حجم نمونهٔ بهینه در مدل‌سازی چند سطحی: بررسی تأثیر حجم نمونه بر اثرهای ثابت و تصادفی با استفاده از داده‌های تیمز

■ زهرا نقش* ■ زهرا هاشمی**

چکیده:

با توجه به افزایش کاربرد مدل‌های چند سطحی، تعیین تعداد نمونه از دغدغه‌های اصلی محققان این حوزه است. لذا هدف اصلی این پژوهش بررسی حجم نمونهٔ بهینه در مدل‌های دو سطحی است. بدین منظور از داده‌های تیمز (۲۰۱۱) در پایهٔ هشتم استفاده شد که تعداد آن‌ها ۶۰۲۹ دانش‌آموز از ۲۳۸ مدرسه است. متغیر وابستهٔ این پژوهش پیشرفت ریاضیات دانش‌آموزان پایهٔ هشتم و متغیر مستقل در سطح دانش‌آموز، متغیر مدت زمان انجام تکلیف است و در سطح مدرسه-معلم هم مدت زمانی است که معلم برای انجام تکلیف به دانش‌آموز می‌دهد. سه تحلیل دو سطحی با سه حجم متفاوت ۵، ۵۰ و ۲۳۸ نمونه در نظر گرفته شد و اندازهٔ نمونه بر اثرهای ثابت و تصادفی بررسی شد. نتایج نشان داد، با افزایش حجم نمونه در سطح دوم، توان آزمون افزایش و خطای برآورد نیز کاهش می‌یابد. همچنین، با افزایش نمونه، انحراف استانداردها افزایش می‌یابد و پراکندگی به حداکثر خود می‌رسد. لذا به‌منظور داشتن براوردهای دقیق‌تر از کواریانس بین خطاها، افزایش تعداد گروه‌های مورد مطالعه در سطح دوم توصیه می‌شود.

کلید واژه‌ها: **حجم نمونه، مدل‌سازی چند سطحی، تیمز، اثرات ثابت، اثرات تصادفی**

■ تاریخ دریافت مقاله: ۹۸/۱۰/۶ ■ تاریخ شروع بررسی: ۹۸/۱۱/۲۸ ■ تاریخ پذیرش مقاله: ۹۹/۴/۲۹

* استادیار گروه روان‌شناسی تربیتی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، ایران (نویسنده مسئول)..... z.naghsh@ut.ac.ir

** استادیار گروه روان‌شناسی تربیتی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه الزهرا، ایران..... z.hashemi@alzahra.ac.ir

مقدمه

کاربرد مدل‌های چند سطحی در علوم اجتماعی و روان شناسی، برای تحلیل داده‌های آشیانه‌ای یا سلسله‌مراتبی، به سرعت در حال افزایش است، چرا که ساختار داده‌ها در علوم اجتماعی و علوم تربیتی غالباً سلسله‌مراتبی[1] است. بدین معنا که متغیرهایی که در سطح فردی وجود دارند، غالباً درون واحدهای بزرگ‌تری گروه‌بندی شده‌اند. برای مثال، در پژوهش‌هایی که در حوزهٔ آموزش و پرورش صورت می‌گیرند، دانش‌آموزان درون کلاس‌ها گروه‌بندی می‌شوند. برخی متغیرها ویژگی‌های دانش‌آموزان را توصیف می‌کنند و برخی توصیف‌کنندهٔ ویژگی‌های کلاس هستند. متغیرهایی که ویژگی دانش‌آموزان را بررسی می‌کنند، معمولاً به‌عنوان متغیرهای مربوط به سطح یک طبقه‌بندی می‌شوند و متغیرهای توصیف‌کنندهٔ کلاس یا معلم یا مدرسه، متغیرهای سطوح بالاترند. متغیرهای توصیف کنندهٔ کلاس ممکن است تجمیع[2] شدهٔ متغیرهای سطح دانش‌آموز باشند یا می‌توانند متغیرهایی باشند که از سطح کلاس به دست آمده‌اند (نقش، ۱۳۹۶). انواع مدل‌های چند سطحی وجود دارند که می‌توانند بر حسب تعداد سطوح (مثلاً دو یا سه سطح)، نوع طرح (مثلاً متقاطع[3] یا طولی با اندازه‌گیری مکرر[4])، مقیاس متغیر وابسته (مثلاً پیوسته یا طبقه‌ای) و تعداد متغیرهای وابسته (تک متغیره یا چند متغیره) متفاوت باشند. این مدل‌ها برای پاسخگویی به انواع سؤالات تحقیق به کار می‌روند و شامل پارامترهای اثرات ثابت و ضرایب تصادفی سطح یک و مؤلفه‌های واریانس-کواریانس هستند.

یکی از موضوعات مهم در مدل‌های چند سطحی، تعیین حجم نمونه است. نمونه‌گیری در این تحلیل‌ها چندمرحله‌ای[5] است. یک نمونه از سطح بالاتر (مثلاً مدرسه یا سازمان) و نمونهٔ دیگر از سطح پایین‌تر (مثلاً دانش‌آموزان یک مدرسه یا کارکنان یک سازمان) است و برای هر سطح حجم نمونهٔ بهینه مطرح است. مثلاً در یک مطالعهٔ سه سطحی، دانش‌آموزان درون کلاس‌ها آشیانه شده و کلاس‌ها درون مدرسه‌ها. و مشاهدات ممکن است شامل ۶۰ مدرسه، ۱۵۰ کلاس و ۳۳۰۰ دانش‌آموز باشند؛ یعنی به طور متوسط هر کلاسی با ۲۲ دانش‌آموز و هر مدرسه با ۲/۵ کلاس، حجم نمونه را تشکیل می‌دهد. از این رو، داشتن نمونهٔ کافی[6] یکی از مهم‌ترین مسائل در زمینهٔ مدل‌سازی چند سطحی است.

پژوهشگران زیادی از جمله مس و هاکس[7] (۲۰۰۴ و ۲۰۰۵) و اسنیجرز[8] (۲۰۰۵) به اهمیت حجم نمونه در هر سطح تحلیل، به منظور به‌دست آوردن محاسبات دقیق و بدون تورش از ضرایب رگرسیون، خطای استاندارد، و نیز توان آزمون اشاره کرده‌اند. روش حداکثر احتمال (ML) که معمولاً در تحلیل چند سطحی از آن استفاده می‌شود، با این فرض است که اندازهٔ نمونه به اندازهٔ کافی بزرگ باشد. از یک طرف، مشکل اصلی در تحلیل‌های چند سطحی، معمولاً اندازهٔ نمونه در سطح گروه است و از طرف دیگر حجم نمونه در سطح گروه به طور متوسط کوچک‌تر از حجم نمونه در سطح فردی است و افزایش تعداد گروه‌ها در این تحلیل‌ها مشکل است (مس و هاکس، ۲۰۰۵). لذا تعیین حجم نمونهٔ بهینه و کافی در سطح دوم اهمیت بسیار زیادی دارد. حجم نمونهٔ کافی و مناسب را می‌توان به صورت حجم

کمینه که تضمین‌کنندهٔ عدم تورش (یا به عبارت دقیق‌تر، حجم پایین مورد قبول تورش) باشد، تعریف کرد. چنین تعریفی با توصیف اسنیجرز و باسکر[۹] (۱۹۹۳) سازگار است که از اصطلاح «بهینه از نظر شرایط[۱۰]» برای توضیح حجم نمونه استفاده کرده‌اند و براساس آن می‌توان به حداقل خطای استاندارد برای برآورد پارامترهای ثابت و تصادفی دست یافت. اگرچه مطالعات قبلی در زمینهٔ تعیین حجم نمونه در مدل‌های چند سطحی بسیار است، ولی با وجود این، هنوز اتفاق نظر در زمینهٔ اینکه حجم نمونه در شرایط گوناگون شبیه‌سازی چه اندازه باید باشد، وجود ندارد (مس و هاکس، ۲۰۰۵). در ادامه دستورالعمل‌ها و راهنمایی‌های مربوط به مدل‌های دو سطحی با استفاده از حجم متوازن مرور می‌شود.

پژوهشگران زیادی از جمله گلداستین و سیلور[۱۱] (۱۹۸۹) و کوهن[۱۲] (۱۹۹۸) در حوزه‌های گوناگون علوم در زمینهٔ تعیین اندازهٔ نمونه در مدل‌های چند سطحی فعالیت کرده‌اند (لانگفورد[۱۳]، ۱۹۸۷). به علاوه، بر اساس انحراف معیار پارامترهای یک مدل دو سطحی، اسنیجرز و باسکر (۱۹۹۳) نیز فرمول‌هایی برای تعیین اندازهٔ نمونه و ارتباط آن با توان آزمون‌های مربوط به پارامترهای مدل ارائه کرده‌اند. از طرف دیگر، افشار توس[۱۴] (۱۹۹۵) با استفاده از الگوریتم خودگردان‌سازی[۱۵]، به ارزیابی تأثیر اندازهٔ نمونه در سطح دوم، در شرایطی که حجم نمونه در سطح اول کم است، پرداخته است. براون، گلالیزاده و پارکر[۱۶] (۲۰۰۹) نیز به کمک شبیه سازی ترکیبات نمونه‌ای متفاوت، اما بدون لحاظ کردن عامل هزینه، موضوع تعیین اندازهٔ نمونه را برای انواع مدل‌های چندسطحی مطالعه کرده و نرم افزاری را نیز در این زمینه تولید کرده‌اند. رویکرد اکثر فعالیت‌های صورت گرفته در زمینهٔ تعیین اندازهٔ نمونهٔ بهینه در مدل‌های چندسطحی برای تأثیر اندازهٔ نمونه‌های مختلف و برآورد پارامترهای مدل با استفاده از شبیه‌سازی بوده است.

نتایج پژوهش‌های صورت گرفته منجر به وضع قوانینی برای تعیین حجم نمونه شد. یکی از این قوانین، قانون سرانگشتی است که حداقل ۳۰ واحد را در هر سطح تحلیل مناسب می‌داند. قانون سرانگشتی را افرادی از قبیل هاکس (۱۹۸۸) مس و هاکس(۲۰۰۲) مس و هاکس(۲۰۰۴) مطرح کرده‌اند. کرفت[۱۷] (۱۹۹۶) قانون و اصل ۳۰/۳۰ را توصیه می‌کند که به معنای حداقل ۳۰ مشاهده در هرگروه و حداقل ۳۰ واحد در هر سطح تحلیل برای برآورد بدون تورش تمام پارامترها و خطاهای استانداردشان است. مس و هاکس (۲۰۰۵) نیز بیان می‌کنند، استفاده از قانون ۳۰/۳۰ به جز در برآورد خطای استاندارد اثرات تصادفی، بدون تورش و مفید است. هاکس (۱۹۹۸) حجم بهینهٔ نمونه را برای مدل‌هایی که اثرات تقاطعی[۱۸] را بررسی می‌کنند، به حداقل ۲۰ مشاهده (برای سطح ۱) و ۵۰ گروه (برای سطح ۲) توسعه داد. اگرچه برای رسیدن به نتایج بدون تورش، هم تعداد گروه‌ها و هم تعداد مشاهدات در هر گروه از اهمیت زیادی برخوردار است، حساسیت تأثیرات تصادفی و ثابت[۱۹] (و خطاهای استانداردآن‌ها) متفاوت است و در مواقعی که دقت برآوردهای مؤلفه‌های واریانس به‌شدت تحت تأثیر تعداد گروه‌ها قرار می گیرد، برآوردهای تأثیرات ثابت حساسیت کمتری نسبت به پراکندگی

داده نشان می‌دهد (نیوسام و نیشیشیبا[20]، ۲۰۰۲؛ کلارک و ویتون[21]، ۲۰۰۷). نتایج مشابهی توسط نیوسام و نیشیشیبا (۲۰۰۲) و کلارک و ویتون (۲۰۰۷) به‌دست آمده که همگی برآوردهای بدون تورش تأثیرات ثابت را حتی برای نمونه‌های کوچک تأیید کرده‌اند. از آنجاکه برآوردهای مؤلفه‌های واریانس غالباً در مرکز توجه و علاقه در مدل‌های چند سطحی هستند، پیشنهادات بیشتر در مورد تأثیرات تصادفی مورد توجه بیشتر قرار گرفت. ماک[22] (۱۹۹۵) مشاهده کرد که پنج گروه در سطح دوم، تورش قابل توجهی از واریانس را باعث می‌شوند. این درحالی است که کلارک و ویتون (۲۰۰۷) پیشنهاد کرده‌اند، حداقل ۱۰ مشاهده در هر گروه برای حداقل ۱۰۰ گروه نیاز است. آن‌ها توصیه می‌کنند، اگر واریانس شیب محاسبه شود، حداقل ۲۰۰ گروه با حداقل ۲۰ مشاهده در هرگروه مورد نیاز است. اگرچه برای محاسبه دقیق مؤلفه‌های واریانس (اغلب کمتر برآورد شده) حداقل ۱۰۰ واحد مورد نیاز است، در عمل به‌دست آوردن چنین نمونه‌ای مشکل است (مس و هاکس ۲۰۰۴). با توجه به آنچه گفته شد، بر اساس تمامی راهنمایی‌های ذکر شده، بیان می‌کنند که به‌جای تعداد بالای مشاهدات در هر واحد، به‌نظر می‌رسد بالابودن تعداد گروه‌ها برای رسیدن به برآوردهای دقیق‌تر اهمیت بیشتری دارد.

در کنار مطالعاتی که برای برآورد حجم بهینه به بررسی تعداد گروه‌ها پرداخته‌اند، مطالعات شبیه سازی زیادی نیز بررسی اثر حجم نمونهٔ کوچک بر نتایج مدل‌های چند سطحی، از قبیل برآوردهای واریانس، برآوردهای اثرات ثابت، خطای استاندارد و همگرایی در سطوح گوناگون را در کانون توجه خویش قرار داده‌اند. ماک (۱۹۹۵) در یک مطالعهٔ شبیه سازی، اثر حجم نمونه را بر برآوردهای واریانس بررسی کرده است. وی دریافت که برآوردهای واریانس در طرح‌های با تعداد کم نمونه (تعداد واحد کمتر) در واحدهای سطح ۲ تورش‌دار هستند. کلارک و ویتون (۲۰۰۷) در مطالعهٔ مونت کارلو[23] بر مدل سطح ۲ تأکید کردند و شرایط با تعداد ۵۰ واحد سطح ۲ و ۲۰۰ نفر در سطح ، یعنی نسبت ۱ به ۴ را بررسی کردند. آن‌ها دریافتند، تورش مثبتی در تفسیر و برآوردهای واریانس شیب وجود دارد. آن‌ها اشاره کردند که حداقل ۱۰ مشاهده برای هر گروه و حداقل ۱۰۰ گروه، به منظور برآورد واریانس عرض از مبدأ و به دست‌آوردن ارزش‌های درست نیاز است. برای واریانس شیب حداقل ۲۰ مشاهده برای هر گروه و حداقل ۲۰۰ گروه مورد نیاز است و نسبت ۲ به ۲۰ را پیشنهاد دادند.

مس و هاکس (۲۰۰۴، ۲۰۰۵) به بررسی شرایط نمونهٔ ۳۰ به ۱۰۰ در مقایسه با نسبت تعداد ۵ به ۵۰ پرداختند. آن‌ها دریافتند، در شرایط ۳۰ به ۱۰۰ تورش کمتری در برآوردهای واریانس وجود دارد، ولی دشواری‌های مربوط به نتیجه‌گیری در مورد مؤلفه‌هایی که تعداد واحدهای سطح ۲ آن‌ها فقط ۳۰ تا بود، دشوار است.

حجم نمونهٔ کافی در دقت خطای استاندارد نیز اهمیت زیادی دارد؛ اگرچه بررسی‌های اندکی در این زمینه انجام شده‌اند (مس و هاکس، ۲۰۰۵). در تحقیق شبیه سازی، معمول‌ترین راه برای اعتبار سنجی تخمین خطای استاندارد، چک کردن دقت آزمون معناداری یا پوشش بازهٔ فاصلهٔ اطمینان است (که

با استفاده از توزیع نرمال و استاندارد و نیز توزیع گاما انجام می‌شود). بر این اساس، براون و دراپر[۲۴] (۲۰۰۰) با استفاده از برآوردهای IGLS و RIGLS نشان دادند، فاصلهٔ اطمینان ۹۵ درصدی برای حداقل ۴۸ گروه بدون تورش است (کمتر از ۹۵ درصد).

به همین نحو، مس و هاکس (۲۰۰۵) نیز گزارش دادند، تأثیر منفی ۳۰ گروه برای خطاهای استاندارد ضرایب تأثیر ثابت ناچیز است (۶ و ۶/۴ درصد برای ضرایب رگرسیون و محور مختصات) و برای خطاهای استاندارد شاخص واریانس بیشتر است (حدود ۹ درصد برای محور مختصات دو سطحی و واریانس‌های شیب). همچنین، در یک آزمایش بسیار گسترده، مونت کارلو (۵۷۶۰ شرایط) بل، مورگان، کرومرو و فرون[۲۵] (۲۰۱۰) متوجه شدند، برای هر نوع متغیر پیش‌بینی‌کننده (که به‌عنوان یک اثر ثابت در نظر گرفته می‌شود)، فاصلهٔ اعتماد برآورد شده تقریباً ثابت و از مقدار مربوط به برآوردهای دو سطحی بیشتر است.

در زمینهٔ تأثیر منفی پراکندگی داده‌ها بر همگرایی، در بین پژوهشگران توافق وجود ندارد. اگرچه بل و همکاران (۲۰۱۰) و نیز مس و هاکس (۲۰۰۴) به این نتیجه رسیدند که مشکلی در زمینهٔ همگرایی مدل با استفاده از برآوردکننده‌های ML و RIGLS وجود ندارد، ولی براساس نتایج پژوهش باسینگ[۲۶] (۱۹۹۳)، در صورت کوچک بودن حجم نمونه، مشکلاتی در تعمیم نتایج داده‌ها و برآورد اثرات (ثابت، تصادفی، متقابل و..) یا روش برآورد به وجود خواهد آمد.

همان‌گونه که اشاره شد، در مطالعات پیشین، برای تعیین حجم نمونهٔ کافی و بهینه در مدل‌سازی چند سطحی، غالباً از روش شبیه سازی[۲۷] استفاده شده است. روش‌های دیگر برای انجام چنین هدفی، استفاده از فرمول تقریب[۲۸]، نسبت اندازهٔ تأثیر و خطاهای استاندارد به توان آماری آزمون معناداری (اسنیجرز و باسکر ۱۹۹۳) است. چرا که توان آماری[۲۹] به حجم نمونه و جنبه‌های دیگر طراحی از جمله اندازهٔ اثر، ارزش‌های پارامتر و سطح معناداری بستگی دارد.

همان‌طور که در مطالعهٔ اسنیجرز (۲۰۰۵) نشان داده شده است، روش محاسبهٔ حجم نمونهٔ کافی و مناسب به تخمین‌هایی پارامتری بستگی دارد که محقق به آن علاقه‌مند است. موربیک، ون برکلن و برگر[۳۰] (۲۰۰۱) چنین فرمول‌هایی را برای محاسبهٔ طراحی و طرح بهینه (حجم نمونه) برای مدل‌های دو سطحی ارائه دادند که در آن‌ها از معیار بهینه‌سازی دی و آل[۳۱] که از جمله الگوریتم‌های تکراری برای یافتن طرح‌های بهینه[۳۲] هستند، استفاده شده است. اگرچه به‌نظر می‌رسد که فرمول تقریب سریع‌تر است، ولی محدودیت‌های آن (از قبیل عدم تعمیم) باعث شده که روش مونت کارلو برای محاسبهٔ کفایت حجم نمونه روشی منعطف‌تر باشد.

با توجه به آنچه مطرح شد، به دلیل ماهیت مدل‌های چندسطحی ،تعیین اندازهٔ نمونه به اندازهٔ آن در سطوح گوناگون وابسته است. برای مثال اگر بخواهیم تحت استانداردهای بین‌المللی و با کمک یک مدل دوسطحی، تعدادی داده‌های آموزشی جمع کنیم، نیازمند تعیین اندازهٔ نمونه بر اساس دو سطح

هستیم: تعداد مدرسهٔ لازم در بررسی نمونه در سطح دوم، و تعداد دانش‌آموزان مورد نیاز در هر مدرسه در سطح اول.

به اعتقاد پژوهشگران این حوزه، در مدل‌های با ساختار سلسله‌مراتبی، داشتن واحدهای تا حد ممکن زیاد در بالاترین سطح اهمیت بیشتری دارد (افشارتوس، ۱۹۹۵). از طرف دیگر، همچنان که اشاره شد، چالش دیگری که در برآورد اندازهٔ حجم نمونه مطرح می‌شود، این است که تعیین اندازهٔ نمونه به نوعی به هدف اصلی مدل‌بندی نیز وابسته است (اسنیجر و باسکر، ۱۹۹۳). به علاوه در بعضی از مسائل تعیین اندازهٔ نمونه، محقق به مطلوبیت برآورد پارامتر، قرار گرفتن برآورد در نواحی خاص یا محدودیت‌های بیزی علاقه مند است. از این رو، مقالهٔ حاضر با انجام شبیه‌سازی به صورت نمونه یری در یک مدل دو سطحی، تأثیر تعداد واحدهای سطح دوم را روی برآورد اثرهای ثابت و تصادفی ارزیابی می‌کند. هدف این پژوهش ارائهٔ نحوهٔ برازش مدل و ارزیابی کارایی آن نیست، بلکه هدف اصلی ارزیابی برآورد پارامترهای اثرات ثابت، خطای استاندارد، توان آزمون، و مؤلفه‌های واریانس_کواریانس با توجه به اندازهٔ نمونه‌های مختلف از اندازهٔ کل نمونه‌ها در هر سطح اول و دوم است. به عبارت دیگر، با تفسیر ترکیب نمونه‌ای متفاوت از سطوح مختلف مدل دو سطحی، نحوهٔ تأثیر آن‌ها را در برآورد پارامترهای مدل، مطالعه خواهیم کرد. می‌توان این هدف را به طریق دیگر و با طرح سؤالی به صورت زیر بررسی کرد: آیا تعداد نمونه‌های در نظر گرفته شده در هر مدل در برازش مدل کافی بوده است؟

■ روش پژوهش

جمعیت نمونهٔ این پژوهش را دانش‌آموزان ایرانی پایهٔ هشتم در سال تحصیلی ۹۰-۱۳۸۹ تشکیل می‌دهند که در مطالعهٔ تیمز ۲۰۱۱ شرکت کرده‌اند. تعداد آن‌ها ۶۰۲۹ دانش‌آموز از ۲۳۸ مدرسه است. تیمز ۲۰۱۱ شبیه به مطالعات قبلی تیمز (۱۹۹۵، ۱۹۹۹ و ۲۰۰۳ و ۲۰۰۷)، برای اطمینان از اینکه داده‌های نمونه، معرف جامعهٔ دانش‌آموزان ملی است، از روش نمونه‌گیری خوشه‌ای طبقه‌ای دو مرحله‌ای[۳۳] استفاده می‌کند (به این معنا که سهم هر یک از خوشه‌ها[۳۴] در نمونه، با حجم آن در جامعه متناسب است). افزون بر آن، با بهره‌گیری از وزن‌های نمونه‌گیری[۳۵] اطمینان حاصل می‌شود که شاخص‌های آماری به‌دست‌آمده از نمونه، معرف جامعهٔ مورد نظر هستند. در مرحلهٔ اول، مدرسه‌ها با روش احتمال متناسب با حجم[۳۶] نمونه گیری شدند. سپس درون هر مدرسهٔ منتخب، از میان همهٔ کلاس‌های پایهٔ هشتم، یک کلاس با روش تصادفی سیستماتیک انتخاب شدند و در نهایت همهٔ دانش‌آموزان با احتمال مساوی از کلاس‌های نمونه‌گیری‌شده در آزمون شرکت کردند. در ایران در سطح مدرسه یا سطح ۲، تنها یک کلاس از هر مدرسه انتخاب شد. لذا تعداد مدرسه‌ها مساوی با تعداد کلاس‌هاست. تعداد نمونهٔ تیمز ۲۰۱۱ از ۶۰۲۹ دانش‌آموز (۲۸۱۶ دختر و ۳۲۱۳ پسر) تشکیل شده است (پژوهشگاه مطالعات

آموزش و پرورش، ۱۳۹۱). در این پژوهش، سه تحلیل با سه حجم نمونهٔ متفاوت در سطح دوم استفاده شـد. در حالت اول، ۲۳۸ مدرسـه یعنی مجموع مدرسه‌های تحلیل تیمز و در حالت‌های دوم و سوم با استفاده از الگوریتم شبیه‌سازی، نمونه‌ای از بین این ۲۳۸ مدرسه انتخاب و حجم نمونه‌ای برابر با ۵ و ۵۰ انتخاب شده است.

تیمـز برای جمع‌آوری اطلاعات در مورد زمینه‌های آموزشـی بـرای تدریس و یادگیری ریاضیات و علوم، سـه پرسـش‌نامه دارد: پرسش‌نامهٔ دانش‌آموز، پرسـش‌نامهٔ معلم و پرسش‌نامهٔ مدرسه. در این پژوهش، به منظور جمع‌آوری اطلاعات، از پرسـش‌نامهٔ دانش‌آموز و معلم اسـتفاده شـده است. متغیر وابستهٔ این پژوهش پیشرفت ریاضیات دانش‌آموزان پایهٔ هشتم است. متغیر مستقل در سطح دانش‌آموز متغیر مدت زمان انجام تکلیف و در سطح مدرسه/معلم مدت زمانی است که معلم برای انجام تکلیف به دانش‌آموز می‌دهد.

محاسـبهٔ اعتبار[۳۷] و روایی[۳۸] سـؤال‌های تیمز براساس شـاخص‌های روان‌سنجی از طریق انجام آزمون‌هـای مقدماتی[۳۹] در کشـورهای شـرکت‌کننده انجام می گیرد و پس از تعیین درجهٔ دشـواری و قدرت تشخیص برای هر یک از سؤال‌های چند گزینه‌ای و پاسخ باز، به تفکیک هر یک از کشورها، در قالب گزارش آماری منتشر می‌شود و در اختیار کشورهای عضو قرار می گیرد. پس از تجزیه و تحلیل شـاخص‌های روان‌سنجی، سؤال‌هایی که از نظر اعتبار و روایی شرایط لازم را نداشته باشند، حذف و سـؤال‌های دیگر جایگزین می‌شوند. بنابراین، شاخص‌های آماری مربوط به روایی و اعتبار سؤال‌های تیمز برای تمام کشورهای شرکت‌کننده از جمله ایران محاسبه می‌شود. در این پژوهش، از تحلیل چند سطحی (دو سطحی) با استفاده از نرم افزار HLM استفاده شد. در تیمز، علاوه بر وزن‌های نمونه‌گیری در سطح دانش‌آموز، چندین وزن نمونه‌گیری در سطح کلاس و مدرسه وجود دارد. از جمله وزن معلم ریاضیـات[۴۰] (MATWGT) و وزن معلم علوم[۴۱] (SCIWGT) هر دو از وزن‌های نمونه‌گیری مهم در سطح کلاس هستند. در این پژوهش از MATWGT استفاده شده است.

■ یافته‌ها

در مطالعهٔ حاضر از سه تحلیل با سه حجم متفاوت نمونه استفاده شد. در حالت اول، کل نمونهٔ سطح دوم (۲۳۸ مدرسه) مورد استفاده قرار گرفته است و در حالت‌های دوم و سوم با استفاده از الگوریتم شبیه سـازی، به انتخاب نمونه از بین این ۲۳۸ مدرسـه پرداخته شـد. به این ترتیب که: فرض کنیم i امین زیر نمونه عبارت است از نمونه‌گیری بدون جایگذاری از یک مجموعهٔ $\{X_{۱}...X_{n}\}$ با هدف تشکیل یک زیرنمونه به اندازهٔ b، آنگاه تعداد زیر نمونه‌های ممکن عبارت است از:

$$C(b, n) = \binom{n}{b}$$

که معمولاً عددی بسیار بزرگ است. از این تعداد زیر نمونه، تعدادی به صورت تصادفی انتخاب می‌شوند و نمونهٔ مورد نظر برای انجام تحلیل در نظر گرفته می‌شود. چنین رویکردی به انتخاب نمونه به روش زیرنمونه‌گیری[۴۲] معروف است. در روش زیرنمونه‌گیری، با استفاده از نمونه‌گیری بدون جایگذاری، اقدام به نمونه‌گیری می‌شود (پولیتیس، رومانو و وولف[۴۳]، ۱۹۹۹). در این مطالعه، زیرنمونه‌هایی به اندازهٔ ۵، ۵۰، ۲۳۸ در نظر گرفته شدند. نتایج حاصل، همراه با تفسیر آن‌ها، به تفصیل در ادامه خواهد آمد. برای دسته‌بندی مناسب، نتایج به دو دستهٔ تأثیر اندازهٔ نمونه بر اثرهای ثابت و تصادفی تقسیم شده‌است.

بررسی تأثیر اندازهٔ نمونه بر برآورد اثرهای ثابت: اثرهای ثابت مدل عبارت‌اند از: ضرایب متغیرهای زمان انجام تکلیف و مدت زمان انجام تکلیف در هر دو سطح. در جدول ۲، میانگین برآورد پارامترهای مربوط به ضرایب متغیرهای تبیینی مقدار تکلیف در هفته و مدت زمان انجام تکلیف و همچنین متوسط مقدار تکلیف در هفته و مدت زمان انجام تکلیف آورده شده است. خطای برآورد و در نهایت توان آماری برای ترکیب اندازهٔ نمونه‌ها مختلف است.

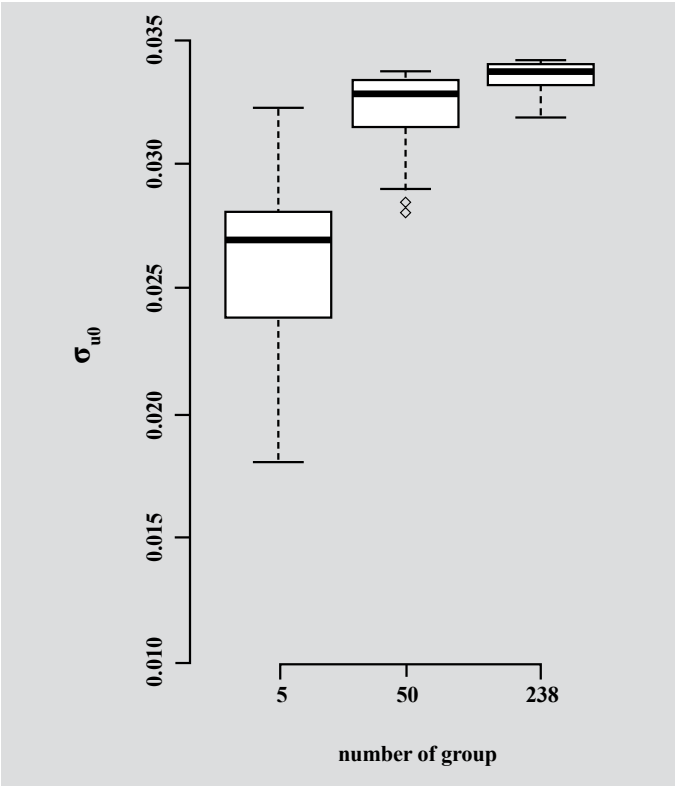
جدول ۱. میانگین، خطای برآورد و توان آزمون برای اندازهٔ گروه‌های سطح دوم

سطح	متغیر	تعداد گروه‌ها (واحدهای سطح دوم)	میانگین	خطای برآورد	توان
اول	مدت زمان انجام تکلیف	۵	۳/۲۳	۰/۰۱۱۲	۰/۷۶
		۵۰	۳/۴۰	۰/۰۰۳	۰/۹۲
		۲۳۸	۳/۵۴	۰/۰۰۱	۰/۹۵
دوم	مدت زمان انجام تکلیف	۵	۲/۷۵	۰/۰۱۲	۰/۸۲
		۵۰	۲/۶۴	۰/۰۰۶	۰/۹۴
		۲۳۸	۲/۶۷	۰/۰۰۴	۰/۹۹

همان‌گونه که در جدول ۱ ملاحظه می‌شود، با افزایش اندازهٔ نمونه در سطح دوم، توان آزمون افزایش می‌یابد و در نهایت به یک میل می‌کند. به علاوه، با افزایش اندازهٔ نمونه، خطای برآورد نیز کاهش می‌یابد.

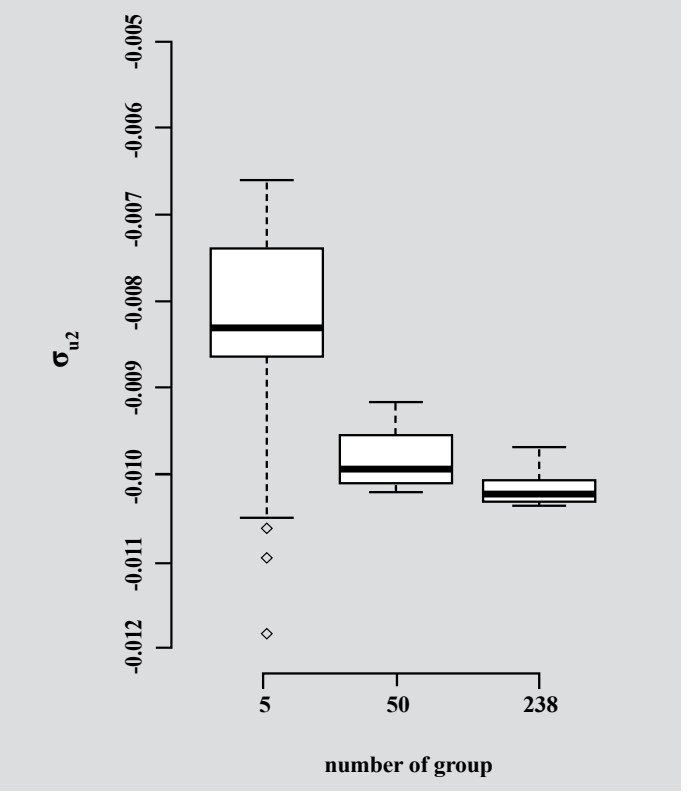
بررسی تأثیر اندازهٔ نمونه بر برآورد اثرهای تصادفی: در فرایند زیرنمونه‌گیری که در بخش قبل توصیف شده است، بررسی تأثیر تعداد واحدهای سطح دوم بر برآورد اثرهای تصادفی نیز مدنظر

قرار گرفته‌است. اثرهای تصادفی مدل، واریانس خطای نمونه‌گیری گروه‌های منتسب به سطح دوم، از تغییرات عرض از مبدأ و ضریب میانگین ساعت مطالعه و کواریانس بین این دو خطا ناشی می‌شوند.



شکل ۱. نمودار جعبه‌ای اثرات تصادفی سطح دوم

شکل ۱ نمودار جعبه‌ای انحراف استانداردهای واریانس خطای نمونه‌گیری گروه‌هاست. در محور افقی، تعداد گروه‌های در نظر گرفته‌شده برای هر زیرنمونه‌گیری و در محور عمودی مقدار انحراف استاندارد متناظر با هر یک نشان داده شده‌است. همان طور که مشاهده می‌شود، با افزایش تعداد گروه‌ها مقدار انحراف استانداردها افزایش و پراکندگی آن‌ها کم می‌شود. همان‌طور که مشاهده می‌شود، افزایش انحراف استانداردها و کاهش پراکندگی در مورد کل نمونهٔ برآورد شده، یعنی ۲۳۸ نمونه، به حداکثر خود رسیده است. ولی در نمونه‌های کمتر در سطح دوم برآورد کمتر اثرات تصادفی، یعنی انحراف استاندارد، ملاحظه می‌شود. لذا به منظور داشتن برآوردهای دقیق‌تر، افزایش تعداد گروه‌های مورد مطالعه در سطح دوم توصیه می‌شود.



شکل ۲. نمودار جعبه‌ای اثرات تصادفی سطح دوم (کواریانس بین خطاها)

شکل ۲ نمودار جعبه‌ای را برای برآورد کواریانس بین خطاها بر اساس روش زیرنمونه‌گیری از تعداد گروه‌ها نشان می‌دهد. در محور افقی، تعداد گروه‌های در نظر گرفته شده برای هر زیرنمونه‌گیری و در محور عمودی مقدار کواریانس متناظر با هر یک نشان داده شده‌است. ملاحظه می‌شود که با اندازهٔ نمونهٔ کم در سطح دوم ابتدا مشکل بیش برآورد کواریانس اتفاق می افتد. اما با افزایش نمونه، برآورد کواریانس به مقدار واقعی نزدیک‌تر می‌شود. لذا به منظور داشتن برآوردهای دقیق‌تر از کواریانس بین خطاها، افزایش حجم نمونه در سطح دوم توصیه می‌شود.

■ بحث و نتیجه‌گیری ■

هدف از این پژوهش بررسی تأثیر حجم نمونه‌های مختلف برای سطح‌های اول و دوم در یک مدل شبیه‌سازی دو سطحی در برآوردهای اثرات ثابت، خطای استاندارد، توان آزمون، و مؤلفه‌های واریانس_کواریانس بوده است. با توجه به آنچه به تفصیل بیان شد،

چالش‌های موجود در این حوزه را می‌توان در سه بخش کلی خلاصه کرد: ۱. رویکردهای مختلف تعیین اندازهٔ حجم نمونه (روش محاسبهٔ اندازهٔ حجم نمونه)؛ ۲. توجه به تفاوت تأثیر اندازهٔ حجم نمونه در هر سطح در برآورد پارامترها؛ ۳. تأثیر متفاوت حجم نمونه در برآورد انواع پارامترها. به‌طور کلی، رویکردهای متفاوتی در برآوردهای پارامترهای مختلف مطرح شده‌اند. در این زمینه، مطالعاتی که برای تعیین حجم نمونهٔ کافی و بهینه انجام شده است، یا از طریق مدل‌های شبیه‌سازی شده صورت گرفته و یا روش‌های دیگری برای تعیین حجم نمونه به کار رفته‌اند. برای مثال، لانگفورد (۱۹۸۷)، گلدستاین و سیلور[۴۴] (۱۹۸۹) و کوهن (۱۹۹۸) با استفاده از فرمول تقریب (اسنیجرز و باسکر، ۱۹۹۳)، بر اساس قانون سرانگشتی ۳۰/۳۰ (کرفت، ۱۹۹۶؛ هاکس، ۱۹۹۸؛ مس و هاکس، ۲۰۰۲؛ مس و هاکس، ۲۰۰۴) و بر اساس انحراف معیار پارامترهای مدل دوسطحی و معادله هزینه (اسنیجر و بوسکر، ۱۹۹۳)، با استفاده از الگوریتم خودگردان‌سازی (افشار توس، ۱۹۹۵) و استفاده از معیار بهینه‌سازی دی و آل (موربیک و همکاران،۲۰۰۱)، به کمک شبیه‌سازی ترکیبات نمونه‌ای متفاوت و بدون لحاظ‌کردن عامل هزینه (براون و همکاران، ۲۰۰۹). از میان روش‌های فوق، این پژوهش بر یک مدل شبیه‌سازی استوار بود که با توجه به چالش دوم مبنی بر اهمیت سطح دوم در برآورد بدون تورش ضرایب اثرات تصادفی و واریانس شیب (افشارتوس، ۱۹۹۵؛ اسنیجر، ۲۰۰۵)، سه تحلیل جداگانه در سه اندازهٔ متفاوت تعداد گروه‌ها در سطح دوم (۵، ۵۰، و ۲۳۸) انجام گرفته است.

با وجود اهمیت تعداد مشاهدات و تعداد گروه‌ها برای رسیدن به نتایج بدون تورش، همان‌طور که در مطالعهٔ اسنیجرز (۲۰۰۵) نشان داده شده است، روش محاسبهٔ حجم نمونهٔ کافی و مناسب، به تخمین‌های پارامتری بستگی دارد که محقق به آن علاقه‌مند است، چرا که حساسیت تأثیرات تصادفی و ثابت (و خطاهای استانداردشان) متفاوت است. با وجود اینکه دقت برآوردهای مؤلفه‌های واریانس به‌طور قوی تحت تأثیر تعداد گروه‌ها قرار می گیرد، برآوردهای تأثیرات ثابت حساسیت کمتری نسبت به پراکندگی داده نشان می‌دهد (نیوسام و نیشیشیبا، ۲۰۰۲؛ کلارک و ویتون، ۲۰۰۷).

نتایج این پژوهش در برآورد اثرات ثابت نیز نشان می‌دهد که با افزایش اندازهٔ نمونه در سطح دوم، توان آزمون افزایش می‌یابد و در نهایت به یک میل می‌کند. به علاوه، با افزایش اندازهٔ نمونه، خطای برآورد نیز کاهش می‌یابد. این موضوع نشان دهندهٔ تأثیر مستقیم تعداد واحدهای سطح دوم یا به عبارت دیگر تعداد گروه‌های مورد بررسی بر توان آزمون است. در واقع، یکی از دلایل افزایش توان در مقابل افزایش تعداد واحدهای سطح دوم همین کاهش خطای برآورد است. از طرف دیگر، با توجه به اینکه مقادیر حاصل از

زیرنمونه‌گیـری برای ارزیابی اثر پارامتر ثابت متغیرها به برآورد اثر تحت داده‌های اصلی بسیار نزدیک و خطای برآورد آن‌ها بسیار کوچک است، می‌توان پایا بودن این اثر ثابت را تحت اندازهٔ نمونه‌های مختلف نتیجه گرفت. لازم به ذکر است، متغیر چه در سطح اول و چه در سطح دوم باشد، عملکردی تقریباً مشابه دارد. این موضوع به محقق کمک می‌کند اگر به دنبال آزمونی با توان حداقل ۹۰درصد است، به جای ۹۹ گروه می‌تواند به حداقل ۵۰ گروه اکتفا کند. این نتایج در راستای نتایج پژوهش‌های کلارک و ویتون (۲۰۰۷) است. در نهایت، نتایج این پژوهش نشان می‌دهد، همچنان‌که اسنیجر (۲۰۰۵) بیان می‌کند، حجم گروه نسـبت به تعداد گروه‌ها در زمینهٔ قدرت آزمون اهمیت کمتری دارد که با یافته‌های مطالعات مربوطه متناسـب است. تنها محدودیت حجم نمونهٔ کوچک برای توان آزمون، همان واریانس‌های شیب تصادفی است.

با توجه به آنچه در مدل‌های شبیه‌سـازی گذشت، تأثیرات تصادفی در مدل‌های چند سـطحی مورد توجه بیشـتری قرار گرفته‌اسـت. اثرهای تصادفی مـدل، واریانس خطای نمونه‌گیری گروه‌ها منتسب به سطح دوم، ناشی از تغییرات عرض از مبدأ و ضریب میانگین ساعت مطالعه و کواریانس بین این دو خطا هستند. نتایج این پژوهش نیز نشان داد، با اندازهٔ نمونهٔ کم در سـطح دوم، ابتدا مشـکل بیش برآوردی اتفاق می افتد. اما با افزایش نمونه، برآورد کواریانس به مقدار واقعی نزدیک‌تر می‌شود. این نتایج همسو با نتایج مطالعات ماک (۱۹۹۵) بیان می‌کند، حجم نمونهٔ پنج گروه در سطح دوم تورش قابل توجهی از واریانس می‌دهد. این درحالی است که کلارک و ویتون (۲۰۰۷) پیشنهاد کردند، حداقل ۱۰ مشاهده در هر گروه برای حداقل گروه‌های ۱۰۰ موردی نیاز است. به اعتقاد آن‌ها حداقل ۲۰۰ گروه با حداقل ۲۰ مشاهده در هرگروه در محاسبهٔ مؤلفه‌های واریانس نیاز است. این تعداد برای مس و هاکس (۲۰۰۴، ۲۰۰۵) به نسـبت ۱۰۰/۳۰ می‌رسد. از این رو توصیه می‌شود، برای رسیدن به برآورد مناسبی از کواریانس بین خطاهای منتسب به متغیر توضیحی، اندازهٔ نمونه اختیاری منتسب به سطح دوم تا حدودی بزرگ اختیار شود؛ اگرچه این وضعیت حتی حادتر از آثار تصادفی سطح اول است.

از آنجـا که هدف این پژوهش بررسـی حجم بهینه در سـطوح مختلف تحلیل در برآورد اثرات ثابت و تصادفی، توان آزمون، خطای برآورد و برآورد مؤلفه‌های واریانس-کواریانس بوده اسـت، نتایج این پژوهش اطلاعاتـی را در حوزهٔ نظر و عمل در اختیار پژوهشگران قرار می‌دهد که بتوان با در نظر گرفتن حداقل حجم اندازهٔ نمونه در سطح‌های اول و دوم و در عین حال داشتن برآوردهای بدون تورش و مناسب از پارامترهای آماری، به نتایج مطلوب رسید.

منابع

- پژوهشگاه مطالعات آموزش وپرورش (۱۳۹۱). *نتایج تیمز و پرلز، ۲۰۱۱*. تهران: نویسنده
- نقش، زهرا. (۱۹۹۶). تحلیل چند سطحی؛ راهکای برای خطاهای حاصل از تجمیع داده ها: استفاده از داده های سطح دانش‌آموز و معلم تیمز ۲۰۱۱. *فصلنامهٔ مطالعات اندازه‌گیری و ارزشیابی آموزشی، ۷*(۱۸)، ۱۴۶–۱۲۷.
- Afshartous, D. (1995). *Determination of Sample Size for Multilevel Model Design*. Paper Presented at the AERA meeting in San Francisco, CA.
- Bell, B. A., Morgan, G. B., Kromrey J. D., Ferron, J. M. (2010). The impact of small cluster size on multilevel models: a Monte Carlo examination of two-level models with binary and continuous predictors. *JSM Proceedings, Survey Research Methods Section, 1*(1), 4057-4067.
- Browne, W. J., Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*, 391-420.
- Browne, W. J, Golalizadeh, M. & Parker, R. (2009). *A Guide to Sample Size Calculations for Random Effect Models via Simulation and ML Pow Sim Software Package*. Bristol: Bristol University Press.
- Busing F. (1993). *Distribution characteristics of variance estimates in two-level models*. Unpublished manuscript, Leiden University, Researchgate.net.
- Clarke, P. & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research, 35*(3), 311-351.
- Cohen, M. (1998). Determining Sample Size for Surveys with Data Analyzed by Hierarchical Linear Models. *Journal of Official Statistics, 14*(3), 267-257.
- Goldstein, H. & Silver, R. (1989). Multilevel and Multivariate Models in Survey Analysis. In C. J Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of Complex Surveys* (pp. 221-235). New York: John Wiley and Sons.
- Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York: Springer Verlag,
- Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University at Los Angeles. ERIC Number: ED371033.
- Longford, N.T. (1987). A First Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Effects. *Biometrika, 74*(4), 812-827.
- Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.
- Moerbeek, M., Van Breukelen, G.J.P., & Berger, M.P.F. (2001). Optimal Experimental Designs for Multilevel Models with Covariates. *Communications in Statistics-Theory and Methods, 30*(12), 2683-2697.
- Mok M. (1995). *Sample size requirements for 2-level designs in educational research*. Unpublished manuscript, Macquarie University.
- Newsom J. T., Nishishiba M. (2002). *Nonconvergence and sample bias in hierarchical linear modeling of dyadic data*. Unpublished Manuscript, Portland State University.
- Snijders, T. A. B. (2005). Power and Sample Size in Multilevel Linear Models'. In B.S. Everitt and D.C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 570–1573). Chicester etc.: Wiley, 2005.
- Snijders, T. A. B., & Bosker R. J. (1993). Standard Errors and Sample Sizes for Two-Level Research. *Journal of Educational Statistics, 3*(18), 237-259.
- Snijder, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling Multilevel Statistical Models*. London etc.: Sage Publication.
- Politis, D.N., Romano, J.P. & Wolf, M. (1999). *Subsampling* .New York: Springer.

پی‌نوشت‌ها

1. Hierarchical
2. Aggregation
3. Cross-sectional
4. Longitudinal with repeated measures
5. Multistage
6. Sufficient sample
7. Mass and Hox
8. Snijders
9. Snigders and Bosker
10. Conditionally optimal
11. Goldstein and Silver
12. Cohen
13. Longford
14. Afshartous
15. bootstrapping
16. Browne, Golalizadeh and Parker
17. Kreft
18. Cross
19. Fixe and random effects
20. Newsom, Nishishiba
21. Clarke and Wheaton
22. Mok
23. Monte carlo study
24. Browne and Draper
25. Bell, Morgan, Kromrey and Ferron
26. Busing
27. Simulation
28. Approximate formula
29. Power of statistical
30. Moerbeek, Breukelen and Berger
31. D-optimality and L-optimality criteria
32. Optimal designs
33. Tow- stage Stratified Cluster Design
34. Strata
35. Sampling Weights
36. Probability Proportional to Size (PPS
37. Reliability
38. Validity
39. Field test
40. Mathematics weight
41. Science weight
42. Subsampling
43. Politis,Romano, Wolf
44. Goldstein, Silver