



پژوهشکده آمار

جمهوری اسلامی ایران  
مرکز آمار ایران  
پژوهشکده‌ی آمار

# برآورد چگالی داده‌ها و آماره‌ها

نخستین ۱۳۸۴



# برآورد چگالی داده‌ها و آماره‌ها

فرشید جمشیدی

زهره فلاح محسن خانی

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 بیش تر 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

صفحه	عنوان
۱	۱- مقدمه
۷	۲- روش‌های موجود در برآورد چگالی
۱۱	۱-۲- هیستوگرام
۱۲	۲-۲- برآوردگر ساده
۱۴	۳-۲- برآوردگر کرنل
۱۴	۲-۳-۱- معرفی برآوردگر کرنل
۲۰	۲-۳-۲- معیارهای اختلاف: میانگین مربع خطا و میانگین انتگرال مربع خطا
۲۱	۲-۳-۳- به‌کارگیری محاسبه‌ی $MSE$ و $MISE$ در مورد برآوردگر کرنل
۲۱	۲-۳-۴- خواص مجانبی برای $MSE$ و $MISE$ در مورد برآوردگر کرنل
	۲-۳-۵- توازن بین جملات انتگرال مربع اریبی و انتگرال واریانس در
۲۳	فرمول میانگین انتگرال مربع خطای ایجاد شده توسط پارامتر هموارسازی
۲۴	۲-۳-۶- پهنای پنجره‌ی ایده‌آل
۲۴	۲-۳-۷- روش انتخاب پارامتر هموارسازی
۲۵	۲-۳-۷-۱- انتخاب ذهنی
۲۵	۲-۳-۷-۲- مراجعه به یک توزیع استاندارد
۲۹	۲-۳-۷-۳- کم‌ترین مربعات اعتبارسنجی متقابل

۳۴	۲-۳-۷-۴- درستی‌نمایی اعتبارسنجی متقابل
۳۶	۲-۳-۸- انتخاب هسته
۳۹	۲-۴- برآوردگر کرنل تطبیقی
۳۹	۲-۴-۱- تعریف و خواص کلی
۴۲	۲-۴-۲- انتخاب خودکار پارامتر هموارسازی
۴۳	۲-۴-۳- چند مثال
۴۶	۲-۵-۵- برآوردگر نزدیک‌ترین همسایه
۴۶	۲-۵-۱- تعریف و خواص
۴۹	۲-۵-۲- انتخاب پارامتر هموارسازی
۵۰	۲-۶- برآوردگر کرنل متغیر
۵۳	۳- بوت‌استرپ
۵۵	۳-۱- استفاده از روش‌های برآوردگر چگالی داده‌ها برای برآورد چگالی آماره‌ها
۵۵	۳-۲- بوت‌استرپ چیست؟
۵۶	۳-۳- برآورد بوت‌استرپ انحراف معیار
۶۱	۳-۴- بوت‌استرپ پارامتری
۶۲	۳-۵- بوت‌استرپ در داده‌هایی با ساختار پیچیده
۶۳	۳-۵-۱- حالت یک نمونه‌ای
۶۴	۳-۵-۲- حالت دو نمونه‌ای
۶۷	۳-۵-۳- ساختارهای داده‌ای کلی‌تر
۷۴	۳-۵-۴- بوت‌استرپ بلوک‌های متحرک
۷۶	۳-۶- بوت‌استرپ در آمارگیری‌های پیچیده



۷۷	۳-۶-۱- بوت استرپ در نمونه‌های طبقه‌بندی شده
۷۸	۳-۶-۲- بوت استرپ در نمونه‌گیری خوشه‌ای دو مرحله‌ای
۸۰	۳-۶-۳- بوت استرپ در نمونه‌گیری خوشه‌های دو مرحله‌ای طبقه‌ای
۸۳	<b>۴- شبیه‌سازی</b>
۸۶	۴-۱- توزیع نرمال
۸۶	۴-۱-۱- نزدیک‌ترین همسایه
۹۱	۴-۱-۲- نزدیک‌ترین همسایه تعمیم یافته
۹۵	۴-۱-۳- کرنل معمولی
۱۰۰	۴-۱-۴- کرنل تطبیقی
۱۰۷	۴-۲- توزیع لگ نرمال
۱۰۷	۴-۲-۱- نزدیک‌ترین همسایه
۱۰۹	۴-۲-۲- نزدیک‌ترین همسایه تعمیم یافته
۱۱۱	۴-۱-۳- کرنل معمولی
۱۱۳	۴-۱-۴- کرنل تطبیقی
۱۱۶	۴-۳- توزیع $t$ -استودنت
۱۱۷	۴-۳-۱- نزدیک‌ترین همسایه
۱۱۹	۴-۳-۲- نزدیک‌ترین همسایه تعمیم یافته
۱۲۲	۴-۳-۳- کرنل معمولی
۱۲۵	۴-۳-۴- کرنل تطبیقی
۱۲۸	۴-۴- توزیع نرمال آمیخته
۱۲۹	۴-۴-۱- نزدیک‌ترین همسایه

۱۳۱	۲-۴-۴- نزدیک‌ترین همسایه تعمیم یافته
۱۳۳	۳-۴-۴- کرنل معمولی
۱۳۵	۴-۴-۴- کرنل تطبیقی
۱۳۹	۵-۴- شبیه‌سازی یک طرح نمونه‌گیری طبقه‌ای خوشه‌ای دو مرحله‌ای و بوت‌استرپ
۱۳۹	۱-۵-۴- شبیه‌سازی طرح نمونه‌گیری
۱۴۱	۲-۵-۴- بوت‌استرپ طرح نمونه‌گیری
۱۴۸	۳-۵-۴- برآورد چگالی
۱۴۸	۱-۳-۵-۴- برآورد چگالی میانگین $X$
۱۵۰	۲-۳-۵-۴- برآورد چگالی برآورد نسبتی $\frac{Y}{X}$
۱۵۳	۵- طرح اشتغال و بیکاری
۱۵۶	۱-۵- برآورد چگالی نرخ اشتغال
۱۵۸	۲-۵- فاصله اطمینان نرخ اشتغال
۱۵۹	۳-۵- سخن آخر
۱۵۹	مراجع
۱۶۵	پیوست‌ها

# فصل اول

مقدمه



تابع چگالی احتمال مفهومی اساسی در آمار و احتمال است. در واقع به وسیله تابع چگالی احتمال است که می توان به رفتار تصادفی متغیرهای تصادفی پی برد. خصوصاً در آمار بدون دانستن تابع چگالی احتمال آماره های حاصل از یک نمونه تصادفی، امکان انجام استنباط از نمونه ی 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 نمونه ی تصادفی  $X_1, \dots, X_n$  گرفته شده است و از روی آن توزیع آماره ی  $T(\mathbf{X})$  معلوم باشد. در این حالت مشکلی برای انجام استنباط به کمک آماره ی  $T(\mathbf{X})$  از نمونه به روی جامعه وجود ندارد.

حالت دوم: مجهول بودن توزیعی که از آن نمونه ی تصادفی  $X_1, \dots, X_n$  گرفته شده است اما توزیع آماره  $T(\mathbf{X})$  به صورت تقریبی بنا به قضیه ی 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 نمونه ی تصادفی  $X_1, \dots, X_n$  گرفته شده است اما به دست آوردن توزیع آماره ی  $T(\mathbf{X})$  به دلیل غیر خطی بودن  $T(\mathbf{X})$  بر حسب  $\sum X_i$  و پیچیده بودن شکل تابعی آن، از روش های تحلیلی امکان پذیر 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  $X_1, \dots, X_n$  گرفته شده است و به دلیل غیر خطی بودن آماره ی  $T(\mathbf{X})$  بر حسب  $\sum X_i$  و عدم امکان استفاده از قضیه ی 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 بزرگ تر 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  $X_1, \dots, X_n$  از آن نمونه‌گیری شده است به طور مکرر نمونه‌هایی شبیه‌سازی شده و  $T(\mathbf{X}_i^*)$ ،  $i=1, \dots, B$ ، 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  $T(\mathbf{X}_i^*)$ 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 توأم استفاده می‌شود به این ترتیب که ابتدا از نمونه تصادفی  $X_1, \dots, X_n$  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  $T(\mathbf{X}_i^*)$ ها،  $i=1, \dots, B$ ، 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  $X_1, \dots, X_n$  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  $\mu$  و  $\sigma^2$  هستند. در این حالت هدف، برآورد  $\mu$  و  $\sigma^2$  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<sup>۱</sup> و هاجز<sup>۲</sup> (۱۹۵۱) به منظور رهایی روش‌های آنالیز تشخیص<sup>۳</sup> 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.

---

<sup>1</sup> Fix

<sup>2</sup> Hodges

<sup>3</sup> Discriminate Analysis





## فصل دوم

روش‌های موجود در برآورد چگالی

