Mathematical Researches
Print ISSN: 2588-2546
Online ISSN: 2588-2554
Homepage: https://jemr.khu.ac.ir/

Optimal Stratification Using Decision Trees

M. Moradi[1] ✉, E. Kasani [2]

1. Corresponding Author, Department of Statistics, Faculty of Sciences, Razi University, Kermanshah, Iran. ✉ E-mail: moradi_m@razi.ac.ir
2. Department of Statistics, Faculty of Sciences, Razi University, Kermanshah, Iran. E-mail: E-mail e.kasani@razi.ac.ir

Article Info

Article type:
Research Article

Article history:
Received: 5 May 2024
Received in revised form: 30 May 2024
Accepted: 16 June 2024
Published online: 10 July 2024

Keywords:
Strata boundaries,
Homogeneity of units within strata,
Stratified sampling Efficiency,
Penalty.

ABSTRACT

Introduction

Stratification sampling is one of the most widely used sampling methods. In some cases, it is up to the researcher to determine the boundaries of the classes, and in some cases, the society is already classified. The optimal classification for a situation is obtained from the boundary of the classes, based on those boundaries, the variance of the average (or total) estimator of the community reaches its minimum value. In traditional methods, the variance of the estimator is considered as a function of the boundaries of the response variable, and in order to achieve the minimum of the variance, equations are obtained, which are often solved by numerical methods.

Considering that one of the boundaries of stratum h, i.e. y_h, appear only in two terms W_hS_h and $W_{h+1}S_{h+1}$, we have

$$\frac{\partial(\sum_h W_hS_h)}{\partial y_h} = \frac{\partial(W_hS_h)}{\partial y_h} + \frac{\partial(W_{h+1}S_{h+1})}{\partial y_h}$$

Now, if $f(y)$ is the density function of y, the weight of the stratum can be written based on it, as follows;

$$\mathrm{W_h} = \int_{\mathrm{y_{h-1}}}^{\mathrm{y_h}} f(t)dt \qquad\qquad \frac{\partial(W_h)}{\partial y_h} = f(y_h)$$

Therefore, the derivative can be written as follows.

$$\frac{\partial(W_hS_h)}{\partial y_h} = S_h\frac{\partial(W_h)}{\partial y_h} + W_h\frac{\partial(S_h)}{\partial y_h} = \frac{f(y_h)}{2}\frac{(y_h-\bar{Y}_h)^2+S_h^2}{S_h}$$

Similarly, a similar formula is obtained for the next stratum, and finally for $h = 1,2,\dots,H-1$, we have

$$\frac{f(y_h)}{2}\frac{(y_h-\bar{Y}_h)^2+S_h^2}{S_h} = \frac{f(y_h)}{2}\frac{(y_h-\bar{Y}_{h+1})^2+S_{h+1}^2}{S_{h+1}}$$

These equations do not have a closed solution and are solved using approximate numerical methods.

The first disadvantage of this optimal classification style is not considering independent variables. The second disadvantage is complicated equations that

do not have a closed and understandable answer. Since the objectives of multiple data mining and the role of classification are not limited to the estimation of one parameter, it is necessary to seek to classify the observations into homogeneous subgroups so as to help more interpretability of the data.

The decision tree, which is basically a prediction method, partitions the data into subgroups with the highest purity based on quantitative or qualitative auxiliary variables in successive branches. The simple interpretability of the decision tree is more prominent than its predictive accuracy.

Material and Methods

In this paper, we have tried to build the optimal classification based on the new criterion, $C = Var(\bar{y}_{st}) + \lambda H,$ that is a combination of variance and a penalty for increasing the number of strata, so that important auxiliary variables in the formation of the decision tree determine the limits of the strata. In the introduced criterion the penalty term is λH, where H is the number of strata and λ is a tunning parameter. For zero value of λ, only the first term of the criterion, which is the variance, will play a role. As the tuning parameter λ increases, the number of optimal starts is reduced from 128 to 84, 6, and 2.

Results and discussion

By pruning the saturated tree successively until reaching the root node, different values of the variance of the estimator are obtained. By drawing a plot including the variance versus the number of strata, similar to the Figure 2 on the top- left, a curve is obtained that has the highest concavity in the corner. Using the information obtained from this corner and the number of strata related to it will lead us to choose the appropriate value of λ.

The classification process starts from the saturated tree and with successive pruning until reaching the tree stump, the number of strata decreases, the optimal classification is achieved based on the introduced combined criteria.

Conclusion

By applying the established criteria to the real population of Razi University students' academic progress, which includes the response variable of academic progress and twelve auxiliary variables in Table (1), the optimal classification with six strata was the result. The decision tree that produces this classification is shown in Figure 3.

How to cite: Moradi, M., & Kasani, E. (2024). Optimal Stratification Using Decision Trees, *Mathematical Researches*, **10** (2), 51 – 65.

Publisher: Kharazmi University

شاپا چاپی: ۲۰۴۶-۲۵۸۸
شاپا الکترونیکی: ۲۵۵۴-۲۵۸۸

پژوهش‌های ریاضی

Homepage: https://mmr.khu.ac.ir/

طبقه‌بندی بهینه با استفاده از درخت تصمیم

محمد مرادی[1]✉، الناز کسانی[2]

۱. نویسنده مسئول، گروه آمار، دانشکده علوم، دانشگاه رازی، کرمانشاه، ایران. رایانامه: abcdef@khu.ac.ir
۲. گروه آمار، دانشکده علوم، دانشگاه رازی، کرمانشاه، ایران. رایانامه: abcdef@khu.ac.ir

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۲/۱۶
تاریخ بازنگری: ۱۴۰۳/۳/۱۰
تاریخ پذیرش: ۱۴۰۳/۳/۲۷
تاریخ انتشار: ۱۴۰۳/۰۴/۲۰

واژه‌های کلیدی:
حدود طبقات،
همگنی واحدها داخل طبقات،
کارایی نمونه‌گیری طبقه‌بندی،
جریمه.

چکیده

نمونه‌گیری طبقه‌بندی یکی از پرکاربردترین روش‌های نمونه‌گیری است. در برخی موارد تعیین حدود طبقات به عهده محقق است و در مواردی جامعه از قبل طبقه‌بندی شده است. طبقه‌بندی بهینه به ازای وضعیتی ازحدود طبقات حاصل می‌شود که بر اساس آن حدود، واریانس برآوردگر میانگین (یا مجموع) جامعه به کمترین مقدار خود برسد. در روش‌های سنتی، واریانس برآوردگر را تابعی از حدود طبقات متغیر پاسخ در نظر می‌گیرند و جهت رسیدن به کمینه واریانس معادلاتی حاصل می‌شود که غالبا از روش‌های عددی به جواب می‌رسد. عیب اول این سبک طبقه‌بندی بهینه، در نظر نگرفتن متغیرهای مستقل است. عیب دوم، معادلات پیچیده‌ای است که جوابی بسته و قابل فهم ندارند. از آنجا که اهداف داده‌کاوی چندگانه و نقش طبقه‌بندی فقط محدود به برآورد یک پارامتر نیست، باید به دنبال رده‌بندی مشاهدات به زیرگروه‌های همگن بود طوری که به تفسیرپذیری بیشتر داده‌ها نیز کمک شود. درخت تصمیم که در اصل روشی جهت پیش‌بینی است، در انشعابات متوالی داده‌ها را براساس متغیرهای کمکی کمی یا کیفی به زیرگروه‌هایی با بیشترین خلوص افراز می‌کند. تفسیرپذیری ساده درخت تصمیم برجسته‌تر از توان پیش‌بینی آن است.

در این مقاله، سعی کرده‌ایم طبقه‌بندی بهینه را بر اساس معیار جدیدی که ترکیبی از واریانس و جریمه‌ای بر زیاد شدن تعداد طبقات است بسازیم طوری که متغیرهای کمکی با اهمیت در تشکیل درخت تصمیم حدود طبقات را مشخص می‌کنند. فرایند طبقه‌بندی از درخت اشباع‌شده شروع و با هرس‌های متوالی تا رسیدن به کنده درخت، تعداد طبقات کم می‌شود، طبقه‌بندی بهینه بر اساس معیار ترکیبی معرفی شده حاصل می‌شود.

استناد: مرادی، محمد؛ و کسانی، الناز (۱۴۰۳). طبقه‌بندی بهینه با استفاده از درخت تصمیم. *پژوهش‌های ریاضی*، ۱۰ (۲)، ۵۱ – ۶۵.

ناشر: دانشگاه خوارزمی

۱. مقدمه

هنگامی که جامعه آماری از چند زیر جامعه تشکیل شده باشد و مدل‌سازی یا استنباط آماری برای هر کدام از زیر جامعه‌ها جداگانه مورد نیاز باشد، طبقه‌بندی کردن جامعه به شکل هر زیر جامعه یک طبقه و گرفتن نمونه‌های مستقل از طبقات تحت عنوان نمونه‌گیری طبقه‌بندی مرسوم بوده است. امروزه در مباحث داده‌کاوی[1] و مواجه با انبوهی از مه داده‌ها[2] از طبقه‌بندی کردن داده‌ها بیشتر به منظور کم کردن پیچیدگی استخراج الگو، از زیر مجموعه‌های کوچکتر با رفتار مشابه تحت عنوان یک طبقه به جای استخراج یک الگوی واحد از کل داده، استفاده می‌شود.

نحوه طبقه‌بندی کردن به صورت بهینه به معیار مورد نظر تحلیل بستگی دارد. در گذشته بیشتر هدف براورد یک پارامتر با درستی[3] قابل قبول بوده است، پس منطقی است که معیار واریانس و هدف رسیدن به کمینه آن باشد. در مواجه با مه‌داده هدف پیچیده‌تر و فراتر از فقط برآورد یک پارامتر مانند میانگین یا مجموع جامعه است و استخراج و تعمیم نتایج تحلیل تا حد امکان باید ساده و قابل فهم باشد. با این استدلال، در طبقه‌بندی داده‌ها در کنار معیار درستی باید نیم نگاهی به ساده بودن ساختار طبقه‌بندی داشت. به عنوان مثال، طبقه‌بندی داده‌ها به چهار طبقه بر اساس دو عامل جنسیت و تحصیلات در دو رده تحصیلات کم و زیاد یک طبقه‌بندی ساده و طبقه‌بندی با تعداد طبقات بیشتر و متغیرهای کمکی مانند سن (۰تا ۵، ۵ تا ۱۵،۱۵ تا ۳۰ و ۳۰ و بالاتر)، درآمد (۰ تا ۵۰۰، ۵۰۰ تا ۲۰۰۰، ۲۰۰۰ تا ۵۰۰۰ و ۵۰۰۰ به بالا) و غیره مثالی از یک طبقه‌بندی با ساختاری پیچیده‌تر است.

ساراندال و همکاران ۱۹۹۲ نمونه‌گیری طبقه‌بندی را بدین شکل تعریف کرده‌اند «... در یک طرح نمونه‌گیری طبقه‌ای، جامعه به زیرجامعه‌های غیرهم‌پوشانی به نام طبقات تقسیم می‌شود. یک نمونه احتمالی در هر طبقه انتخاب می‌شود. انتخاب در طبقات مختلف مستقل است. طبقه‌بندی افرازی از واحدهای جامعه است و معمولاً با توجه به مقادیر یک یا چند متغیر کمکی موجود برای همه واحدهای جامعه تعریف می شود.»

دالنیوس در سال ۱۹۵۰ بر اساس متغیر پاسخ اقدام به تعیین حدود طبقات نمود طوری که واریانس برآوردگر تحت تخصیص نیمن کمینه شود. سینگ و سوخاتمه در سال ۱۹۶۹بر اساس متغیر کمکی و تحت تخصیص‌های نیمن و متناسب اقدام به تعیین حدود طبقه‌بندی بهینه نمودند. سینگ و سوخاتمه در سال ۱۹۷۱ طبقه‌بندی بهینه را برای براوردگر نسبتی و رگرسیونی مورد بررسی قرار دادند.

خان و همکاران سال ۲۰۱۴ مسئله تعیین طبقه‌بندی بهینه یک متغیرپاسخ را بر اساس متغیر کمکی که از توزیع یکنواخت پیروی کند، به عنوان یک مسئله برنامه‌ریزی غیرخطی در نظر گرفته و با استفاده از فن برنامه نویسی پویا[4] حل کردند.

ردی و همکاران ۲۰۱۸ به جای استفاده از متغیرهای کمکی مانند مناطق جغرافیایی یا سایر شرایط طبیعی مانند سن، جنسیت و غیره، طبقه‌بندی بهینه را جهت به حداکثر رساندن دقت برآورد در تخصیص نیمن و با استفاده از روش برنامه‌نویسی پویا بر روی شاخص‌های سلامت نظیر تخمین میزان هموگلوبین در زنان بررسی کم‌خونی فقر آهن بررسی کردند.

[1] Data mining
[2] Big data
[3] Accuracy
[4] Dynamic

گوپت و اهامد سال۲۰۲۲تحت مدل ابرجامعه[1] رگرسیون ناهمسان، طبقه‌بندی بهینه برای یک متغیر کمکی تعمیم یافته با تخصیص متناسب را مورد مطالعه قرار دادند.

دانیش و همکاران سال ۲۰۲۳ از دو متغیرپاسخ و یک متغیر کمکی به عنوان متغیر طبقه‌بندی استفاده کردند و نمونه را با استفاده از متغیر طبقه‌بندی با ترکیبی از برآوردگرهای نسبتی و حاصل‌ضرب انتخاب کردند. تحت مجموعه‌ای از ابر جامعه‌ها، حداقل معادلات را از طریق کمینه‌سازی واریانس انباشته متغیرهای پاسخ و با رویکرد برنامه‌نویسی پویا به دست آوردند.

در تمام کارهای انجام شده فوق، هدف تعیین حدود طبقات بر اساس متغیرپاسخ یا متغیر کمکی کمی است که همبستگی بالایی با متغیر پاسخ داشته باشد. در ابرجامعه‌ها برای متغیر پاسخ توزیع احتمالی در نظر گرفته می‌شود و بیشتر جنبه نظری دارد زیرا داده‌های واقعی ممکن است از توزیع‌های فرض شده پیروی نکنند. با توجه به پیچیده شدن حل معادلات، غالبا از برنامه نویسی پویا جهت تقریب جواب معادلات استفاده شده است.

در این مقاله بدون فرض توزیع خاصی برای متغیر پاسخ و با استفاده از متغیرهای کمکی کمی یا کیفی که در ساخت درخت تصمیم قابل استفاده هستند، اقدام به طبقه‌بندی جامعه می‌کنیم. پس از ساختن درخت تصمیم که الگوریتمی ساده و قابل فهم برای همه دارد، تصور حدود طبقات کوچک با تعداد زیاد و سپس ادغام آن‌ها جهت رسیدن به تعداد کمتری طبقه بزرگ کار ساده‌ای است که می‌تواند مورد استقبال کاربران نه چندان مسلط به ریاضیات هم قرار گیرد.

در بخش دوم، جامعه موردمطالعه معرفی می‌شود که دانشجویان دانشگاه رازی هستند. موفقیت تحصیلی آن‌ها را از داده‌های آموزشی سیستم گلستان سال‌های ۱۳۷۵ تا ۱۴۰۱ مورد بررسی قرار داده‌ایم. به همراه متغیر پاسخ موفقیت تحصیلی، متغیرهای کمکی جنسیت، دانشکده، معدل دیپلم، مقطع و مقیاس بومی بودن هم جمع‌اوری شده اند که در تعیین طبقات قابل فهم مدیران اجرایی می‌توانند کارآمد و مفید باشند. در بخش سوم، روش‌های موجود تعیین طبقه‌بندی بهینه بر اساس متغیر پاسخ بیان شده است. در بخش چهارم، طبقه‌بندی بهینه را بر اساس درخت تصمیم و معیاری ترکیبی از درستی و تعداد طبقات معرفی کرده‌ایم و در نهایت بخش پنجم نتیجه گیری نوشته شده است.

۲. جامعه آماری و متغیرهای مربوطه

در برآورد میزان پیشرفت تحصیلی دانشجویان، غیر از اهمیت درستی برآوردها، براورد این پارامتر به تفکیک زیرجامعه‌های قابل فهم و به تعداد کم هم اهمیت دارد. در این بخش، جامعه آماری، متغیر پاسخ پیشرفت تحصیلی و متغیرهای کمکی که می‌توان از آن‌ها در طبقه‌بندی جامعه استفاده کرد، مورد بررسی قرار می‌گیرند.

[1] Superpopulation

۲,۱ جامعه هدف،

آن دسته از دانشجویان دانشگاه رازی هستند که اطلاعات آموزشی آن‌ها در سیستم گلستان ثبت شده باشد. گزارش‌های اولیه شامل ۶۰۹۲۲ دانشجو در مقطع کاردانی، کارشناسی، کارشناسی‌ارشد و دکتری از سال ۱۳۵۷ تا ۱۴۰۱ بود. شاخص‌پیشرفت تحصیلی در مقاطع مختلف تفاوت زیادی با هم دارند. به عنوان مثال، در مقطع کارشناسی شاخص‌های آموزشی از قبیل معدل فارغ‌التحصیلی، تعداد ترم مشروطی و تعداد واحد افتاده اهمیت زیادی دارند درحالی‌که در مقاطع بالاتر در کنار شاخص آموزشی، شاخص پژوهشی هم اهمیت پیدا می‌کند. بر این اساس و با توجه به اینکه اکثریت دانشجویان پذیرفته شده مقطع کارشناسی بودند، دانشجویان تحصیلات تکمیلی را مورد بررسی قرار ندادیم، همچنین اطلاعات سیستم گلستان از سال ۱۳۷۵ به قبل ناقص بود و به ناچار آن‌ها را نیز در نظر نگرفتیم. با این اصلاحات، جامعه آماری مورد مطالعه به ۳۰۴۳۴ دانشجو رسید که در مقاطع کاردانی و کارشناسی از سال ۱۳۷۵ تا ۱۴۰۱ در این دانشگاه تحصیل کرده‌اند.

۲,۲ متغیرهای مورد مطالعه

داده‌های موجود در سیستم گلستان اغلب به‌صورت پردازش نشده و ناقص هستند و قبل از انتقال به نرم‌افزارهای آماری نیاز به پاکسازی و ویرایش دارند. برخی از نواقص موجود در داده‌ها شامل، وجود داده‌های قدیمی، اضافی، گمشده و... است. پردازش اولیه داده‌ها در نرم‌افزار اکسل انجام شد تا انحرافات، مقادیر گمشده، مقادیر ثبت نشده و... مشخص شوند. بعد از آماده‌سازی داده‌ها در نرم‌افزار اکسل آماده ورود به نرم‌افزار آماریRشدند. علاوه بر متغیرهایی که در سیستم ثبت شده بود متغیر مقیاس بومی بودن با استفاده از متغیر استان محل زندگی، محاسبه و به مجموعه داده اضافه شد. مشحصات مربوط به تعدادی از متغیرهایی که در این مقاله استفاده شده‌اند در جدول ۱ آمده است.

مقیاس بومی متغیری رتبه‌ای است و با استفاده از کد استان و شهرستان محل سکونت به ۶ گروه استان کرمانشاه، شهرستان‌های کرمانشاه، استان‌های همجوار، استان‌های دور، استان‌های خیلی دور و خارج ازکشور تبدیل شد.

پیشرفت تحصیلی در بخش آموزش چند بعدی است و تنها براساس معدل خام قابل ارزیابی نیست. به عنوان مثال، دو هم‌کلاسی را در نظر بگیرید که در ترمی همزمان یک درس را اخذ می‌کنند. یکی از آن‌ها با نمره ۱۶ درس را گذرانده و دیگری در آن درس می‌افتد. اما در ترم‌های بعد دانشجوی افتاده هم با نمره ۱۶درس را می‌گذراند، در محاسبه معدل درس مورد نظر برای هر دو نفر ۱۶ لحاظ می‌شود درحالی‌که میزان پیشرفت آن دو نفر در پاس‌کردن این درس برابر نبوده است. بر مبنای همین استدلال نسبت واحدهای پاس شده به اخذ شده هم جزو شاخص پیشرفت در نظر گرفته شده است. همچنین فرض کنید دانشجویی چند ترم متوالی همه واحدهای اخذ شده را پاس ولی مشروط شده و در شرف اخراج از دانشگاه قرار می‌گیرد اما در ادامه کار با تلاش معدل خود را بهبود می‌دهد و فرضا با معدل ۱۵ فارغ‌التحصیل می‌شود. همزمان هم‌کلاسی او را در نظر بگیرید که به طور یکنواخت هر ترم با معدل ۱۵ و بدون مشروطی فارغ‌التحصیل شود. میزان پیشرفت این دو نفر با شاخصی مانند نسبت ترم‌های مشروطی به ترم‌های ثبت نام شده، منطقی‌ترقابل مقایسه خواهد بود. به‌دلیل اینکه تعداد ترم اخذ شده در مجموعه داده‌ها وجود نداشت متغیر y_3همان تعداد ترم مشروطی در نظر گرفته شد.

جدول ۱. مشخصات متغیرهای مهم از بین متغیرهای اخذ شده ازسیستم گلستان

ردیف	نوع	نام متغیر	نوع متغیر	درصدگمشده
۱	عمومی	کد استان محل سکونت	کیفی	۴۲٪
۲	عمومی	ملیت(ایرانی - غیر ایرانی)	کیفی	۰٪
۳	عمومی	جنسیت(مرد -زن)	کیفی	۰٪
۴	آموزشی	تعداد واحداخذ شده	کمی	۰٪
۵	آموزشی	تعداد واحد گذرانده	کمی	۰٪
۶	آموزشی	تعداد مشروطی	کمی	۰٪
۷	آموزشی	معدل کل	کمی	۰٪.۳۱
۸	آموزشی	معدل دیپلم	کمی	۹۰٪.۲۶
۹	آموزشی	مقطع	کیفی	۰٪
۱۰	آموزشی	آخرین وضعیت دانشجو(فارق التحصیل، اخراجی، انتقالی و...)	کیفی	۴۱٪
۱۱	آموزشی	کددانشکده	کیفی	۰٪
۱۲	آموزشی	کد رشته	کیفی	۰٪

نکته دیگر معدل خام است که از یک رشته به رشته دیگر و حتی یک رشته در چند سال مختلف متغیر است. جهت بی‌اثر کردن اثر رشته و سال ورود، معدل دانشجویان در زیر گروه ورودی-رشته‌های مختلف استاندارد شده است. به منظور تلخیص سه متغیر معدل استاندارد شده، نسبت واحد پاس شده و نسبت ترم مشروطی در یک بعد، از روش مولفه‌اصلی استفاده شده است.

$$y_1 = \text{معدل کل استاندارد شده به تفکیک ورودی-رشته},\ y_2 = \frac{\text{تعداد واحد پاس شده}}{\text{تعداد واحد اخذ شده}},\ y_3 = \text{تعداد ترم مشروطی}$$

در روش مولفه‌اصلی، مولفه‌اصلی اول ترکیبی خطی از متغیرها است که بیشترین واریانس ممکن را دارد و به عبارتی بیشترین اطلاعات ممکن را در مورد سه متغیر پاسخ در خود حفظ می‌کند. جهت محاسبه پیشرفت تحصیلی، راه دیگری تحت عنوان میانگین موزون با وزن برابر وجود دارد. از معایب این روش این است که متغیرهای کم اهمیت و پراهمیت از لحاظ میزان واریانس را یکسان در نظر می‌گیرد و به اندازه روش مولفه اصلی توان تفکیک پیشرفت تحصیلی افراد را ندارد.

شاخص پیشرفت تحصیلی براساس معدل خام، با توجه به تغییرات معدل در ورودی‌ها و رشته‌های مختلف، شاخص مناسبی نیست. پس شاخص استاندارد شده براساس ورودی-رشته مبنای محاسبات بعدی قرار گرفت و به منظور تفسیرپذیری بیشتر آن را در دامنه صفر تا صد مقیاس‌بندی کرده‌ایم.

$$\boldsymbol{p = 0.5188y_1 + 0.0741y_2 - 0.8516y_3}$$

$$\boldsymbol{y = \frac{p - min(p)}{max(p) - min(p)} * 100}$$

۳. طبقه‌بندی بهینه با استفاده از متغیر پاسخ

فرض کنید متغیر پاسخ y در بازه (a,b) قرار گرفته و هدف ساختن H طبقه است. در این بخش، به دنبال یافتن آن حدود $a \leq y_1, y_2, \dots, y_{H-1} \leq b$ هستیم که واریانس برآوردگر $\bar{y}_{st}$ را کمینه کند. اگر از هر طبقه با اندازه N_h نمونه‌ای به اندازه n_h گرفته شود، واریانس برآوردگر $\bar{y}_{st}$ به صورت زیر خواهد بود.

$$Var(\bar{y}_{st}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h} = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

که $\sigma_h^2 N_h = S_h^2(N_h - 1)$، جملات σ_h^2 و S_h^2 هر دو به عنوان واریانس طبقه hام در نظر گرفته می‌شوند و در طبقات بزرگ تفاوت قابل چشم‌پوشی با هم دارند. در تخصیص نیمن که اندازه نمونه هر طبقه متناسب با اندازه طبقه و واریانس طبقه در نظر گرفته می‌شود، فرمول واریانس را می‌توان به صورت زیر ساده کرد.

$$Var(\bar{y}_{st}) = \frac{1}{n}\left(\sum_h \frac{N_h}{N} S_h\right)^2 - \frac{1}{N}\sum_h W_h S_h^2 = \frac{1}{n}\left(\sum_h \frac{N_h}{N} S_h\right)^2 - \frac{1}{N}\sum_h \frac{N_h}{N} S_h^2$$

$$= \frac{1}{n}\left[\sum_h W_h S_h\right]^{2^2} - \frac{1}{N}\sum_h W_h S_h^2$$

که $W_h = {}^{N_h}/_{N}$ وزن طبقه است. با فرض اینکه جامعه آنقدر بزرگ است که می‌توان ضریب تصحیح جامعه متناهی را عدد یک فرض کرد، جمله دوم حذف می شود و با کمینه کردن $\sum_h W_h S_h$ به هدف خود می‌رسیم. همچنین با توجه به اینکه حدود طبقه h یعنی y_h فقط در دو جمله $W_h S_h$ و $W_{h+1} S_{h+1}$ ظاهر می‌شود، داریم

$$\frac{\partial(\sum_h W_h S_h)}{\partial y_h} = \frac{\partial(W_h S_h)}{\partial y_h} + \frac{\partial(W_{h+1} S_{h+1})}{\partial y_h}$$

حال اگر $f(y)$ تابع چگالی yباشد، می‌توان وزن طبقه را بر اساس آن نوشت.

$$\mathrm{W_h} = \int_{\mathrm{y_{h-1}}}^{\mathrm{y_h}} f(t)dt$$

$$\frac{\partial(W_h)}{\partial y_h} = f(y_h)$$

لذا مشتق را می‌توان به صورت زیر نوشت.

$$\frac{\partial(W_h S_h)}{\partial y_h} = S_h \frac{\partial(W_h)}{\partial y_h} + W_h \frac{\partial(S_h)}{\partial y_h} = \frac{f(y_h)}{2} \frac{(y_h - \bar{Y}_h)^2 + S_h^2}{S_h}$$

به طور مشابه برای طبقه بعدی نیز می‌توان نوشت،

$$\frac{\partial(W_{h+1} S_{h+1})}{\partial y_h} = -\frac{f(y_h)}{2} \frac{(y_h - \bar{Y}_{h+1})^2 + S_{h+1}^2}{S_{h+1}}$$

حدود طبقات بهینه به ازای $h = 1,2, \ldots, H-1$از حل معادلات زیر به دست می‌آید.

$$\frac{f(y_h)}{2} \frac{(y_h - \bar{Y}_h)^2 + S_h^2}{S_h} = \frac{f(y_h)}{2} \frac{(y_h - \bar{Y}_{h+1})^2 + S_{h+1}^2}{S_{h+1}}$$

این معادلات جواب بسته‌ای ندارند، زیرا جملات $\bar{Y}$ و S^2 هردو به y_h وابسته‌اند. دالنیوس و هاج سال ۱۹۵۷ راه‌حلی تقریبی برای رسیدن به کمینه $\sum_h W_h S_h$ ارایه کردند. فرض کنید تابع $Z(y)$ از رابطه زیر حاصل شود.

$$Z(y) = \int_a^{\mathrm{y}} \sqrt{\mathrm{f(t)}}\,\mathrm{dt}$$

اگر تعداد طبقات زیاد و طول طبقات کم باشد، منطقی است که تابع $\mathrm{f}(y)$ درون طبقه ثابت و توزیع یکنواخت در نظر گرفته شود.

$$\mathrm{W_h} = \int_{\mathrm{y_{h-1}}}^{\mathrm{y_h}} f(t)dt \cong f_h(\mathrm{y_h} - \mathrm{y_{h-1}})$$

$$S_h \cong \frac{\mathrm{y_h} - \mathrm{y_{h-1}}}{\sqrt{12}}$$

$$Z_h - Z_{h-1} = \int_{\mathrm{y_{h-1}}}^{\mathrm{y_h}} \sqrt{\mathrm{f(t)}}\,\mathrm{dt} \cong \sqrt{\mathrm{f_h}}(\mathrm{y_h} - \mathrm{y_{h-1}})$$

که f_h مقداری ثابت از تابع fدر طبقه hام است. با استفاده از روابط بالا داریم،

$$\sqrt{12}\sum_{h} W_h S_h \cong \sum_{h} f_h\,(\mathrm{y_h}-\mathrm{y_{h-1}})^2 = \sum_{h}(\mathrm{Z_h}-\mathrm{Z_{h-1}})^2$$

چون $Z(b)-Z(a)$ مقداری ثابت است، سمت راست معادله بالا زمانی کمینه می‌شود که همه جملات $\mathrm{Z_h}-\mathrm{Z_{h-1}}$ با هم برابر باشند.

از مطالب فوق روش طبقه‌بندی بهینه را بدین شکل معرفی کرده‌اند که از فراوانی ها جذر گرفته و سپس مقدار تجمعی طبقات محاسبه شود. سپس مقدار کل به تعداد طبقات مورد نظر تقسیم شود. بازه‌های مساوی تشکیل یافته از مقدار تجمعی جذر فراوانی‌ها، حدود بهینه طبقات را مشخص می‌سازد.

۴. طبقه‌بندی بهینه با استفاده از درخت تصمیم

در بخش قبلی، یکی از مرسوم‌ترین روش‌های طبقه‌بندی بهینه معرفی شد. روش‌های دیگر هم تا حدی مشابه همین روش هستند. در این بخش به کمک ساختن درخت تصمیم ابتدا اقدام به افراز مشاهدات به زیرگروه‌های همگن به تعداد زیاد نموده و سپس با هرس کردن[1] درخت، تعداد گروه‌های پایانی[2] را کاهش می‌دهیم. هرکدام از این گروه‌ها هم‌ارز زیرمجموعه‌ای از فضای متغیرهای کمکی است و یک طبقه را مشخص می‌کنند.

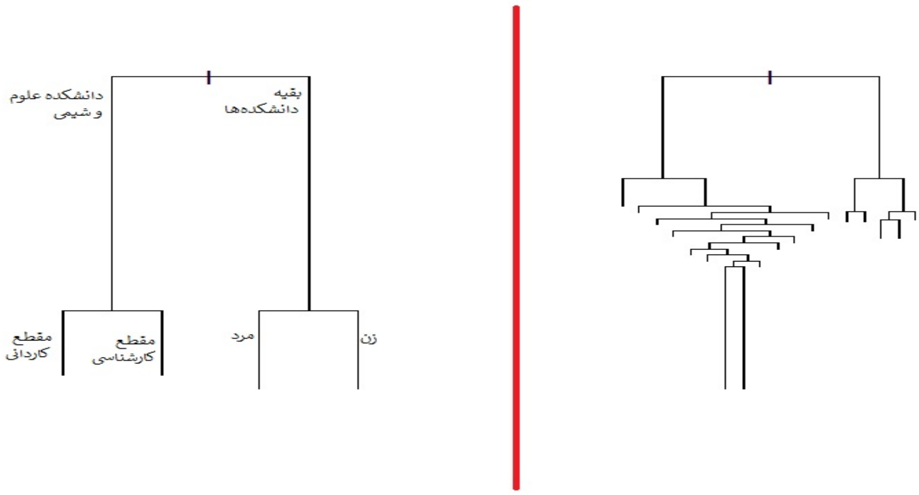
درخت‌های تصمیم را می‌توان برای هردو مسئله رگرسیون و رده‌بندی[3] به کار برد، لذا در مباحث نمونه‌گیری از ساختن طبقات بر اساس درخت می‌توان به هردو منظور برآورد میانگین یا مجموع جامعه و برآورد نسبت جامعه به ترتیب با درخت رگرسیونی و درخت رده‌بندی کمک گرفت.

در این مقاله از روش انشعاب دوتایی بازگشتی[4]برای ساخت درخت استفاده می‌شود. رویکرد این روش از بالا به پایین است، زیرا از بالای درخت شروع می شود (در این نقطه همه مشاهدات متعلق به یک طبقه واحد هستند) و سپس به طور متوالی فضای پیش بینی افراز می‌شود. هر انشعاب از طریق دو شاخه جدید در پایین‌تر درخت نشان داده می‌شود، تا در نهایت تعداد مشاهدات هر انشعاب به مقداری از قبل تعیین شده برسد و فرایند انشعاب‌سازی متوقف می‌شود. به درختی که بیشترین انشعابات روی آن انجام شده درخت اشباع[5] شده می‌گویند.

[1] Pruning
[2] Terminal nodes
[3] Classification
[4] Recursive binary splitting
[5] Saturated tree

هنگام پیش‌بینی با درخت تصمیم جهت اجتناب از بیش‌برازشی[1]با هرس کردن درخت اقدام به کوچک‌تر کردن آن می‌شود. در این تحقیق، به طور مشابه، اجازه می‌دهیم درخت تا حد اشباع، بزرگ شود و تعداد زیادی گره پایانی به دست آید. هرکدام از این گره‌ها زیر گروهی از مشاهدات را در بر دارد که برای ما نقش یک طبقه دارند. سپس با هرس کردن شاخه‌های اضافی، که به نوعی ادغام گره‌های پایانی در همدیگر است، به طبقاتی با اندازه بزرگتر دست می‌یابیم. در هرس کردن آخرین انشعاب‌های تولید شده مورد توجه قرار می‌گیرند، به ازای صرفنظر کردن از هر کدام از انشعاب‌ها مقدار مجموع توان دوم خطا در درخت رگرسیون (و نرخ خطا در درخت رده‌بندی) محاسبه می‌شود، انشعابی که با حذفش بیشترین درستی و یا کمترین خطا نتیجه شود به عنوان گزینه هرس در نظر گرفته می‌شود. دو مورد از طبقه‌بندی در چهار و هیجده طبقه در شکل (۱) نمایش داده شده است.

این عمل را تا رسیدن به کنده درخت (گره ریشه)[2] ادامه می‌دهیم که مفهوم حذف طبقه‌بندی و ادغام همه مشاهدات در یک زیر گروه است. در ادامه بر اساس معیار دقت برآوردگر که جمله جریمه تعداد طبقات هم به آن اضافه می‌شود، طبقه‌بندی بهینه را به دست می‌آوریم.



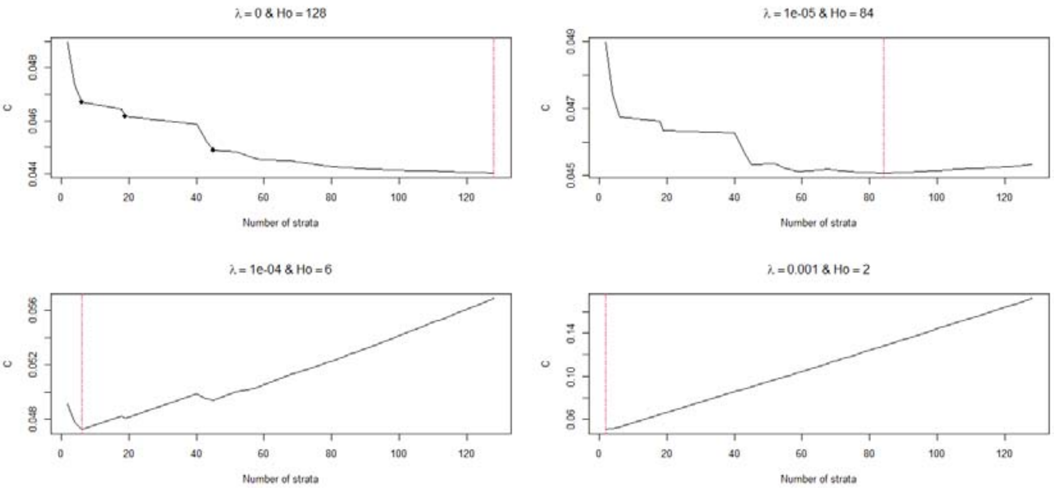
شکل ۱: دو درخت تصمیم هرس شده. سمت چپ جامعه را به چهار طبقه و سمت راستی جامعه را به هیجده طبقه افراز می‌کند.

[1]Overfitting
[2]Root node

تاثیر تعداد طبقات بر واریانس برآوردگر را با فرض استفاده از تخصیص متناسب مورد بررسی قرار می‌دهیم. در تخصیص متناسب اندازه نمونه هر طبقه متناسب با اندازه طبقه در نظر گرفته می‌شود، یعنی $n_h = n^{N_h}/_N$. در تخصیص متناسب واریانس برآوردگر به صورت زیر ساده می‌شود.

$$Var(\bar{y}_{st}) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} = \frac{N-n}{Nn} \sum_h W_h S_h^2$$

در فرمول ساده شده بالا دو جمله وزن طبقه W_h و واریانس طبقه S_h^2 با تغییر طبقات تغییر می‌کنند. با زیاد شدن تعداد طبقات مقادیر وزن کم ولی تعدادشان افزایش می‌یابد، لذا با کم و زیاد شدن تعداد طبقات W_hها اثر چندانی بر $Var(\bar{y}_{st})$ نخواهند داشت. اما جملات واریانس طبقه S_h^2 نقش اساسی در تغییرات $Var(\bar{y}_{st})$ خواهند داشت، زیرا $\sum_h W_h S_h^2$ عملا میانگین موزونی از واریانس طبقات است. در درخت تصمیم با زیاد شدن تعداد انشعاب‌ها، ناخالصی[1] مشاهدات داخل گره‌ها کاهش می‌یابد. به عبارتی دیگر مشاهدات داخل گره‌ها همگن‌تر می‌شوند که همان مفهوم کاهش واریانس طبقه و در نهایت کاهش میانگین موزون $\sum_h W_h S_h^2$ را دارد. به راحتی می‌توان تصور کرد که در طبقه‌بندی حاصل از درخت تصمیم با افزایش تعداد انشعاب‌ها، واریانس $Var(\bar{y}_{st})$ کاهش پیدا می‌کند.



شکل ۲: تغییرات معیار C به عنوان تابعی از تعداد طبقات به ازای چهار مقدار متفاوت λ. خط افقی نقطه‌چین قرمز مکان تعداد طبقه بهینه Ho در هر نمودار را نشان می‌دهد. نقاط توپر شکل بالا سمت چپ کنج‌های منحنی را در نقاط ۶، ۱۹ و ۴۵ نمایش داده اند.

[1]Impurity

از طرفی نمونه‌گیری کردن از تعداد زیادی طبقه، چندان مطلوب نیست و در عمل سعی می‌شود تعداد طبقات تاحد امکان کم باشند. بر این اساس معیاری تعریف می‌کنیم که در کنار کاهش واریانس $Var(\bar{y}_{st})$ جریمه‌ای هم بر زیاد شدن تعداد طبقات اعمال کنیم. معیار ترکیبی C به صورت زیر است.

$$C = Var(\bar{y}_{st}) + \lambda H \qquad 1-4$$

که H تعداد طبقات و در اصل همان اندازه درخت است و λ پارامتر تنظیم است. با انتخاب $\lambda = 0$ یعنی هیچ محدودیتی در تعداد طبقات وجود نخواهد داشت و هدف اصلی فقط کمینه‌کردن واریانس است. با انتخاب مقدار بزرگی برای λ اهمیت بیشتر به تعداد کم طبقات داده می‌شود نه واریانس. مقدار λ می‌تواند به صورت پیش فرض بر اساس میزان اهمیت تعداد طبقات به دقت تعیین شود. همچنین می‌توان بر اساس نقطه کنج منحنی مشابه شکل ۲ بالا سمت راست، تعداد طبقات بهینه و سپس از رابطه (۴-۱) مقدار بهینه λ را تعیین کرد.

مقدار λ باید با دقت طوری انتخاب شود که همزمان که واریانس تا حد مطلوبی کاهش داده شده، تعداد طبقات هم به صورت غیرضروری افزایش نیابد.

بر اساس جامعه آماری معرفی شده در بخش (۲) که متغیر پاسخ پیشرفت تحصیلی و دوازده متغیر جدول (۱) کمکی هستند، در شکل ۲، به ازای مقادیر $\lambda = 0, 0.00001, 0.0001, 0.001$ تعداد طبقات بهینه بر اساس معیار C محاسبه و مشخص شده است. همانطور که از نمودار معلوم است، با افزایش ضریب تنظیم λ، تعداد طبقات بهینه از ۱۲۸ به ۸۴ و ۶ و ۲ کاهش پیدا کرده است.

تا حدی از روی نمودار بالا سمت چپ که در آن λ برابر صفر است، می‌توان تعداد طبقه بهینه را تعیین کرد. منحنی C در نقاط ۶،۱۹ و ۴۵ داری سه کنج است. البته کنج بودن در نقطه ۶ مشهودتر است زیرا شیب منحنی قبل و بعد از آن نقطه تفاوت زیادی دارد. هنگامی که تعداد طبقات از ۲ تا ۶ افزایش می یابد، شیب منحنی C ، که به ازای $\lambda = 0$ همان واریانس برآوردگر است، شدیدا کاهش پیدا می‌کند ولی افزایش تعداد طبقات از ۶ به بعد به آن اندازه کاهش واریانس به وجود نمی‌آورند. در جدول ۲ میزان کاهش واریانس به ازای افزودن یک طبقه از تعداد طبقه ۲ تا ۶، ۶ تا ۱۹، ۱۹ تا ۴۵ و ۴۵ تا ۱۲۸ محاسبه شده است. بیشترین کاهش واریانس تا رسیدن به شش طبقه حاصل می‌شود و از طبقه شش به بالا به ازای افزایش هر طبقه مقدار ناچیزی از واریانس کم می‌شود. تشکیل شش طبقه با استفاده از درخت تصمیم هرس شده در شکل ۳ نمایش داده شده است.

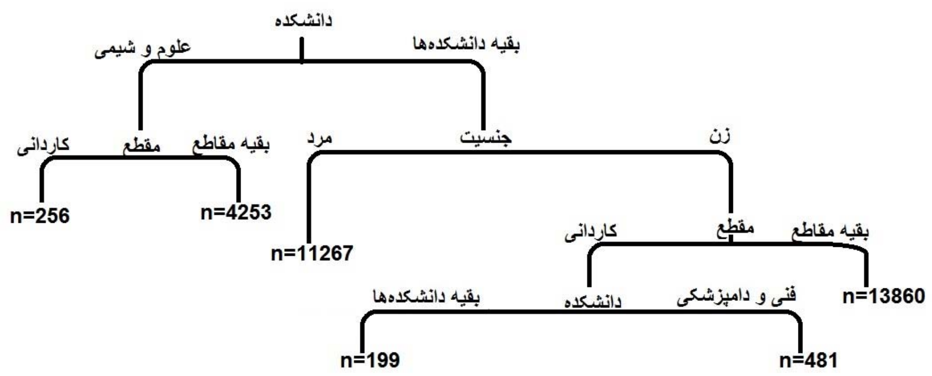
در نمودار پایین سمت چپ نیز به ازای $\lambda = 0.0001$ تعداد طبقه بهینه برابر شش حاصل شده است.

تفاوت کلیدی که این روش با روشهای قبلی دارد، محدود نبودن معیار بهینگی به پارامتر واریانس و دخالت داشتن عامل تعداد طبقات است. همچنین طبقه‌بندی کردن مشاهدات بر اساس حدود متغیر پاسخ در اجرا اغلب میسر نیست. در طبقه‌بندی با درخت تصمیم، هردو نوع متغیر کمکی کمی و کیفی که ممکن است در طبقه‌بندی طبیعی جامعه مورد

توجه باشند در این روش قابل استفاده هستند. به عنوان مثال، هنگام بررسی رضایت مشتریان یک شرکت به ویژه در اولین نظرسنجی، شاید اطلاع چندانی از حدود رضایت در دست نباشد ولی به راحتی مشتریان بر اساس متغیرهای کمکی موجود قابل طبقه‌بندی هستند.

جدول۲ . میزان کاهش واریانس در ازای افزودن یک طبقه در کنج‌های نمودار بالا سمت چپ شکل ۲.

وضعیت	تعداد طبقات	واریانس برآوردگر	کاهش واریانس به نسبت وضعیت قبلی	کاهش واریانس در ازای افزایش هر طبقه
۱	۲	0.048947		
۲	٦	0.04669	-0.00226	-0.00056
۳	۱۹	0.04616	-0.00053	-4.1E-05
٤	٤٥	0.044895	-0.00127	-4.9E-05
٥	۱۲۸	0.044045	-0.00085	-1E-05



شکل ۳: درخت با شش گره پایانی که طبقه بندی بهینه را بر اساس معیار ۴-۱ نتیجه داده است.

۵. نتیجه‌گیری

هنگامی که حدود طبقات بر اساس انشعابات درخت تصمیم ساخته می‌شود مزیت تفسیرپذیری هم اضافه می‌شود. از نتایج بخش چهارم معلوم شد که این هدف با شش طبقه و بر اساس متغیرهای کمکی دانشکده، مقطع و جنسیت قابل دستیابی است. نتایج این طبقه‌بندی بهینه بدین صورت است که طبقه اول شامل ۲۵۶ نفر دانشجویان دانشکده‌های علوم و شیمی است که در مقطع کاردانی یا کاردانی معادل فارغ التحصیل شده‌اند، طبقه دوم شامل ۴۲۵۳ نفر از دانشجویان

دانشکده‌های علوم و شیمی است که در مقاطع دیگر (غیر از کاردانی یا کاردانی معادل) فارغ التحصیل شده‌اند، طبقه سوم ۱۱۲۶۷ نفر از دانشجویان پسر دانشکده‌های دیگر (غیر از شیمی و علوم) هستند. دانشجویان دختر دانشکده‌های دیگر (غیر از شیمی و علوم) بر اساس دو عامل مقطع و دانشکده در سه طبقه بعدی قرار گرفته‌اند. دانشجویان دختر غیر کاردانی به تعداد ۱۳۸۶۰ نفر در یک طبقه و دانشجویان کاردانی دانشکده فنی و دامپروری به تعداد ۴۸۱ نفر در یک طبقه و دانشجویان کاردانی بقیه دانشکده‌ها به تعداد ۱۹۹ نفر در طبقه‌ای دیگر جای گرفته‌اند.

اگر هدف از طبقه‌بندی فقط تفسیرپذیری بود شاید طبقه‌بندی بر اساس دانشکده و جنسیت(یعنی دو برابر تعداد دانشکده‌ها)، بهترین گزینه بود که در آن صورت به درستی قابل قبولی دست پیدا نمی‌کردیم. ولی اگر هدف را فقط درستی بالا در تعیین می‌کردیم باید به تعداد ۱۲۸ طبقه یا بیشتر تشکیل می‌دادیم که مشکلات پیچیدگی و دشواری تفسیرپذیری را با خود داشت.

References

1. B. K. Gupt and Md. I. Ahamed, Optimum stratification for a generalized auxiliary variable proportional allocation under a superpopulation model, Communications in Statistics - Theory and Methods, **51(10)** (2022), 3269–3284.
2. T. Dalenius, The Problem of Optimum Stratification, Scandinavian Actuarial Journal, (**3–4**) (1950), 203–213.
3. F. Danish, R. Jan, M. Daniyal, and K. Tawiah, Optimum Stratification Using Dynamic Programming with a Mixture of Ratio and Product Estimators under Super Population Model, Mathematical *Problems in Engineering* (2023)
4. M. G. M. Khan, V. D. Prasadand, and D. K. Rao, On Optimum Stratification, World Academy of Science, Engineering and Technology International Journal of Mathematical and Computational Sciences. **8(3)** (2014), 508-512
5. K.G. Reddy, M. G. M. Khan and S. Khan, Optimum strata boundaries and sample sizes in health surveys using auxiliary variables. PLoS One. Apr 5;**13(4)** (2018), e0194787.
6. C. E. Särndal, B. Swensson, and J. Wretman, *Model assisted survey sampling,* Springer-Verlag Publishing, (1992).
7. R. Singh, and B.V. Sukhatme, Optimum stratification, Ann Inst Stat Math **21** (1969), 515–528.
8. R. Singh, and B.V. Sukhatme, Optimum stratification with ratio and regression methods of estimation, Ann Inst Stat Math **25** (1973), 627–633.