جمهوری اسلامی ایران
مرکز آمار ایران
پژوهشکده‌ی آمار

براورد چگالی داده‌ها و آماره‌ها

زمستان ۱۳۸۴

براورد چگالی داده‌ها و آماره‌ها

فرشید جمشیدی

زهره فلاح‌محسن‌خانی

مهتاب کشاورز

پژوهشکده‌ی آمار

زمستان ۱۳۸۴

به نام خداوند جان و خرد

پیشگفتار

تابع چگالی احتمال مفهومی اساسی در آمار و احتمال است که با دانستن آن می‌توان به رفتار تصادفی برآوردگرها پی برد. البته تعیین توزیع برآوردگرها به سادگی امکان‌پذیر نیست. خصوصاً در حالتی که توزیع آماره‌ی غیر خطی از نمونه‌ی تصادفی که توزیع آن مشخص نیست مورد نظر باشد.

دانستن توزیع آماره نه تنها امکان انجام استنباط از نمونه‌ی تصادفی به جامعه را فراهم می‌کند، بلکه دقت برآوردها را نیز در اختیار کاربران قرار می‌دهد.

در این راستا پژوهشکده‌ی آمار با تشکیل یک گروه مطالعاتی، طرح پژوهشی «برآورد چگالی داده‌ها و آماره‌ها» را در دستور کار خود قرار داد. این طرح به تفصیل روش‌های موجود برآورد چگالی داده‌ها را معرفی می‌نماید. بعد از بررسی روش‌های برآورد چگالی روی داده‌های شبیه‌سازی شده‌ی منتج از توزیع‌های معلوم، یک طرح نمونه‌گیری پیچیده نیز شبیه‌سازی می‌شود و پس از محاسبه تکرارهای بوت‌استرپ یک برآوردگر مفروض آن طرح، روش‌های برآورد چگالی روی تکرارهای محاسبه شده اعمال و سپس تابع چگالی آن برآوردگر، برآورد می‌شود. در انتها نیز برآورد چگالی نرخ اشتغال از داده‌های فصل چهارم آمارگیری اشتغال و بیکاری سال ۱۳۸۲ ارائه می‌شود.

در گروه مطالعاتی طرح مذکور، آقای فرشید جمشیدی به عنوان مجری طرح و خانم‌ها زهره فلاح‌محسن‌خانی و مهتاب کشاورز به عنوان همکار عضویت داشتند که بدین‌وسیله از زحمات یکایک این افراد تشکر و قدردانی می‌شود.

در طول انجام این طرح، سرکار خانم مهرنوش میرمحمد با دقت و حوصله‌ی بسیار، زحمت حروفچینی تمام مدارک و مستندات طرح را به عهده داشته‌اند که در این‌جا لازم است از تلاش بی‌وقفه ایشان تشکر و قدردانی شود.

اگرچه در انجام پژوهش، داور محترم نظرات اصلاحی خود را اعلام نموده و از این نظرات در جهت بهبود هر چه بیش‌تر طرح استفاده شده است. اما از خوانندگان محترم تقاضا دارد موارد احتمالی اشکالات و ابهامات موجود در نشریه را به گروه پژوهشی طرح‌های فنی و روش‌های آماری پژوهشکده‌ی آمار منعکس نمایند.

گروه پژوهشی طرح‌های فنی و روش‌های آماری

پژوهشکده‌ی آمار

فهرست مطالب

فصل اول

مقدمه

تابع چگالی احتمال مفهومی اساسی در آمار و احتمال است. در واقع به‌وسیله تابع چگالی احتمال است که می‌توان به رفتار تصادفی متغیرهای تصادفی پی برد. خصوصاً در آمار بدون دانستن تابع چگالی احتمال آماره‌های حاصل از یک نمونه تصادفی، امکان انجام استنباط از نمونه‌ی تصادفی به روی جامعه وجود ندارد. در انجام استنباط از نمونه به روی جامعه ۴ حالت ممکن است که پیش آید.

حالت اول: دانستن توزیعی که از آن نمونه‌ی تصادفی $X_1,...,X_n$ گرفته شده است و از روی آن توزیع آماره‌ی $T(\mathbf{X})$ معلوم باشد. در این حالت مشکلی برای انجام استنباط به کمک آماره‌ی $T(\mathbf{X})$ از نمونه به روی جامعه وجود ندارد.

حالت دوم: مجهول بودن توزیعی که از آن نمونه‌ی تصادفی $X_1,...,X_n$ گرفته شده است اما توزیع آماره $T(\mathbf{X})$ به صورت تقریبی بنا به قضیه‌ی حد مرکزی معلوم باشد. این در حالتی است که $T(\mathbf{X})$، تابعی خطی از $\sum_{i=1}^{n} X_i$ باشد. در این حالت نیز در انجام استنباط مشکلی وجود ندارد.

حالت سوم: دانستن توزیعی که از نمونه‌ی تصادفی $X_1,...,X_n$ گرفته شده است اما به‌دست آوردن توزیع آماره‌ی $T(\mathbf{X})$ به دلیل غیر خطی بودن $T(\mathbf{X})$ بر حسب $\sum X_i$ و پیچیده بودن شکل تابعی آن، از روش‌های تحلیلی امکان‌پذیر نبوده یا مشکل باشد. در این حالت به دلیل مجهول بودن توزیع $T(\mathbf{X})$ انجام استنباط با مشکل روبرو است.

حالت چهارم: ندانستن توزیعی که از آن نمونه تصادفی $X_1,...,X_n$ گرفته شده است و به دلیل غیر خطی بودن آماره‌ی $T(\mathbf{X})$ بر حسب $\sum X_i$ و عدم امکان استفاده از قضیه‌ی حد مرکزی، امکان تعیین توزیع $T(\mathbf{X})$ وجود نداشته باشد. در این حالت به دلیلی که بعداً ذکر خواهد شد انجام استنباط با مشکلی بزرگ‌تر از حالت سوم روبرو خواهد بود.

برای حل مشکل حالت سوم، الگوریتم مونت کارلو غالباً مفید خواهد بود. به این ترتیب که از توزیع معلومی که $X_1,...,X_n$ از آن نمونه‌گیری شده است به طور مکرر نمونه‌هایی شبیه‌سازی شده و $T\left(\mathbf{X}_i^*\right)$، $i=1,...,B$، که در آن B تعداد تکرارهای نمونه‌گیری از توزیع معلوم و $\mathbf{X}_i^*$ نمونه‌های شبیه‌سازی شده از توزیع معلوم می‌باشند. قسمتی از این تحقیق مروری بر روش‌های برآورد چگالی داده‌هایی است که توزیع آن‌ها مجهول می‌باشد. با به‌کارگیری این روش‌ها روی $T\left(\mathbf{X}_i^*\right)$ها به چگالی برآورد شده‌ی $T(\mathbf{X})$ دست پیدا می‌کنیم.

برای حل مشکل حالت چهارم، کار کمی پیچیده‌تر است. در واقع، عمدتاً با این مشکل روبرو هستیم. برای حل این مشکل از روش‌های باز نمونه‌گیری و روش‌های برآورد چگالی تواماً استفاده می‌شود به این ترتیب که ابتدا از نمونه تصادفی $X_1,...,X_n$ به وسیله‌ی روش بوت‌استرپ که روش مناسبی از بین روش‌های باز نمونه‌گیری است باز نمونه‌های $\mathbf{X}_i^*$ گرفته شده سپس روش‌های برآورد چگالی روی $T\left(\mathbf{X}_i^*\right)$ ها، $i=1,...,B$، پیاده می‌شود تا نهایتاً برآورد چگالی آماره‌ی $T(\mathbf{X})$ به دست آید.

دانستن توزیع آماره‌ی $T(\mathbf{X})$ علاوه بر آن که امکان انجام استنباط‌های (ناپارامتری) از نمونه‌ی تصادفی $X_1,...,X_n$ به روی جامعه را فراهم می‌آورد بلکه همراه با روش‌های برآورد واریانس که توسط شمس و همکاران (۱۳۸۴) معرفی شده است، معیار مناسبی را از دقت برآوردها در اختیار کاربران قرار می‌دهد.

روش‌های برآورد چگالی عمدتاً به دو گروه اصلی تقسیم می‌شوند: ۱- برآورد پارامتری ۲- برآورد ناپارامتری. در برآورد پارامتری فرض می‌شود که داده‌ها از یک خانواده توزیع احتمال مانند نرمال با پارامترهای مجهول μ و σ^2 هستند. در این حالت هدف، برآورد μ و σ^2 از روی داده‌ها است. در برآورد ناپارامتری خود تابع چگالی f مجهول می‌باشد و در این حالت خود داده‌ها باید برآورد f را تعیین کنند. در این تحقیق روش‌های برآورد ناپارامتری تابع چگالی مدنظر است.

برآورد ناپارامتری تابع چگالی ابتدا توسط فیکس[1] و هاجز[2] (۱۹۵۱) به منظور رهایی روش‌های آنالیز تشخیص[3] از فرضیاتی راجع به توزیع، پیشنهاد شد. از آن پس برآورد چگالی و ایده‌های وابسته به آن در بسیاری از زمینه‌ها به‌کار گرفته شده است.

در فصل دوم این تحقیق روش‌های موجود و عملی در برآورد چگالی شرح داده می‌شوند. در فصل سوم روش بوت‌استرپ به عنوان روشی مناسب برای باز نمونه‌گیری به منظور برآورد چگالی در طرح‌های نمونه‌ای پیچیده توضیح داده می‌شود. فصل چهارم به شبیه‌سازی مجموعه داده‌هایی از توزیع‌های معلوم و به‌کارگیری روش‌های مختلف برآورد چگالی روی آن‌ها و مقایسه چگالی برآورد شده با چگالی واقعی و همچنین شبیه‌سازی یک طرح نمونه‌گیری پیچیده و انجام روش بوت‌استرپ به منظور به‌دست آوردن تکرار از آماره‌ای مفروض در آن طرح و برآورد چگالی آن آماره اختصاص دارد. سرانجام در فصل پنجم روش‌های بحث شده در این تحقیق بر روی داده‌های یک طرح نمونه‌گیری منتخب از طرح‌های مرکز آمار ایران و برآورد چگالی یک برآوردگر که در آن طرح محاسبه شده است، پیاده می‌شود.

---

[1] Fix
[2] Hodges
[3] Discriminate Analysis

فصل دوم

روش‌های موجود در برآورد چگالی