0.01$ provides very high rates of identifying equivalence.
- Lower δ_B values might not perform well.
- The Brier equivalence test improves its performance as n grows. It also outperforms the t-test, but only for large effect and large n .

Sim 2: Controlling α

- We construct datasets with effect size a and observe their performance with respect to given α
- General setting: Similar to simulation #1
- We use $n = 100, 1000, 10000$, $a = 0.1, 0.25, 0.4$ and $\alpha = 0.05, 0.1$
- For each method we choose sensitivity level δ according to our recommendations

Sim 2.1: Coefficients Equivalence Method

		$a = 0.1$			$a = 0.25$			$a = 0.4$		
	n / δ_β	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
$\alpha = 0.05$	100	0.03	0.04	0.05	0.04	0.04	0.06	0.04	0.04	0.07
	1000	0.04	0.04	0.13	0.01	0.05	0.17	0.01	0.05	0.18
	10000	0	0.05	0.68	0	0.04	0.69	0	0.05	0.72
$\alpha = 0.1$	100	0.1	0.1	0.12	0.08	0.11	0.13	0.08	0.09	0.14
	1000	0.06	0.09	0.27	0.03	0.1	0.29	0.02	0.08	0.27
	10000	0	0.1	0.81	0	0.1	0.8	0	0.12	0.84

The coefficients equivalence method does control α upon choosing $\delta_\beta \leq a$

Sim 2.2: Log-Odds Equivalence Method

		$\alpha = 0.1$			$\alpha = 0.25$			$\alpha = 0.4$		
n / r		5%	10%	15%	5%	10%	15%	5%	10%	15%
$\alpha = 0.05$	100	0.1	0.31	0.52	0.06	0.18	0.35	0.01	0.06	0.14
	1000	0.27	0.74	0.93	0	0.06	0.3	0	0	0
	10000	0.17	0.95	1	0	0	0.1	0	0	0
$\alpha = 0.1$	100	0.11	0.31	0.5	0.07	0.19	0.35	0.02	0.05	0.14
	1000	0.3	0.7	0.91	0	0.06	0.32	0	0	0
	10000	0.14	0.96	1	0	0	0.1	0	0	0

The log-odds equivalence method does control α for $\alpha = 0.25, 0.4$ and $n = 10^3, 10^4$. for small effect size ($\alpha = 0.1$) the chosen r values are too large.

Sim 2.3: Brier Score Equivalence Method

		$\alpha = 0.1$			$\alpha = 0.25$			$\alpha = 0.4$		
	n / δ_B	0.001	0.01	0.1	0.001	0.01	0.1	0.001	0.01	0.1
$\alpha = 0.05$	100	0	0	1	0	0	0.98	0	0	0.94
	1000	0	0	1	0	0	1	0	0	1
	10000	0	1	1	0	0.86	1	0	0	1
$\alpha = 0.1$	100	0	0	1	0	0	1	0	0	0.97
	1000	0	0.2	1	0	0.01	1	0	0	1
	10000	0	1	1	0	0.97	1	0	0	1

As sample size grows, the Brier equivalence method fails to control α .

- Introduction
- Equivalence Testing
- Background
- Methods
- Simulation Results
- **MATAL Results**
- Discussion

The MATAL System

- A computer-based test battery for the diagnosis of learning disabilities (dyslexia, dysgraphia and dyscalculia) & ADHD
- Includes 22 tests
- Uses logistic regression models for the prediction of each disability
- Introduced 2007
- Around 5000 cases per year
- Re-written 2016-2018

MATAL Data

- A large, well-researched dataset ($n = 1046$)
- Data was split by gender ($n_f = 591, n_m = 455$)
- Each sub-sample was used to construct logistic regression models for dysgraphia and dyscalculia
- Each gender-based models were tested for both datasets
- $\alpha = 0.05, \delta_\beta = 0.1, r = 7.5\%, \delta_B = 0.05$

Dysgraphia

- **Deviance test:** gender-based models are equivalent
- **Coefficients equivalence test:** gender-based models are equivalent
- **Hosmer-Lemeshow test:** gender-based models are not equivalent
- **Log-odds equivalence test:** gender-based models are not equivalent
- **Brier score t-test:** gender-based models are equivalent
- **Brier score equivalence test:** gender-based models are equivalent

fMRI studies: no gender differences in brain activation were observed on most writing tasks (Berninger & O'Malley May, 2011)

	Brier Scores	
	Male model	Female model
Male data	0.1227	0.1283
Female data	0.0950	0.0922

Dyscalculia

- **Deviance test:** gender-based models are not equivalent
- **Coefficients equivalence test:** gender-based models are not equivalent
- **Hosmer-Lemeshow test:** gender-based models are not equivalent
- **Log-odds equivalence test:** gender-based models are not equivalent
- **Brier score t-test:** gender-based models are equivalent
- **Brier score equivalence test:** gender-based models are equivalent

“Significantly more girls than boys could be defined as having developmental dyscalculia”
(Devine et al, 2013)

	Brier Scores	
	Male model	Female model
Male data	0.1172	0.1215
Female data	0.1199	0.1107

- Introduction
- Equivalence Testing
- Background
- Methods
- Simulation Results
- MATAL Results
- **Discussion**

Usage

- The usage of the proposed methods is subject to the research goal of the scientist
- Each method suggests a different form of equivalence:
 - Coefficients equivalence suggests descriptive equivalence – both models describe the phenomenon in an equivalent manner
 - Log-odds equivalence suggests individual predictive equivalence – both models yield the equivalent predictions (per sample)
 - Brier score equivalence suggests overall predictive equivalence – both models have an equivalent average prediction accuracy
- Using more than one method requires multiplicity correction to α

Advantages

- A new approach for comparing logistic regression models
- Each method provides a different insight
- Comprehensive view of the models using method combinations
- Can enhance meta-analyses
- Using equivalence tests ensures performance for large sample sizes
- Easy to use tests, similar to well-known tests
- The Brier scores method can be used without training data

Limitations

- Usage of normal approximations for logistic regression estimates
- Usage of normal approximation for the Brier score
- The log-odds equivalence method might be too strict
- The Brier score equivalence method might be too lenient
- It tends to reject the inequivalence hypothesis
- It sometimes contradicts the other methods
- The researcher needs to decide what is an acceptable equivalence margin, a “key methodological issue” of equivalence testing (Greene et al., 2008)

What's Next?

- Additional comparison steps, such as comparison of outputted probabilities
- Meaningful interpretation of each step (as in Vandenberg & Lance, 2000)
- Incorporating equivalence testing in the measurement invariance framework
- Extending the current methods for other classifier types
- Using the proposed methods in psychometric context

Thank you