فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
دوره ششم، شماره ۲۲، تابستان ۹۵، ص ۲۶۱ تا ۲۸۷

Educational Measurement
Allameh Tabataba'i University
Vol. 6, No. 22, Summer 2016

روش بهینه هموارسازی داده‌ها در همتراز سازی: موردمطالعه آزمون تولیمو و آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش آموزش کشور

علی مقدم زاده[۱]

تاریخ دریافت: ۹۴/۰۹/۱۸
تاریخ پذیرش: ۹۵/۰۷/۱۵

چکیده

این پژوهش با هدف یافتن بهترین روش هموارسازی داده‌ها در روش‌های مختلف همترازسازی انجام شد. به این منظور از داده‌های آزمون تولیمو و آزمون‌های جامع کنکورهای آزمایشی شرکت تعاونی سازمان سنجش آموزش کشور در سال ۹۲-۹۱ استفاده شد. برای تحلیل داده‌های آزمون‌های جامع کنکورهای آزمایشی شرکت تعاونی سازمان سنجش آموزش کشور صرفاً از سؤالات مشترک دروس عمومی رشته‌های ریاضی- فیزیک، علوم تجربی و علوم انسانی استفاده شد. برای یافتن بهینه‌ترین روش هموارسازی داده‌ها در همترازسازی از حجم نمونه‌های مختلف و آزمون‌های با تعداد سؤالات مختلف (طول متفاوت) استفاده شد. در آزمون تولیمو از طرح همترازسازی NEAT و در آزمون‌های جامع از طرح گروه‌های همسان استفاده شد. نتایج تحلیل‌ها به‌طورکلی نشان داد که برای هموارسازی داده‌ها در آزمون تولیمو، مدل اول (مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی- چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب شد و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم X) در سه حجم نمونه ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری مدل اول به دست آمد. به همین ترتیب، در فرم Y آزمون تولیمو در حجم نمونه ۲۰۰ و ۱۰۰۰ نفر بهترین مدل، مدل اول و در حجم نمونه ۵۰۰ نفر مدل دوم برگزیده شد. در داده‌های آزمون جامع سنجش، (هم فرم X و هم فرم Y) در حجم‌های نمونه مختلف ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری بهترین

---

۱. عضو هیأت علمی دانشگاه تهران، نویسنده مسئول amoghadamzadeh@ut.ac.ir

مدل، مدل سوم مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) است چون مدلی که کمترین مقدار AIC را داشته باشد برازش بهتری دارد. همچنین نتایج پژوهش گویای آن است که به‌موازات افزایش حجم نمونه، برازش مربوط به هموارسازی کرنل نیز بهبود یافته است و بهبود هموارسازی کرنل با افزایش طول آزمون همراه بوده است.

واژگان کلیدی: همترازسازی، پس هموارسازی (PSE)، روش کرنل (KE)، آزمون تولیمو، سؤالات لنگر، نظریه کلاسیک آزمون (CTT)، Circle arc

مقدمه

نمرات حاصل از اجرای یک آزمون اغلب به‌عنوان یک قسمت از اطلاعات موردنیاز برای اخذ تصمیمات مهم دربارهٔ افراد به کار می‌روند. برخی از این تصمیمات در سطوح فردی و برخی در سطوح اجتماعی (مانند انتخاب شغل) و یا ملی (پذیرش در دانشگاه) اتخاذ می‌شوند (هومن، ۱۳۸۰؛ آناستازی و اوربینا، ۱۹۹۷)؛ اما بدون توجه به نوع تصمیمی که ممکن است گرفته شود این تصمیم باید بر مبنای صحیح‌ترین اطلاعات ممکن اتخاذ گردد، یعنی هر چه اطلاعات صحیح‌تر باشند، تصمیمات بهتری هم گرفته می‌شود. برای تصمیم‌گیری در برخی از این زمینه‌ها لازم است آزمون‌هایی در موقعیت‌های مختلف اجرا شوند. مثلاً مؤسسات علمی و دانشگاه‌ها برای گزینش دانشجو و سازمان‌های اداری و شرکت‌ها برای استخدام کارکنان موردنیاز، اقدام به برگزاری آزمون‌ها در زمان‌های مختلفی می‌نمایند، به‌طوری‌که داوطلبین می‌توانند یکی از زمان‌های تعیین‌شده را انتخاب و در آزمون شرکت کنند و یا در برنامه‌هایی که نیاز به سنجش اولیه و ثانویه وجود دارد و یا در برنامه‌های سنجش مستمر سالانه و نظایر این‌ها که هدف از اجرای مکرر آزمون، اندازه‌گیری صفت، ویژگی و یا مهارت خاصی است شرکت کنند. اگر مجموعه سؤالات آزمون تغییر نکند (از یک فرم آزمون استفاده شود) به علت افشای سؤالات یا تست آشنایی و یا تمرین و تکرار، اندازه‌های حاصل از آزمون‌های بعدی ممکن است بیشتر از اندازه‌های واقعی آزمودنی‌ها در شرایط عادی به دست آید. برای جلوگیری از این امر و حفظ امنیت سؤالات و حذف اثر تست آشنایی و یا تمرین و تکرار می‌توان از مجموعه

سؤالات متفاوت در هر بار اجرای آزمون استفاده نمود. به مجموعه‌ای از سؤالات که بر اساس محتوا و ویژگی‌های آماری خاص تهیه می‌شوند یک فرم از آزمون می‌گویند (میلمن و گرین[1]، ۱۹۸۹). استفاده از فرم‌های مختلف یک آزمون ایجاب می‌نماید که فرم‌ها دارای محتوا و ویژگی‌های آماری یکسان باشند تا صفت خاصی را به‌طور یکسان اندازه‌گیری نمایند (لرد، ۱۹۸۰، ترجمه دلاور و یونسی، ۱۳۹۱). بر این اساس، سازندگان و پرورش‌دهندگان آزمون‌ها از ویژگی‌های آزمون به‌عنوان خط راهنما استفاده می‌کنند تا فرم‌های آزمون ازنظر محتوا و ویژگی‌های آماری تا حد ممکن شبیه هم باشند. این فرم‌ها را فرم‌های موازی یا معادل گویند.

سنجش‌های آموزشی و تربیتی غالباً در زمان‌های مختلفی اجرا می‌شوند و فرم‌های مختلف یک آزمون اغلب برای بیشینه کردن امنیت آزمون ساخته می‌شوند (تانگ و کولن[2]، ۲۰۰۵). فرم‌های مختلف آزمون در یک برنامه سنجش می‌توانند ازلحاظ محتوا مشابه باشند (که معمولاً این‌چنین هم است) ولی سؤالات متفاوتی داشته باشند (که نوعاً هم این‌گونه است). با طراحی فرم‌های جایگزین می‌توان مانع این شد که آزمودنی‌ها به سؤالات یکسانی در زمان‌های مختلفی پاسخ دهند. از این طریق می‌توان مسئله نحوه ارائه سؤال[3] را کنترل کرد، بدین‌صورت که آزمودنی‌ها دیگر از آن دسته سؤالات آزمون که قبلاً با آن مواجه شده‌اند آگاه نمی‌شوند. البته هنوز یک مسئله باقی می‌ماند. بااینکه فرم‌های متفاوت آزمون را می‌توان در زمان‌های مختلفی مورداستفاده قرار داد، ولی دشواری این فرم‌ها ممکن است تا حدی با یکدیگر متفاوت باشد. این شرایط هم ازنظر آزمودنی‌ها و هم ازنظر مسئولان برگزاری آزمون، ناعادلانه است.

یکی از مسائل مهم در ادبیات مربوط به همترازسازی، روش‌های پیش هموارسازی[4] و روش‌های پس هموارسازی[5] است. در روش‌های پیش هموارسازی توزیع نمرات هموار

1. Millman & Greene
2. Tong & Kolen
3. item exposure
4. Presmoothing equating
5. Postsmoothing

می‌شوند که در هموارسازی توزیع‌ها دقت در برآورد توزیع‌ها بسیار مهم است. یک ویژگی مهم که ارتباط نزدیکی با دقت دارد شاخص حفظ مشخصات توزیع (گشتاورهای درجه ۱ تا ۴) است؛ بنابراین شاخص، توزیع هموارشده حداقل برخی از همان شاخص‌های مرکزی را که در توزیع مشاهده‌شده وجود دارد دارا است. برای مثال اگر دو شاخص اصلی مانند میانگین و انحراف استاندارد در توزیع هموارشده و هموار نشده با یکدیگر برابر باشند آن‌وقت می‌گوییم روش هموارسازی مورداستفاده، دو شاخص اصلی را حفظ کرده است.

یک روش هموارسازی از یک مدل لگاریتم خطی چندجمله‌ای برای هموار کردن توزیع‌های نمرات استفاده می‌کند. روش دوم نیز مدل قوی نمره واقعی است. در این روش یک توزیع به‌صورت توزیع نمرات واقعی درمی‌آید و خطای نمره واقعی محاسبه می‌گردد. در هر دو روش بعد از هموار کردن توزیع‌ها نمرات واقعی درمی‌آید و خطای نمره واقعی محاسبه می‌شود. در هر دو روش بعد از هموار کردن توزیع‌ها نمرات فرم X و Y با یکدیگر همتراز می‌شوند. (کولن و برنان، ۲۰۰۴).

فرایند پس هموارسازی، فرایند عبور دادن یک خط مستقیم یا منحنی از بین نقاطی است که روابط همترازسازی هم صدک بین آن‌ها برقرار گردیده است. درواقع در این روش‌ها معادل‌های هم صدک مورد هموارسازی قرار می‌گیرند. کار همترازسازی هم صدک با در دست داشتن جداولی که در هر یک فراوانی نمرات برای آزمودنی‌های هر دو گروه ذکر شده‌اند شروع می‌شود و درنهایت با به دست دادن یک جدول که در آن نمرات معادل هر فرم آزمون در فرم دیگر قید گردیده به پایان می‌رسد. لازم به ذکر است نمرات معادل معمولاً نمرات اعشاری هستند حتی اگر نمرات اصلی عدد صحیح باشند. کولن و برنان (۲۰۰۴) به توصیف چند روش پس هموارسازی پرداخته‌اند. همچنین فربنک (۱۹۸۷) به هفت روش پس هموارسازی ازجمله لگاریتم خطی، رگرسیون با جملات درجه چهارم و درجه سوم و ... اشاره نموده است (کولن و برنان، ۲۰۰۴؛ فربنک، ۱۹۸۷).

اگرچه بااینکه روش‌های قبل از هموارسازی لگاریتم خطی چندجمله‌ای[1] و روش قبل از هموارسازی spline درجه سوم[2] در مطالعات بسیاری موردبررسی قرار گرفته است (کوی‌ای، ۲۰۰۶؛ کوی‌ای و کولن[3]، ۲۰۰۸، ۲۰۰۹؛ فربنک[4]، ۱۹۸۷؛ هانسون، ۱۹۹۰، ۱۹۹۱، ۱۹۹۶؛ هانسون، زنگ و کولتون[5]، ۱۹۹۴؛ هالاند و تایر، ۱۹۸۱، ۲۰۰۰؛ کولن، ۱۹۸۴، ۱۹۹۱؛ کولن و جارجورا[6]، ۱۹۸۷؛ لیوینگستون، ۱۹۹۳؛ موزس و هالاند[7]، ۲۰۰۷، ۲۰۰۹a، ۲۰۱۰، موزس و ون داویر، ۲۰۰۶، زِنگ، ۱۹۹۵)، ولی هیچ‌کدام در این پژوهش مرور نشده و موردتوجه نبوده‌اند. در روش قبل از هموارسازی لگاریتم خطی چندجمله‌ای، تنها مطالعات درباره روش قبل از هموارسازی لگاریتم خطی تک متغیری مرور شده است. به‌علاوه، برای هر دو روش قبل از هموارسازی، تنها پیشینه‌ای موردنظر بوده است که نتایج هموارسازی با پارامترهای هموارسازی متفاوت را مقایسه کرده‌اند.

هانسون (۱۹۹۰) سه روش پیش هموارسازی، یعنی روش کرنل، روش پیش هموارسازی لگاریتم خطی چندجمله‌ای (هابرمن، ۱۹۷۴b) و روش دوجمله‌ای بتا ۴ پارامتری[8] (لرد، ۱۹۶۵) را به‌طور مفصل معرفی کرد. وی اثربخشی این سه روش را در برآورد توزیع‌های جامعه بر اساس تعداد زیادی از داده‌های مشاهده‌شده با توجه به سه آزمون که اولی شامل ۵۹ سؤال چندگزینه‌ای بود که نسخه موازی و کوچک‌تر یک آزمون کامل ۲۰۰ سؤالی است، یک آزمون ریاضی ۴۰ سؤالی مربوط به سنجش دانشکده‌های امریکا[9] و یک آزمون علوم اجتماعی[10] ۵۲ سؤالی مورد مقایسه قرار داد. در روش پیش هموارسازی لگاریتم خطی چندجمله‌ای، مقادیر C در دامنه ۱۰-۱ موردبررسی قرار گرفت و کمترین مقدار C که توسط روش هابرمن (۱۹۷۴b) رد نشده بود

1. polynomial loglinear presmoothing
2. cubic spline postsmoothing
3. Cui & Kolen
4. Fairbank
5. Hanson, Zeng, & Colton
6. Kolen & Jarjoura
7. Moses & Holland
8. the 4-parameter beta binomial method
9. American College Testing (ACT) Assessment Mathematics test
10. ACT Assessment Social Science test

مورداستفاده قرار گرفت. تعداد آزمودنی‌هایی که در هر آزمون شرکت کرده بودند به ترتیب برابر با ۳۹۱۴۹۲۳۰۰۶۵ و ۲۳۰۰۶۵ نفر بود. هر سه توزیع فراوانی مشاهده‌شده به‌اندازه کافی هموار شده بودند و به نظر می‌رسید که توزیع‌های جامعه باشند. ۵۰۰ نمونه با حجم‌های ۵۰۰، ۱۰۰۰، ۲۰۰۰ و ۵۰۰۰ نفری به‌طور تصادفی از هر توزیع جامعه انتخاب شدند. در هر نمونه، روش کرنل متغیر، روش کرنل ثابت، روش دوجمله‌ای بتای ۴ پارامتری، و روش پیش هموارسازی لگاریتم خطی چندجمله‌ای برای برآورد توزیع‌های جامعه بکار گرفته شدند. پژوهش هانسون (۱۹۹۰) گویای آن است که تمامی روش‌های هموارسازی برآورد توزیع‌های جامعه را برحسب واریانس و میانگین مربع خطا برای تمامی آزمون‌ها و تمامی نمونه‌ها با حجم‌های متفاوت بهبود می‌بخشند، البته دقت روش دوجمله‌ای بتای ۴ پارامتری از سایر روش‌ها بیشتر است به‌ویژه زمانی که حجم نمونه‌ها کمتر از ۲۰۰۰ نفر باشد. در تمامی این سه توزیع جامعه، تمامی روش‌های هموارسازی واریانس را به‌طور معناداری کاهش دادند و اندکی سوگیری بیشتری را معرفی کردند. بر اساس نتایج، مدل لگاریتم خطی چندجمله‌ای، کمترین مقدار سوگیری به‌جز برای آزمون ریاضی در نمونه‌های با حجم کوچک‌تر (۵۰۰ و ۱۰۰۰ نفری) را دارد. هانسون (۱۹۹۰) همچنین پیشنهاد کرده است که نمودارهای توزیع‌های نمرات خام و نمرات برازش یافته و آماره‌ی کلی نیکویی برازش خی‌دو باید در عمل برای انتخاب پارامتر C در روش پیش هموارسازی لگاریتم خطی چندجمله‌ای موردبررسی قرار گیرند.

کولن (۱۹۹۱) همچنین به بررسی این موضوع پرداخت که عملکرد سه روش پیش هموارسازی، یعنی روش کرنل دوجمله‌ای، روش نمره واقعی قوی[1] و روش پیش هموارسازی لگاریتم خطی چندجمله‌ای در هموارسازی توزیع‌های نمره مشاهده‌شده بر اساس آزمون ریاضی ۴۰ سؤالی (۱۹۸۸) با ۳۰۳۹ آزمودنی چگونه است. توضیح هر روش هموارسازی و انتخاب پارامترهای هموارسازی (h برای روش کرنل و C برای روش لگاریتم خطی چندجمله‌ای) بر اساس بررسی تصویری توزیع‌های نمره خام و هموارشده و

1. strong true-score method

گشتاورهای توزیع‌ها همچنین موردبررسی قرار گرفت. در روش کرنل دوجمله‌ای، هرچه پارامتر هموارسازی (h) افزایش یابد گشتاورها و نمودارهای توزیع‌های برازش یافته از توزیع نمونه مشاهده‌شده فاصله بیشتری می‌گیرند. کولن (۱۹۹۱) تأکید کرده است که باید به روش کرنل توجه ویژه‌ای شود چون یک مقدار بزرگ h می‌تواند منجر به افزایش انحراف استاندارد و دنباله‌های طولانی توزیع هموار شده شوند.

پژوهشی که توسط گادفری[1] (۲۰۰۷) در دو بخش انجام شده است به ۱) تأثیر انتخاب مدل لگاریتم خطی[2] در قبل از هموارسازی توزیع‌های نمرهٔ مشاهده‌شده[3] در روش کرنل همترازسازی آزمون[4] و ۲) تفاوت‌های بین همترازسازی کرنل، همترازسازی هم صدک رشته‌ای[5] و روش‌های نمره واقعی یعنی مدرج سازی همزمان[6] و روش تبدیل استاکینگ و لرد[7] پرداخته است. برای اینکه بتوان با شرایط واقعی رقابت کرد داده‌ها به نحوی شبیه‌سازی شدند که در آن دشواری آزمون متفاوت است، حجم نمونه متغیر است، طول آزمون لنگر متفاوت است و دامنه تغییر طول آزمون بین ۲۰ سؤال تا ۱۰۰ سؤال است. دشواری سؤالات لنگر ثابت نگه داشته شد/فرض شد. چون داده‌ها در فرمت یک گروهی[8] (SG) شبیه‌سازی شدند، همترازسازی هم صدک غیر هموار متداول[9] سنتی به‌عنوان یک ملاک استفاده شد که از آن طریق تمامی روش‌های دیگر که از طرح گروه‌های غیر معادل با یک آزمون لنگر[10] (NEAT) استفاده می‌کنند را بتوان مورد مقایسه قرار داد. داده‌ها با استفاده از نرم‌افزار ICEDOG (ETS، ۲۰۰۷) شبیه‌سازی شدند و با استفاده از نرم‌افزار KE (ETS، ۲۰۰۷)، MULTILOG (تیسن[11]، ۲۰۰۳)، ICEDOG (ETS، ۲۰۰۷)، PARSCALE

1. Godfrey
2. loglinear model
3. presmoothing observed score distributions
4. kernel method of test equating
5. chained equipercentile equating
6. concurrent calibration
7. Stocking and Lord's transformation
8. single group (SG) format
9. unsmoothed equipercentile equating
10. non-equivalent groups with an anchor test design (NEAT)
11. Thissen

(موراکی و باک[1]، ۲۰۰۳) و یک سری کدهای برنامه‌نویسی فرترن نوشته‌شده توسط محقق مورد تحلیل قرار گرفتند. نتایج حاکی از این است که تکنیک همترازسازی انتخابی بر نمرات آزمون آزمودنی‌ها در شرایط واقعی مختلف تأثیر دارد.

با توجه به آنچه بیان شد مهم‌ترین هدف کاربردی این پژوهش، استفاده از نتایج آن در همترازسازی نمرات آزمون‌های تولیمو و آزمون‌های جامع کنکورهای آزمایشی شرکت تعاونی سازمان سنجش آموزش کشور است. لذا هدف اصلی این پژوهش بررسی تأثیر مدل لگاریتم خطی (برای مرحله پیش هموارسازی) بر توزیع نمرات بوده است؛ بنابراین سؤال اصلی پژوهش این بوده که بهترین مدل لگاریتم خطی برای هموار کردن داده‌ها در همترازسازی در نمونه‌هایی با حجم‌های متفاوت کدام است؟

روش پژوهش

در حوزهٔ روش‌شناسی پژوهش، محققان در شرایط نسبتاً برابر ترجیح می‌دهند به‌جای داده-های شبیه‌سازی‌شده از داده‌های واقعی استفاده کنند. ازاین‌رو در این پژوهش سعی شده است که از داده‌های واقعی مربوط به اجرای آزمون‌های تولیمو و آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش استفاده شود. با این اطلاعات دربارهٔ رتبه و وضعیت داوطلبان شرکت‌کننده در آزمون زبان تولیمو و منتخبی از دروس آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش تصمیماتی اتخاذ شده است. نظر به اینکه در این پژوهش محقق در پی مقایسه رویکردهای مختلف در همترازسازی نتایج حاصل از اجرای فرم‌های مختلف آزمون، به دست آوردن برآوردهای باثبات‌تر پارامترها (اعم از پارامترهای سؤال و توانایی آزمودنی‌ها) و افزایش اعتبار تصمیم‌ها و دقت اندازه‌گیری خواهد بود، لذا روش این پژوهش به‌طور عام جزو پژوهش‌های توصیفی (غیرآزمایشی و غیر تاریخی) است. پژوهش‌های توصیفی شامل مجموعه روش‌هایی است که هدف آن‌ها توصیف کردن شرایط یا پدیده‌های موردبررسی است؛ اجرای پژوهش توصیفی می‌تواند صرفاً برای

1. Muraki & Bock

شناخت بیشتر شرایط موجود یا یاری‌دادن به فرایند تصمیم‌گیری باشد (سرمد، بازرگان، و حجازی، ۱۳۸۴). روش پژوهش حاضر از حیث نوع تجزیه‌وتحلیل‌های آماری و سنجشی جزو تحقیقات همبستگی (که خود زیرمجموعهٔ روش پژوهش توصیفی قرار می‌گیرد) است. درواقع، چون در پژوهش حاضر با تحلیل‌های چند متغیری مواجهیم از مجموعه همبستگی‌های دو متغیری که ماتریس همبستگی یا ماتریس کوواریانس را می‌سازد استفاده می‌شود؛ بنابراین روش این پژوهش را می‌توان جزو تحقیقات همبستگی از نوع تحلیل ماتریس کوواریانس دانست.

به‌منظور دستیابی به اهداف پژوهش، از اطلاعات واقعی به‌دست‌آمده از اجرای آزمون تولیمو (۱۳۹۲-۱۳۹۱) و داده‌های آزمون‌های جامع کنکور آزمایشی (دروس عمومی مشترک رشته‌های ریاضی، تجربی و علوم انسانی) استفاده شده است. بدین ترتیب، جامعه آماری و گروه نمونه پژوهش حاضر، جامعه آماری و گروه نمونه شرکت‌کننده در آزمون‌های مذکور می‌شود که در سال تحصیلی ۱۳۹۲-۱۳۹۱ به اجرا درآمده است. لازم به ذکر است که در نمونه‌گیری و تعیین حجم نمونهٔ آزمودنی‌ها برای تحلیل‌های موردنظر باید توجه داشت که نظریه‌های روان‌سنجی، نظریه‌های نمونه‌های بزرگ[1] هستند، و جهت برآورد باثبات پارامترها، باید نمونه‌ای با حجم بالا انتخاب گردد. البته با توجه به اینکه حجم نمونه یکی از متغیرهای مهم در دقت نتایج همترازسازی است، در این پژوهش نمونه‌های ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری به‌طور کاملاً تصادفی از مجموعه داده‌های آزمون‌های تولیمو و آزمون‌های جامع کنکور آزمایشی سازمان سنجش انتخاب و سپس پارامترها برآورد شده است.

داده‌های این پژوهش از طریق آزمون‌های سراسری که توسط سازمان سنجش آموزش کشور در سال تحصیلی ۹۲-۹۱ اجرا شده‌اند به دست آمده‌اند. به‌طور دقیق از داده‌های حاصل از آزمون تولیمو در سال ۹۲-۹۱ و داده‌های آزمون‌های جامع کنکور آزمایشی در سال ۹۲-۹۱ استفاده شده است.

1. Large Sample Theory

الف) آزمون تولیمو. هدف از برگزاری آزمون TOLIMO سنجش توانش زبانی افرادی است که زبان مادری آنان زبان انگلیسی نیست. نمرات حاصل از این آزمون را می‌توان برای مقاصد زیر مورداستفاده قرار داد: ۱- پذیرش دانشجو در مقاطع مختلف تحصیلی به‌ویژه کارشناسی ارشد و دکتری در رشته‌های مختلف دانشگاهی؛ ۲- اعطای بورس تحصیلی و یا شرط استخدام قابل‌استفاده توسط مؤسسات، سازمان‌ها و ارگان‌های دولتی و غیردولتی. آزمون TOLIMO هرساله چندین نوبت به‌طور منظم در سطح کشور برگزار می‌شود و کلیه افراد بدون توجه به سن، جنسیت، قومیت یا ملیت می‌توانند در آن شرکت کنند. این آزمون که توسط کارشناسان سازمان در دفتر آزمون‌سازی تهیه می‌شود شامل سه بخش است که در جدول ۱-۳ ساختار کلی آزمون TOLIMO نشان داده شده است.

جدول ۱. ساختار آزمون TOLIMO

بخش‌های آزمون	هدف	بخش	تعداد سؤال
۱ ـ دستور و نگارش مجموع سؤالات = ۴۰) زمان = ۳۰ دقیقه	سنجش توانایی داوطلبان در انتخاب ساختارهای درست زبان انگلیسی	قسمت A ـ یافتن پاسخ صحیح قسمت B ـ یافتن گزینه غلط	۱۵ ۲۵
۲ ـ خواندن و درک مطلب زمان= ۵۵ دقیقه	سنجش توانایی داوطلبان در خواندن و درک صحیح متون علمی در موضوعات گوناگون. سؤالات مربوط به هر متن دقیقاً اجزای مختلف مهارت خواندن ازجمله یافتن موضوع اصلی موردبحث متن، استنتاج، حدس معانی کلمات مشکل به کمک بافت کلامی موجود و غیره را موردسنجش قرار می‌دهد.	۵ یا ۶ متن آکادمیک	۵۰
۳ ـ درک شنیداری (مجموع سؤالات = ۵۰) زمان = ۳۵ دقیقه	سنجش توانایی داوطلبان در زمینهٔ درک شفاهی مکالمات روزمره و آکادمیک و نیز سخنرانی‌های کلاسی و دانشگاهی	قسمت A ـ مکالمات کوتاه قسمت B ـ مکالمات طولانی‌تر قسمت C ـ سخنرانی‌ها	۳۰ ۷ یا ۸ ۱۳ یا ۱۲

بعدازآنکه نمره خام هر داوطلب در هر یک از بخش‌های آزمون مشخص گردید هریک از آن‌ها به نمره‌ای تراز که برای تمامی بخش‌های آزمون بین ۲۸ تا ۶۸ است تبدیل

می‌گردد. سپس سه نمره تراز با یکدیگر جمع و حاصل جمع در عدد ۱۰ ضرب و مجموع حاصل تقسیم‌بر عدد ۳ می‌شود. نمره به‌دست‌آمده نمره کل هر داوطلب محسوب می‌گردد. نمره کل هر داوطلب بین ۲۰۰ تا ۶۷۷ متغیر است. نمره مطلوب در آزمون TOLIMO از طرف دفتر آزمون‌سازی سازمان سنجش آموزش کشور تعریف نمی‌شود بلکه دانشگاه‌ها، ادارات، مؤسسات و غیره هستند که برحسب انتظارات خود حد نمره مطلوب را مشخص می‌سازند. ولی به طول کلی نمره زیر ۴۰۰ ضعیف ارزیابی می‌گردد و اکثر دانشگاه‌های انگلیسی‌زبان برای پذیرش دانشجو در مقطع کارشناسی حداقل نمره ۵۵۰ را ملاک قرار می‌دهند. این دانشگاه‌ها برای پذیرش دانشجو در مقطع کارشناسی ارشد و دکتری عموماً نمره‌ای برابر یا بالاتر از ۶۰۰ می‌خواهند. برای هر داوطلب با توجه به نمرات اکتسابی در هر بخش آزمون و نمره کل، رتبه صدکی خاصی گزارش می‌شود. رتبه صدکی جایگاه نمره دریافتی هر داوطلب را در مقایسه با نمرات اخذشده توسط سایر داوطلبان مشخص می‌سازد.

ب) آزمون‌های آزمایشی جامع شرکت تعاونی کارکنان سازمان سنجش آموزش کشور. شرکت تعاونی خدمات آموزشی کارکنان سازمان سنجش آموزش کشور با بررسی‌های کارشناسی مجموعه آزمون‌های آزمایشی را در نُه نوبت طراحی نموده است. شش نوبت از آزمون‌های آزمایشی به‌صورت مرحله‌ای و سه نوبت آزمون جامع برگزار خواهد شد تا داوطلبان شرکت‌کننده در آزمون‌های مرحله‌ای و جامع، با ارزیابی کاملاً علمی و استاندارد، از وضعیت علمی و تحصیلی خود شناخت پیدا کرده و در هر مرحله از آزمون‌ها نسبت به رفع مشکلات تحصیلی خود اقدام نمایند. دانش‌آموزان پس از اتمام آزمون‌های مرحله‌ای، با شرکت در سه نوبت آزمون‌های آزمایشی جامع و حضور در جلسات مشابه آزمون سراسری و پاسخگویی به سؤالات استاندارد و همتراز با آزمون سراسری سال ۱۳۹۲، آمادگی خود را روزبه‌روز افزایش داده و موقعیت علمی و تحصیلی خود را در سه نوبت تا قبل از برگزاری آزمون سراسری سال ۱۳۹۲، به‌طورجدی محک می‌زنند.

یافته‌ها

جهت تجزیه‌وتحلیل داده‌های جمع‌آوری‌شده از نرم‌افزارهای کامپیوتری EXCEL، SPSS، و R، KE (چن و همکاران[1]، ۲۰۰۷) استفاده شده است. این پژوهش معطوف به سؤالات و آزمون‌هایی است که به‌صورت دو ارزشی نمره‌گذاری شده و روش‌های همترازسازی کرنل و همترازسازی نمره مشاهده‌شده در آن موردتوجه است. برای انجام روش همترازسازی کرنل از نرم‌افزار KE 3.0 (چن و همکاران، ۲۰۰۷) استفاده شده است. چون در همترازسازی کرنل غالباً نوع مدل انتخابی برای پیش هموارسازی که برای داده‌های مورد تحلیل بهترین برازش را داشته باشد نامعلوم است (ساده‌ترین مدلی که با توزیع برازش داشته باشد)، سعی شده تا با استفاده از ملاک آگاهی آکائیک[2] (AIC)، خی‌دوی پیرسون، و مقدار شاخص Residual مدل بهینه انتخاب شود. آماره خی‌دوی نشان می‌دهد که کدام مدل با توزیع داده‌های خام برازش دارد. پس‌ازآن توزیع دو متغیری هموار شده در نرم‌افزار همترازسازی کرنل (ETS، ۲۰۰۷) وارد شد و داده‌ها همتراز شدند.

برای پاسخ به سؤال اصلی پژوهش مبنی بر اینکه "بهترین مدل لگاریتم خطی برای هموار کردن داده‌ها در همترازسازی در نمونه‌هایی با حجم‌های متفاوت کدام است؟" ابتدا نتایج مربوط به آزمون تولیمو و سپس نتایج مربوط به آزمون جامع سنجش ارائه شده است.

الف) آزمون تولیمو. برای پاسخگویی به سؤال پژوهش در داده‌های آزمون تولیمو مدل‌های لگاریتم خطی برای طرح گروه‌های ناهمسان با آزمون لنگر (داده‌های تولیمو) انتخاب و استفاده شد. این مدل‌ها شامل سه مدل به شرح زیر بوده است:

مدل اول: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) هم برای آزمون و هم برای لنگر

1. Chen et al.
2. Akaike Information Criterion

مدل دوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) برای آزمون و لنگر و فرض تعامل بین میانگین‌های آزمون و لنگر و تعامل بین واریانس‌های آزمون و لنگر

مدل سوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) برای آزمون و لنگر و تمام تعامل‌های ممکن بین میانگین و واریانس آزمون و لنگر (در این مدل بین میانگین‌ها و واریانس‌ها هم تعامل وجود دارد).

آزمون X و Y. این داده‌ها به دلیل وجود آزمون لنگر دارای توزیع فراوانی دو متغیری هستند که امکان رسم نمودار مثل طرح گروه‌های یکسان را ندارد، بنابراین در ادامه فقط شاخص‌های برازش گزارش شده است. در ادامه شاخص‌های برازش مربوط به مدل‌های مختلف در آزمون تولیمو فرم‌های X و Y برای حجم نمونه ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری گزارش شده است.

جدول ۲: شاخص‌های برازش سه مدل لگاریتم خطی برای هموار کردن داده‌های آزمون تولیمو فرم‌های X و Y در حجم نمونه مختلف

آزمون X					آزمون Y				
حجم نمونه	مدل	AIC	Residual	df	حجم نمونه	مدل	AIC	Residual	df
۲۰۰	اول	۹۰۷/۸۹۸۴	۵۵۳/۶۱۱۵	۲۲۲۲	۲۰۰	اول	۵۸۶/۱۳۵۲	۲۸۸/۰۳۱۶	۲۲۲۲
	دوم	۹۱۱/۶۹۹۴	۵۵۳/۴۱۲۶	۲۲۲۰		دوم	۵۸۸/۰۸۵۴	۲۸۵/۹۸۱۹	۲۲۲۰
	سوم	۹۱۳/۵۰۰۶	۵۵۱/۲۱۳۷	۲۲۱۸		سوم	۵۸۸/۳۸۹۸	۲۸۲/۲۸۶۲	۲۲۱۸
		χ^2 =۲/۴۰	df =۴	P =۰/۶۶			χ^2 =۵/۷۵	df =۴	P =۰/۲۲
۵۰۰	اول	۱۵۴۹/۹۵۷	۷۹۵/۵۵۶۳	۲۲۲۲	۵۰۰	اول	۱۵۹۰/۲۷۱	۱۰۴۰/۸۵۵	۲۲۲۲
	دوم	۱۵۵۰/۳۹۷	۷۹۱/۹۹۶۳	۲۲۲۰		دوم	۹۵۰/۰۸۲۹	۳۹۶/۶۶۶۹	۲۲۲۰
	سوم	۱۵۵۴/۲۶۷	۷۹۱/۸۶۶۲	۲۲۱۸		سوم	۹۵۳/۹۹۸۴	۳۹۶/۵۸۲۳	۲۲۱۸
		χ^2 =۳/۶۹	df =۴	P =۰/۴۵			χ^2 =۷/۵۱	df =۴	P =۰/۱۱
۱۰۰۰	اول	۲۱۶۸/۸۲۲	۹۹۸/۵۷۵۶	۲۲۲۰	۱۰۰۰	اول	۱۲۴۶/۹۶۵	۴۵۹/۷۳۰۵	۲۲۲۰
	دوم	۲۱۷۱/۵۰۲	۹۹۷/۲۵۵۳	۲۲۲۰		دوم	۱۲۴۹/۹۹۳	۴۵۸/۷۵۷۸	۲۲۲۰
	سوم	۲۱۷۵/۳۲۷	۹۹۷/۰۸۰۱	۲۲۱۸		سوم	۱۲۴۸/۳۴۶	۴۵۳/۱۱۱۷	۲۲۱۸
		χ^2 =۱/۵۰	df =۴	P =۰/۸۳			χ^2 =۶/۶۲	df =۴	P =۰/۱۶

بر اساس اطلاعات جدول بالا مشاهده می‌شود که شاخص AIC و مقدار Residual در سه مدل گزارش شده است.

بر اساس توضیحات مربوط به AIC می‌توان گفت که مدلی که کمترین مقدار AIC و Residual را داشته باشد برازش بهتری دارد. ولی با توجه به نتایج جدول در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. به همین دلیل از آزمون تفاوت خی‌دو بین دو مدل کمک گرفته شده است. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم می‌توان دید که آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (P >۰/۰۵). لذا مدل اول (مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی- چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و در حجم نمونه ۲۰۰ نفری بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم X) است. در نمونه ۵۰۰ نفری مشاهده می‌شود که در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. هرچند بر اساس توضیحات قبل، مدلی که کمترین مقدار AIC و Residual را داشته باشد برازش بهتری دارد. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (۰/۰۵< P). لذا مدل اول (مدل لگاریتم خطی شامل چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم X) است. به همین ترتیب در حجم نمونه ۱۰۰۰ نفری، در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (P >۰/۰۵). لذا مدل اول (مدل لگاریتم خطی شامل چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم X)، مدل اول است.

آزمون ۷. با توجه به نتایج جدول مشاهده می‌شود که در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. بر همین اساس، از آزمون تفاوت خی‌دو بین دو

مدل برای تشخیص بهترین مدل کمک گرفته شده است. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم، آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (۰/۰۵< P). لذا مدل اول (مدل لگاریتم خطی شامل چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم Y)، در حجم نمونه ۲۰۰ نفر مدل اول است. در نمونه ۵۰۰ نفری مشاهده می‌شود که در مدل دوم شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. هرچند بر اساس توضیحات قبل، مدلی که کمترین مقدار AIC و Residual را داشته باشد برازش بهتری دارد. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل دوم و سوم می‌توان دید که آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (۰/۰۵< P). لذا مدل دوم (مدل لگاریتم خطی شامل چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم Y) در حجم نمونه ۵۰۰ نفر است. به همین ترتیب در حجم نمونه ۱۰۰۰ نفری، در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. بر اساس نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم مشاهده می‌شود که آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (۰/۰۵< P). لذا مدل اول (مدل لگاریتم خطی شامل چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب می‌شود و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم Y) برای حجم نمونه ۱۰۰۰ نفر، مدل اول است.

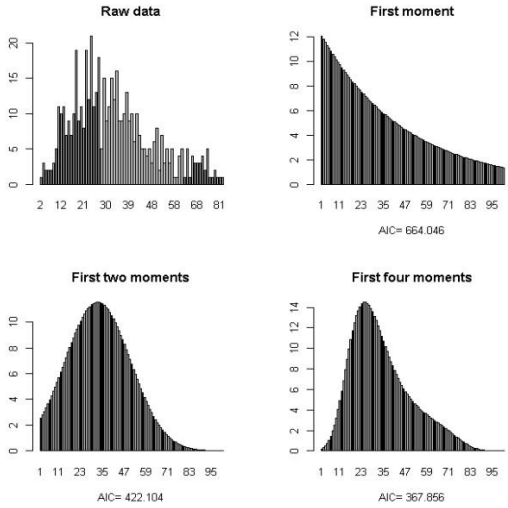
ب) آزمون جامع سنجش. برای پاسخگویی به این سؤال پژوهش در داده‌های آزمون جامع سنجش (دروس عمومی) از مدل‌های لگاریتم خطی برای طرح گروه‌های همسان انتخاب و استفاده شد. این مدل‌ها شامل سه مدل به شرح زیر بوده است:

مدل اول: مدل لگاریتم خطی تنها شامل میانگین (گشتاور اول)

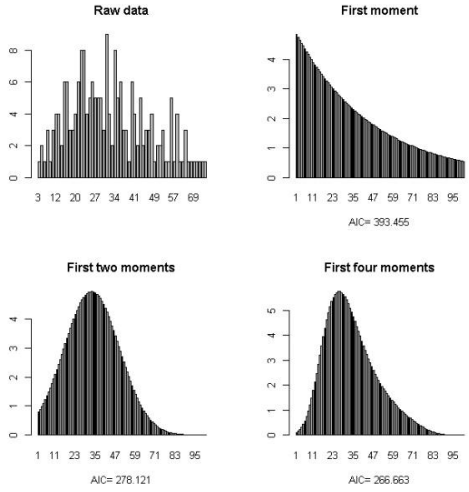
مدل دوم: مدل لگاریتم خطی شامل میانگین و واریانس (دو گشتاور اول)

مدل سوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول)

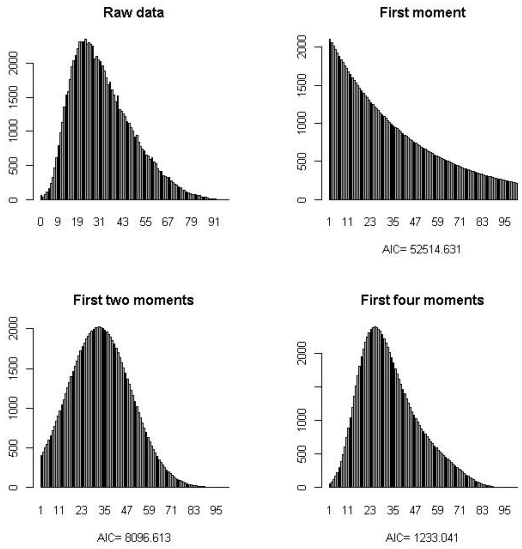
در این داده‌ها به دلیل عدم وجود آزمون لنگر امکان ترسیم نمودار برای طرح گروه‌های یکسان وجود دارد، بنابراین در ادامه فقط نمودارها و شاخص برازش AIC گزارش شده است. در شکل زیر نمودارهای مربوط به آزمون سوم جامع سنجش برای نمونه‌های مختلف ۲۰۰، ۵۰۰، ۱۰۰۰ نفری (فرم X) ارائه شده است:



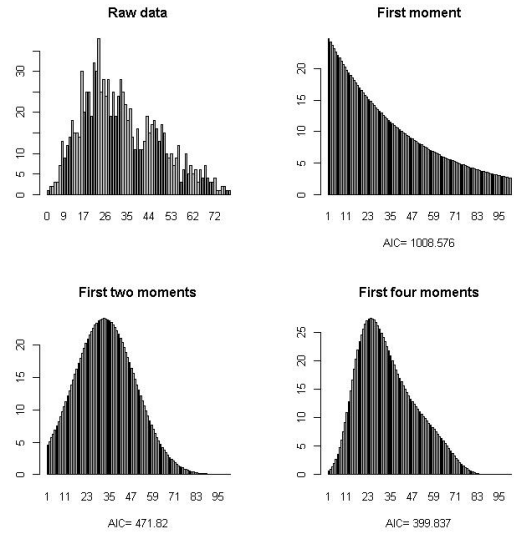
داده‌های آزمون جامع سنجش فرم X در حجم نمونه ۵۰۰ نفر



داده‌های آزمون جامع سنجش فرم X در حجم نمونه ۲۰۰ نفر



داده‌های آزمون جامع سنجش فرم X برای کل حجم نمونه

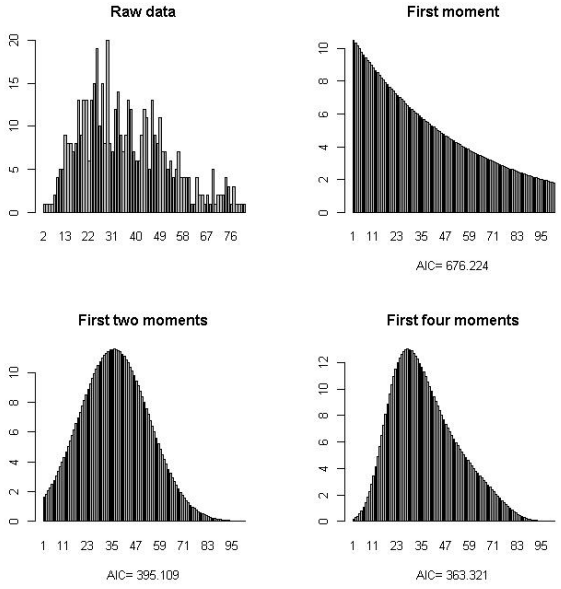


داده‌های آزمون جامع سنجش فرم X در حجم نمونه ۱۰۰۰ نفر

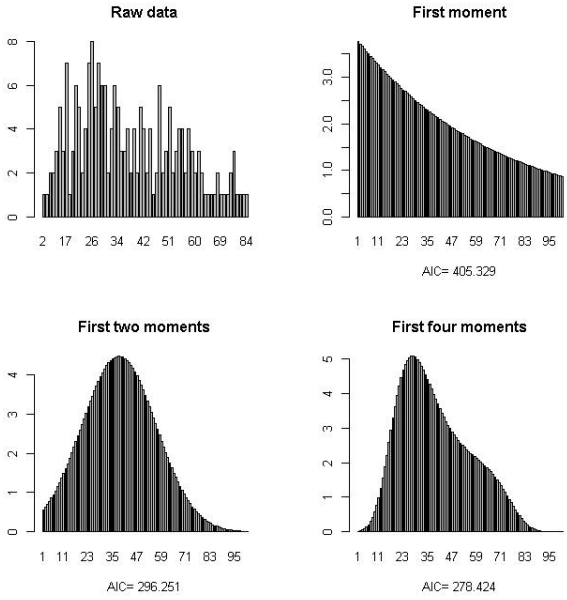
شکل ۱. نمودارهای برازش سه مدل لگاریتم خطی برای هموار کردن داده‌های آزمون جامع سنجش فرم X در حجم نمونه‌های مختلف

با توجه به نمودار داده‌های خام و سه نمودار مربوط به سه مدل موردنظر می‌توان مشاهده کرد که برای داده‌های این آزمون با حجم‌های مختلف ۲۰۰ نفری، ۵۰۰ نفری، ۱۰۰۰ نفری و حتی کل حجم نمونه بهترین مدل، مدل سوم است، چون مدلی که کمترین مقدار AIC را داشته باشد برازش بهتری دارد؛ بنابراین مدل سوم با مقدار AIC برابر با ۲۶۶/۶۶۳ بهترین مدل برای هموارسازی این مجموعه داده‌ها است. در نمونه ۵۰۰ نفری مدل سوم با مقدار AIC برابر با ۳۶۷/۸۵۶ بهترین مدل برای هموارسازی این مجموعه داده‌ها است؛ در نمونه ۱۰۰۰ نفری مدل سوم با مقدار AIC برابر با ۳۹۹/۸۳۷ بهترین مدل برای هموارسازی این مجموعه داده‌ها است؛ و با کل حجم نمونه مدل سوم با مقدار AIC برابر با ۱۲۳۳/۰۴۱ بهترین مدل برای هموارسازی این مجموعه داده‌ها است.

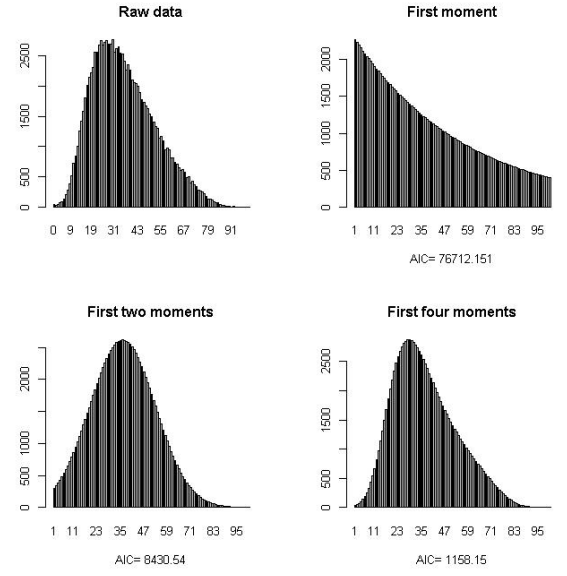
در شکل زیر نمودارهای مربوط به آزمون سوم جامع سنجش برای نمونه‌های مختلف ۲۰۰، ۵۰۰، ۱۰۰۰ نفری (فرم Y) ارائه شده است:



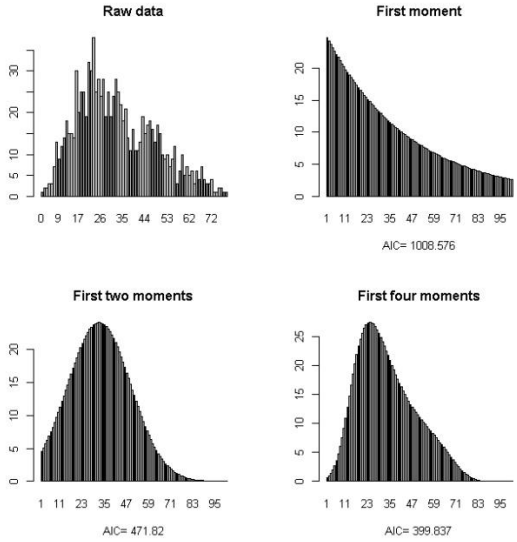
داده‌های آزمون جامع سنجش فرم Y در حجم نمونه ۵۰۰ نفر



داده‌های آزمون جامع سنجش فرم ۷ در حجم نمونه ۲۰۰ نفر



داده‌های آزمون جامع سنجش فرم ۷ برای کل حجم نمونه



شکل ۲. نمودارهای برازش سه مدل لگاریتم خطی برای هموار کردن داده‌های آزمون جامع سنجش فرم Y در حجم نمونه‌های مختلف

با توجه به نمودار داده‌های خام و سه نمودار مربوط به سه مدل موردنظر می‌توان مشاهده کرد که برای داده‌های این آزمون با حجم‌های مختلف ۲۰۰ نفری، ۵۰۰ نفری، ۱۰۰۰ نفری و حتی کل حجم نمونه بهترین مدل، مدل سوم است، چون مدلی که کمترین مقدار AIC را داشته باشد برازش بهتری دارد؛ بنابراین در نمونه ۲۰۰ نفری مدل سوم با مقدار AIC برابر با ۲۷۸/۴۲۴ بهترین مدل برای هموارسازی این مجموعه داده‌ها است؛ در نمونه ۵۰۰ نفری مدل سوم با مقدار AIC برابر با ۳۶۳/۳۲۱ بهترین مدل برای هموارسازی این مجموعه داده‌ها است؛ در نمونه ۱۰۰۰ نفری مدل سوم با مقدار AIC برابر با ۳۹۹/۸۳۷ بهترین مدل برای هموارسازی این مجموعه داده‌ها است؛ و با کل حجم نمونه مدل سوم با مقدار AIC برابر با ۱۱۵۸/۱۵ بهترین مدل برای هموارسازی این مجموعه داده‌ها است. ازاین‌رو به‌طورکلی می‌توان نتیجه گرفت که برای داده‌های آزمون سنجش مدل سوم بهترین مدل برای هموارسازی داده‌ها است.

بحث و نتیجه‌گیری

یکی از اهداف اصلی این مطالعه بررسی تأثیر الگوی لگاریتم خطی مورداستفاده در پیش هموارسازی بر روی نتایج همتراز سازی کرنل بود. به هنگام پیش هموارسازی، برنامه همترازسازی کرنل در نرم‌افزار R برونداد‌های AIC (ملاک آگاهی آکائیک) را فراهم می‌نماید. این اطلاعات می‌تواند به کاربر یا پژوهشگر برای تصمیم‌گیری در رابطه با انتخاب مدل لگاریتم خطی برای این مرحله کمک نماید. بدیهی است چنانچه مدل انتخاب‌شده تفاوت اندک یا هیچ تفاوتی را در نتایج همتراز سازی ایجاد نکند پس فرآیند انتخاب مدل می‌تواند نامربوط و یا ناکارآمد باشد.

برای پاسخگویی به این سؤال پژوهش در داده‌های آزمون تولیمو از مدل‌های لگاریتم خطی برای طرح گروه‌های ناهمسان با آزمون لنگر (داده‌های تولیمو) انتخاب و استفاده شد. این مدل‌ها شامل سه مدل به شرح زیر بوده است:

مدل اول: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) برای آزمون و لنگر

مدل دوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) برای آزمون و لنگر و فرض تعامل بین میانگین‌های آزمون و لنگر و تعامل بین واریانس‌های آزمون و لنگر

مدل سوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول) برای آزمون و لنگر و تمام تعامل‌های ممکن بین میانگین و واریانس آزمون و لنگر (در این مدل بین میانگین‌ها و واریانس‌ها هم تعامل وجود دارد).

به‌طورکلی باید اشاره کرد که روش‌های ارزیابی مدل بسیار پیچیده است، چون نیازمند تعادل بین چندین ملاک مرتبط شامل کفایت توصیفی و اکتشافی[1]، عمومیت[2]، سادگی و ابطال‌پذیری[3] است (جاکوبس و گراینگر[4]، ۱۹۹۴، نقل از لئو و آیتکین، ۲۰۰۸). در ادبیات

1. descriptive and exploratory adequacy
2. Generality
3. Falsifiability
4. Jacobs & Grainger

مدل و مدل‌سازی به تعمیم‌پذیری مدل[1] توجه ویژه‌ای شده است. امساگ‌گری[2] یا همان سادگی مدل مسئله‌ای است که همیشه مطلوب محققان بوده است. مهم‌ترین چالش فرضیه آزمایی یا انتخاب مدل، مشخص کردن مدلی است که عملکرد پیش‌بینی بهتری داشته باشد. (مایونگ، فورستر، و براون[3]، ۲۰۰۰؛ واگنمیکرز و والدورپ[4]، ۲۰۰۶). مدل‌های پیچیده‌تر نسبت به مدل‌های ساده‌تر عموماً برازش بهتری با داده‌های مشاهده‌شده دارند و به همین دلیل باید از انتخاب مدل‌هایی که شاخص برازش[5] بهتری دارند اجتناب کرد، چون این اقدام منجر به بیش‌برازش[6] خواهد شد. انتخاب مدل شامل مقایسه دو یا چند مدل است که هر یک از این مدل‌ها نمایانگر نظریه‌های مختلف دربارهٔ واقعیت است و تصمیم دربارهٔ اینکه کدام‌یک از این مدل‌ها بهترین توصیف از داده‌های موجود هستند. رویکرد غالب در انتخاب مدل در علوم رفتاری و اجتماعی، رویکرد آزمون فرضیه صفر[7] (NHST) است. این رویکرد در طی سال‌ها با انتقادات زیادی هم ازلحاظ کاربردی و هم ازلحاظ فلسفی مواجه شده است (واگنمیکرز[8]، ۲۰۰۷). برخی از برجسته‌ترین محدودیت‌های عملی استفاده از رویکرد NHST برای انتخاب مدل از قرار زیر است: محدودیت این رویکرد برای مقایسه دو مدل؛ نیاز به اینکه این دو مدل آشیانه‌ای باشند؛ و عدم امکان یافتن شواهد برای مدل صفر. یک رویکرد جایگزین برای انتخاب مدل که با این محدودیت‌ها نیز همراه نباشد توسط آکائیک (۱۹۷۴) مطرح شد که به ملاک آگاهی آکائیک[9] (AIC) مشهور

1. Model generalizability
2. Parsimony
3. Myung, Forster, & Browne
4. Wagenmakers & Waldorp
5. Goodness of fit
6. Overfitting
7. Null hypothesis significance testing (NHST)
8. Wagenmakers
9. Akaike Information Criterion (AIC)

شد[۱]. بر اساس شاخص AIC می‌توان گفت که مدلی که کمترین مقدار AIC و Residual را داشته باشد برازش بهتری دارد.

با توجه به اعداد گزارش‌شده در جداول مربوط دیده شد که در مدل اول شاخص AIC کمتر و در مدل سوم شاخص Residual کمتر است. بر همین مبنا به کمک نتایج آزمون تفاوت خی‌دو بین دو مدل اول و سوم ملاحظه شد که آزمون تفاوت خی‌دوی بین دو مدل معنادار نیست (P >۰/۰۵). لذا مدل اول (مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی- چهار گشتاور اول برای آزمون و لنگر) که ساده‌تر است انتخاب شد و بهترین مدل برای هموارسازی داده‌های آزمون تولیمو (فرم X) در سه حجم نمونه ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری مدل اول به دست آمد. به همین ترتیب، در فرم Y آزمون تولیمو در حجم نمونه ۲۰۰ و ۱۰۰۰ نفر بهترین مدل، مدل اول و در حجم نمونه ۵۰۰ نفر مدل دوم برگزیده شد.

برای پاسخگویی به این سؤال پژوهش در داده‌های آزمون جامع سنجش (دروس عمومی) از مدل‌های لگاریتم خطی برای طرح گروه‌های همسان انتخاب و استفاده شد. این مدل‌ها شامل سه مدل به شرح زیر بوده است:

مدل اول: مدل لگاریتم خطی تنها شامل میانگین (گشتاور اول)

مدل دوم: مدل لگاریتم خطی شامل میانگین و واریانس (دو گشتاور اول)

مدل سوم: مدل لگاریتم خطی شامل میانگین، واریانس، چولگی و کشیدگی (چهار گشتاور اول)

با توجه به نمودار داده‌های خام و سه نمودار مربوط به سه مدل موردنظر مشاهده شد که برای داده‌های آزمون سنجش (هم فرم X و هم فرم Y) در حجم‌های نمونه مختلف ۲۰۰،

---

۱. چندی بعد از معرفی AIC، ملاک آگاهی بیزی (BIC) توسط شوارز (۱۹۷۸) مطرح شد که به ملاک آگاهی شوارز (SIC) نیز مشهور است. بااینکه BIC شباهت زیادی به AIC دارد ولی این دو شاخص دو ریشه کاملاً متفاوت دارند: AIC در چارچوب «نظریه آگاهی» مطرح شده است ولی مبنای BIC احتمال پسین مدل است که اساساً یک مفهوم بیزی است.

۵۰۰ و ۱۰۰۰ نفری بهترین مدل، مدل سوم است چون مدلی که کمترین مقدار AIC را داشته باشد برازش بهتری دارد؛ بنابراین مدل سوم بهترین مدل برای هموارسازی این مجموعه داده‌ها انتخاب شد.

به‌طورکلی باید یادآوری کرد که ملاک‌های آگاهی بر مبنای این ایده به وجود آمده‌اند که مدل‌ها تقریبی از واقعیت هستند؛ بنابراین موضوع، یافتن یک «مدل درست» نیست بلکه مسئله اصلی یافتن بهترین مدل تقریبی از بین مجموعه مدل‌های موردنظر است (برنهام و اندرسون[1]، ۲۰۰۴). در این رابطه مدل‌هایی که کمیت AIC در آن‌ها کوچک‌تر باشد، مدل‌های بهتری هستند. زمانی که دو یا چند مدل تقریباً دارای مقدار یکسان ملاک آگاهی هستند، این مدل‌ها همه به‌طور مساوی خوب هستند. در این صورت به‌جای انتخاب بهترین مدل منفرد، می‌توان مجموعه کوچکی از مدل‌های معقول را انتخاب کرد.

درنهایت با توجه به نمودارها و جداول گزارش‌شده و نتایج حاصل می‌توان این‌گونه نتیجه‌گیری کرد که مدل‌های لگاریتم خطی نه‌تنها برای همتراز سازی کرنل سودمند هستند، بلکه احتمالاً در سایر همترازسازی‌های معادل درصدی نمره مشاهده‌شده نیز سودمند هستند. ازآنجاکه داده‌های مورد تحلیل در اینجا واقعی هستند، بنابراین ملاحظات مربوط به خطا و اشتباه و نظایر آن درنتیجه نهایی اعمال شده است.

1. Burnham & Anderson

منابع

سرمد، زهره، بازرگان، عباس، حجازی، الهه. (۱۳۸۴). روش‌های تحقیق در علوم رفتاری. تهران. نشر آگاه.

لرد، فردریک (۱۹۸۰). کاربردهای نظریه سؤال- پاسخ. ترجمه دلاور، علی و یونسی، جلیل. انتشارات رشد.

هومن، حیدر علی. (۱۳۸۰). اندازه‌گیری‌های روانی و تربیتی و فن تهیه تست. تهران: نشر پارسا.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716-723.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Cui, Z. (2006). *Two new alternative smoothing methods in equating: the cubic B-spline presmoothing method and the direct presmoothing*. Unpublished doctoral dissertation, University of Iowa, Iowa City, Iowa.

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32*(4), 334-347

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: the cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 2009, 46*(2), 135-158.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement. 11*(3), 245-262.

Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods.* Unpublished doctoral dissertation, University of North Carolina, Greensboro. Retrieved from ProQuest. (AAT 3273329).

Haberman, S. J. (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrics*, *30*, 589-600.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions*. (ACT Research Report 90-4). Iowa City, IA: American College Testing.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement. 15*(4), 391-408.

Hanson, B. A. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education, 9*(4), 305-321.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating*. ACT Research Report 94-4. Iowa City, IA: American College Testing.

Holland, P. W. & Thayer, D. T. (1981). *Section pre-equating: The Graduate Record Examination.* Program Statistics Research Technical Report No. 81-13, Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133-183.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating Methods and Practices*. New York: Springer-Verlag.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*(1), 25-44.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*(3), 257-272.

Kolen, M. J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equatingunder the common item nonequivalent populations design. *Psychometrika, 52*(1), 43-59.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52*, 362–375.

Livingston, S. A. (1993). *An empirical tryout of Kernel equating* (ETS RR-93-33). Princeton, NJ: Educational Testing Service.

Lord, F. M., (1965). A strong true score theory with applications. *Psychometrika, 30*, 239-270.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability.

Moses, T., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (ETS RR-07-15). Princeton, NJ: Educational Testing Service.

Moses. T., & Holland, P. W. (2009a). *Selection strategies for bivariate loglinear smoothing models and their effects on NEAT equating functions*. (Technical Report 09-04). Princeton, NJ: Educational Testing Service.

Moses. T., & Holland, P. W. (2010). The effects of selection strategies for bivariate loglinear smoothing models on NEAT equating functions. *Journal of Educational Measurement. 47*(1), 76-91.

Moses. T., & von Davier, A. A. (2006). *A SAS Macro for loglinear smoothing: applications and implications*. (Technical Report 06-05). Princeton, NJ: Educational Testing Service.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, 44(1–2).

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Thissen, D. M. (2003). MULTILOG for Windows (version 7.0.2327.3). Scientific Software International, Inc.

Tong, Y., & Kolen, M. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29* (6), 418-432.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, 50(2).

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement, 19*(2), 177-190.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (2003). BILOG MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items [computer software]. Chicago, IL: Scientific Software International.