دانشگاه علامه طباطبائی

دانشکده‌ی اقتصاد

گروه آمار، ریاضی و کامپیوتر

پایان‌نامه برای دریافت درجه‌ی کارشناسی ارشد **آمار ریاضی**

عنوان

نزدیکی پیتمن آماره‌های ترتیبی و مقدارهای رکوردی به چندک‌های جامعه

پژوهشگر

بهمن تارویردی‌زاده

استاد راهنما

دکتر نادر نعمت‌الهی

استاد مشاور

دکتر فرزاد اسکندری

شهریور ۱۳۹۰

بسم الله الرحمن الرحیم

تأیید پایان‌نامه‌ی کارشناسی ارشدتوسط دانشجو

عنوان پایان‌نامه: نزدیکی پیتمن آماره‌های ترتیبی و مقدارهای رکوردی به چندک‌های جامعه

نام دانشجو: بهمن تارویردی‌زاده

شماره‌ی دانشجویی: ۸۸۱۲۵۱۲۸۱۰۴

استاد راهنما: دکتر نادر نعمت‌الهی

این‌جانب **بهمن تارویردی‌زاده** دانشجوی کارشناسی ارشد رشته‌ی **آمار ریاضی** دانشکده‌ی اقتصاد دانشگاه علامه طباطبائی گواهی می‌نمایم پژوهش‌های ارائه شده در پایان‌نامه با عنوان مذکور توسط شخص این‌جانب انجام شده است و درستی مطالب نگارش یافته مورد تأیید می‌باشد. همچنین گواهی می‌نمایم مطالب مندرج در پایان‌نامه تاکنون برای دریافت هیچ نوع مدرک یا امتیازی توسط این‌جانب یا فرد دیگری در هیچ کجا ارائه نشده است و در نگارش متن پایان‌نامه شیوه‌ی نگارش مصوب دانشکده‌ی اقتصاد را به‌طور کامل رعایت نموده‌ام. چنان‌چه در هر زمان خلاف آن‌چه گواهی نموده‌ام مشاهده گردد خود را از آثار حقیقی و حقوقی ناشی از دریافت مدرک کارشناسی ارشد محروم می‌دانم و هیچ‌گونه ادعایی نخواهم داشت.

تاریخ: **امضا دانشجو:**

اهدائیه

تقدیم به همه ی آن هایی که

حقشان بر گردنم است

سپاس گزاری

سپاس خدای را که هر توفیقی در گرو عنایت اوست. اکنون که با یاری او توانسته‌ام تلاشی هر چند ناچیز را در راه کسب دانش به انجام رسانم، بر خود لازم می‌دانم از استاد راهنمای بزرگوارم، جناب آقای دکتر نادر نعمت‌الهی، که به پایان رساندن این تحقیق جز با راهنمایی‌های پدرانه و هدایت‌های بی‌دریغ ایشان میسر نبود، قدردانی نمایم.

از استاد مشاورم جناب آقای دکتر فرزاد اسکندری که تذکراتشان باعث غنای پایان‌نامه شد، تشکر می‌نمایم. همچنین از جناب آقای دکتر خلیل شفیعی و جناب آقای دکتر حمیدرضا نواب‌پور که زحمت داوری این اثر را به عهده داشتند سپاس گزارم.

در پایان، از خانواده‌ام، به‌ویژه مادرم که با حمایت‌های خویش، همواره مرا پشتیبانی کرده‌اند نهایت سپاس و قدرشناسی را دارم.

امیدوارم بتوانم از عهده ادای حق این عزیزان برآیم.

شهریور ۹۰

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

نمادها و علایم اختصاری

تعریف	نماد
تابع توزیع تجمعی	$F(.)$
تابع بقا	$\overline{F}(.)$
تابع چگالی احتمال	$f(.)$
تابع بتا	$B(.,.)$
تابع گاما	$\Gamma(.)$
میانه‌ی توزیع	$M(.)$
چندک pام	ξ_p
هم‌توزیعی	$\stackrel{d}{=}$

چکیده

در بیش‌تر مطالعه‌های آماری، براورد پارامتر مجهول جامعه به‌کمک یک نمونه‌ی تصادفی از جامعه، هدف اصلی است. از آن‌جا که براوردگرهای زیادی را می‌توان برای یک پارامتر از جامعه ارائه داد، لذا باید معیاری وجود داشته باشد که در میان این براوردگرها مناسب‌ترین آن‌ها را انتخاب کند. یکی از این معیارها، معیار نزدیکی پیتمن است که برای مقایسه‌ی دو براوردگر مورد استفاده قرار می‌گیرد.

هدف اصلی این پایان‌نامه، به‌دست آوردن بهترین براوردگر برای چندک جامعه براساس آماره‌های ترتیبی و مقدارهای رکوردی است و معیار انتخاب این براوردگر، معیار نزدیکی پیتمن است.

بر همین اساس، در فصل دوم به‌دنبال یافتن نزدیک‌ترین آماره‌ی ترتیبی به‌میانه‌ی جامعه خواهیم بود. نشان می‌دهیم میانه‌ی نمونه براوردگر پیتمن نزدیک‌ترین به میانه‌ی جامعه تحت شرایط معینی است. در فصل سوم در حالت کلی براوردگر پیتمن نزدیک‌ترین به چندک خاصی از جامعه را بر اساس آماره‌های ترتیبی جستجو می‌کنیم.

در فصل‌های چهارم تا ششم به‌دنبال یافتن براوردگر پیتمن نزدیک‌ترین به چندک خاصی از جامعه در خانواده‌ی مکان مقیاس از توزیع‌ها، به‌ترتیب براساس مقدار رکوردهای بالایی، رکوردهای جاری و k-رکوردها خواهیم بود.

واژگان کلیدی. *آماره‌های ترتیبی؛ براوردگر پیتمن نزدیک‌تر؛ براوردگر پیتمن نزدیک‌ترین؛ چندک؛ خانواده‌ی مکان مقیاس؛ رکوردهای جاری؛ مقدارهای رکوردی؛ نزدیکی پیتمن؛* k*-رکوردها*

فصل ۱
کلیات

۱.۱ مقدمه

یکی از پارامترهای جامعه که اهمیت زیادی در علوم مختلف دارد، چندک جامعه است. براورد چندک‌ها در زمینه‌های گوناگونی چون اقتصاد، قابلیت اعتماد و آزمایش‌های طول عمر، آزمایش‌های مربوط به مرگ و میر، علوم بیولوژیکی، فرسودگی مواد و اجزای یک ماشین، منابع معدنی، داده‌های رادیو متریک هوایی، عملیات شاتل فضایی و آزمایش‌های بالینی چشم پزشکی کاربرد دارند. لذا این موضوع حائز اهمیت است که بتوانیم یک براوردگر مناسب برای چندک‌ها به‌دست آوریم.

چندک pام جامعه‌ای با تابع توزیع F که با ξ_p نشان داده می‌شود، به‌صورت زیر تعریف می‌شود:

$$\xi_p = \inf\{x : F(x) \geq p\}$$

زمانی‌که برای یک مسئله آماری، چند براوردگر متفاوت در اختیار داریم، یافتن بهترین براوردگر دارای اهمیت بوده و مسلم است که قضاوت ما باید براساس معیارهای معقول باشد. در روش‌های آمار کلاسیک پارامتری،

فرض می‌شود جامعه مورد نظر دارای یک تابع احتمال یا چگالی احتمال $f_\theta(.)$ است که به پارامتر مجهول θ وابسته است و θ را کمیتی ثابت در نظر می‌گیرند. سپس براساس یک نمونه تصادفی $X_1, X_2, \ldots, X_n$ از این جامعه، براوردگر $\delta(\mathbf{X}) = \delta(X_1, \ldots, X_n)$ را چنان تعیین می‌کنند که تا حد ممکن به پارامتر مجهول θ نزدیک باشد. معمولا میزان این نزدیکی را با تابع زیان $L(\theta, \delta(\mathbf{X}))$ اندازه‌گیری می‌کنند که نه تنها به پارامتر مجهول θ بستگی دارد بلکه به نمونه تصادفی $\mathbf{X} = (X_1, \ldots, X_n)$ نیز بستگی دارد. بنابراین $L(\theta, \delta(\mathbf{X}))$ نیز متغیری تصادفی است. امید ریاضی تابع زیان یعنی $R(\theta, \delta) = E(L(\theta, \delta(\mathbf{X}))$ را تابع مخاطره گویند. براوردگرهای مختلف را براساس تابع مخاطره آن‌ها با یکدیگر مقایسه می‌کنند. براوردگری مطلوب است که بتواند تا حد ممکن تابع مخاطره را به ازای تمامی مقادیر ممکن پارامتر θ مینیمم کند.

اما برای یافتن بهترین براوردگر برای چندک‌های جامعه، معیارهای دیگری غیر از معیار مخاطره نیز وجود دارند. از جمله این معیارها، معیار نزدیکی پیتمن (PMC[1]) است که برای مقایسه دو به دوی براوردگرها معرفی شده است. معیار نزدیکی پیتمن دارای سابقه‌ای بیش از نیم قرن است و به سبب نا آشنایی با ویژگی‌های آن کمتر مورد استفاده قرار گرفته است. مسائلی از قبیل عدم اتکا به وجود گشتاورها، بر خلاف میانگین توان دوم خطا (MSE[2])، تعبیر ساده و بدیهی که در این معیار نهفته است و وجود مواردی که در آنها معیارهای دیگر قابل استفاده نیستند، باعث شده‌اند که این معیار اهمیت خاصی پیدا کند. یکی دیگر از ویژگی‌های معیار نزدیکی پیتمن این است که به توزیع توأم براوردگرهای رقیب می‌پردازد در حالی که برخی از معیارها مانند معیار MSE توزیع حاشیه‌ای براوردگرها را ملاک عمل قرار می‌دهد. به‌عنوان مثال رکوردهایی را در نظر بگیرید که دوندگان یک مسابقه دو بر جا می‌گذارند، هر چند داشتن نتیجه‌های حاشیه‌ای برای هر دونده در مسابقه‌های انفرادی مفید است اما بیشتر علاقه‌مندیم بدانیم دوندگان در رقابت‌های شانه به شانه (رقابت‌های توأم)، که انرژی بیشتری را صرف می‌کنند، چه رکوردی از خود به جا می‌گذارند. این براورد محتاج داشتن توزیع توأم زمان‌هایی است که توسط دوندگان به‌دست آمده است که در حالت کلی ممکن است از هم مستقل نباشند(نعمت‌اللهی، ۱۳۷۵). در ادامه تعریف‌هایی در مورد معیار نزدیکی پیتمن ارائه می‌دهیم.

تعریف ۱.۱.۱. (معیار نزدیکی پیتمن) اگر T_1 و T_2 دو براوردگر برای پارامتر θ باشند، آن‌گاه معیار نزدیکی

[1] Pitman's Measure of Closeness

[2] Mean Square Error

پیتمن T_1 نسبت به T_2 به‌ازای هر $\theta \in \Omega$ برابر است با :

$$PMC(T_1, T_2|\theta) = Pr(|T_1 - \theta| < |T_2 - \theta|)$$

□

معیار نزدیکی پیتمن فراوانی این‌که براوردگر T_1 نسبت به براوردگر T_2 به مقدار واقعی پارامتر θ نزدیک‌تر است را تعیین می‌کند.

تعریف ۲.۱.۱. (براوردگر پیتمن نزدیک‌تر) اگر T_1 و T_2 دو براوردگر برای پارامتر θ باشند، T_1 براوردگر پیتمن نزدیک‌تر نسبت به T_2 است اگر برای هر $\theta \in \Omega$

$$PMC(T_1, T_2|\theta) \geq PMC(T_2, T_1|\theta) \tag{۱.۱.۱}$$

با نامساوی اکید برای حداقل یک $\theta \in \Omega$. □

چون

$$PMC(T_1, T_2|\theta) = ۱ - PMC(T_2, T_1|\theta)$$

لذا شرط (۱.۱.۱) برای هر $\theta \in \Omega$ به‌صورت زیر نوشته می‌شود

$$PMC(T_1, T_2|\theta) \geq \frac{۱}{۲}$$

با نامساوی اکید برای حداقل یک $\theta \in \Omega$. □

تعریف ۳.۱.۱. (براوردگر پیتمن نزدیک‌ترین) فرض کنید Λ یک کلاس غیرتهی از براوردگرها برای پارامتر θ باشد. در این‌صورت T^* براوردگر پیتمن نزدیک‌ترین در میان همه براوردگرهای موجود در Λ است هرگاه برای هر $T \in \Lambda$ که $T \neq T^*$ و به‌ازای هر $\theta \in \Omega$ داشته باشیم

$$PMC(T^*, T|\theta) \geq \frac{۱}{۲}$$

با نامساوی اکید برای حداقل یک $\theta \in \Omega$.

در این تحقیق سعی می‌کنیم با استفاده از معیار نزدیکی پیتمن براورد چندک‌های جامعه را براساس آماره‌های ترتیبی و آماره‌های رکوردی در یک خانواده‌ی مکان- مقیاس از توزیع‌ها مورد مطالعه قرار دهیم.

۲.۱ مفاهیم و واژه‌های اساسی

در این قسمت برخی از مفاهیم پایه‌ای به کار رفته در پایان‌نامه معرفی می‌شوند.

۱.۲.۱ آماره‌های ترتیبی

در این بخش مفاهیم اولیه‌ی آماره‌های ترتیبی ارائه می‌شود. برای مشاهده‌ی برهان روابط به دیوید و ناگاراجا (۲۰۰۳) و آرنولد و همکاران (۲۰۰۸) مراجعه شود. فرض کنید $X_1, X_2, \ldots, X_n$، n متغیر تصادفی از جامعه‌ای با تابع توزیع تجمعی بطور مطلق پیوسته‌ی $F(x)$ و تابع چگالی احتمال پیوسته‌ی $f(x)$ باشند. آماره‌های ترتیبی متناظر با این متغیرهای تصادفی عبارتند از X_iهایی که به‌ترتیب صعودی مرتب شده‌اند. کوچک‌ترین X_iها را با $X_{1:n}$ ، مشاهده‌ی کوچکتر بعدی را با نماد $X_{2:n}$ ، ...، و در نهایت بزرگ‌ترین مشاهده را با نماد $X_{n:n}$ نشان می‌دهیم. از این رو داریم $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$.

آماره‌های ترتیبی کاربرد فراوانی در زمینه‌های مختلف علوم دارند. بررسی استحکام مواد، کنترل کیفیت، قابلیت اعتماد، کشف مشاهده‌ی دور افتاده، نمونه‌گیری سانسور شده، براورد وضعیت توان‌مند و ... از جمله مواردی هستند که آماره‌های ترتیبی در آن‌ها نقش اساسی ایفا می‌کند.

تابع چگالی توأم $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ به‌صورت زیر به‌دست می‌آید

$$f_{1,2,\ldots,n:n}(x_1, x_2, \ldots, x_n) = n! \prod_{r=1}^{n} f(x_r), \qquad -\infty < x_1 < x_2 < \ldots < x_n < \infty$$

تابع چگالی حاشیه‌ای i-امین آماره ترتیبی $X_{i:n}$، برای هر $1 \leq i \leq n$ برابر است با :

$$f_{i:n}(x) = \frac{n!}{(i-1)!\,(n-i)!} f(x) F(x)^{i-1} \overline{F}(x)^{n-i}, \quad -\infty < x < \infty \qquad (1.2.1)$$

که در آن $\overline{F}(x) = 1 - F(x)$ تابع بقا نامیده می‌شود.

همچنین تابع چگالی توأم $(X_{i:n}, X_{j:n})$ برای هر $1 \leq i < j \leq n$ به‌صورت زیر به‌دست می‌آید

$$f_{i,j:n}(x,y) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} f(x)f(y) \times F(x)^{i-1}(F(y)-F(x))^{j-i-1}\overline{F}(y)^{n-j}, \quad -\infty < x < y < \infty \qquad (1.2.2)$$

۱.۲.۲ رکوردها و خواص آن‌ها

زمانی که با مشاهده‌های متوالی سر و کار داریم، مشاهده‌هایی که از مقدارهای قبلی خود بیشتر یا کمتر می‌باشند، مورد توجه قرار می‌گیرند که این مقدارها را مقدارهای رکوردی می‌نامیم. برای مثال ممکن است به روزها یا هفته‌هایی که میزان بارندگی بیش از گذشته است علاقه‌مند باشیم و یا ممکن است طراح یک سازه بخواهد بداند که در طول عمر مفید سازه مورد طراحی وی، احتمال وقوع زلزله‌ای بزرگتر از آنچه در گذشته اتفاق افتاده چقدر است. همچنین ممکن است مایل به تعیین کمترین فشاری باشیم که سبب تخریب قطعه‌های چوبی مورد استفاده در تولید محصولی می‌شوند. کافیست اولین قطعه را تحت فشار قرار دهیم تا تخریب شود، این میزان فشار را به‌عنوان X_1 ثبت می‌کنیم، سپس قطعه‌های بعدی را در نظر می‌گیریم و X_m را میزان فشار مورد نیاز برای تخریب قطعه mام نامیده و آن را در صورتی ثبت می‌کنیم که

$$X_m < \min\{X_1, \ldots, X_{m-1}\}.$$

بنابر این در صورتی فشار مورد نیاز برای تخریب قطعه‌ای را ثبت خواهیم کرد که کمترین مقدار در میان مقدارهای ثبت شده قبل از خود را داشته باشد. این مقدارها را مقدارهای رکوردی گویند. واضح است که در این آزمایش قطعه‌های کمتری تخریب شده‌اند و فقط قطعه‌هایی مورد تخریب قرار گرفته‌اند که میزان فشار برای تخریب آن‌ها از قطعه‌های قبلی کمتر بوده است. از طرفی موارد زیادی وجود دارند که داده‌های اولیه در دسترس نیستند و ما فقط به داده‌های رکوردی دسترسی داریم. همچنین در بسیاری موارد مشاهده‌ها فقط در صورتی ثبت می‌شوند که یک رکورد باشند، مانند داده‌های مسابقه‌های ورزشی، هواشناسی، ژئوفیزیک، زلزله‌نگاری و غیره. بنابر این این موضوع حائز اهمیت است که بتوان با استفاده از رکوردها، تحلیل و استنباط آماری درباره‌ی جامعه اصلی انجام داد.

فرض کنید $\{X_i, i \geq 1\}$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع با تابع توزیع تجمعی $F(x)$

و تابع چگالی احتمال پیوسته $f(x)$ باشد. مشاهده X_j مقدار رکورد بالایی نامیده می‌شود هرگاه مقدارش از همه مشاهده‌های قبلی بیشتر باشد، به‌عبارت دیگر X_j یک رکورد بالاست اگر برای هر $j > i$، $X_j > X_i$ باشد. همچنین X_j مقدار رکورد پایینی نامیده می‌شود هر گاه برای هر $j > i$، $X_j < X_i$ باشد. طبق قرارداد، فرض می‌کنیم صفرامین رکورد بالایی و پایینی به‌صورت $U_۰ \equiv L_۰ \equiv X_۱$ و برای هر $n \geq ۱$ مقدار nامین رکورد بالایی و پایینی به‌ترتیب با U_n و L_n نمایش داده شوند.

تابع چگالی احتمال حاشیه‌ای nامین رکورد بالایی U_n، برای هر $n \geq ۰$ به‌صورت زیر است

$$f_{U_n}(x) = \frac{\left(-\log \overline{F}(x)\right)^n}{n!} f(x) \tag{۳.۲.۱}$$

همچنین تابع چگالی احتمال توأم U_i و U_j برای هر $۰ \leq i < j$ به‌صورت زیر به‌دست می‌آید

$$f_{U_i,U_j}(x,y) = \frac{\left(-\log \overline{F}(x)\right)^i}{i!} \frac{\left(-\log \overline{F}(y) + \log \overline{F}(x)\right)^{j-i-۱}}{(j-i-۱)!} \frac{f(x)}{\overline{F}(x)} f(y), \quad x < y \tag{۴.۲.۱}$$

برای مشاهده‌ی مطالب بیشتر و برهان روابط می‌توانید به آرنولد و همکاران (۱۹۹۸) مراجعه کنید.

۳.۱ مرور نوشتگان

معیار نزدیکی پیتمن برای اولین بار در سال (۱۹۳۷) توسط پیتمن معرفی شد که بر اساس کمتر بودن زیان، براوردگر مناسب را در مقایسه با دیگر براوردگرها ارائه می‌کند. بعدها افرادی چون لی (۱۹۸۷)، کیتینگ و گوپتا (۱۹۸۴) و رائو و همکاران (۱۹۸۶) ویژگی‌های معیار نزدیکی پیتمن را مورد مطالعه قرار دادند.

مطالعه‌های انجام شده در مورد آماره‌های ترتیبی از ابتدای آن تا اوایل سال ۱۹۶۰، در کتابی توسط سارحان و گرین‌برگ (۱۹۶۲) منتشر شد که به علت داشتن جدول‌های فراوان در آن حتی تا امروز سودمندی خود را حفظ کرده است. بررسی جامع و استادانه‌ی دیوید (۱۹۸۱) یک نقطه‌ی شروع شگفت انگیز است که در آثار مربوط به گالامبوس (۱۹۷۸)، هارتر (۱۹۷۰)، بارنت و لویس (۱۹۸۴)، کاستیلو (۱۹۸۸)، و بالاکریشنان و کوهن (۱۹۹۱) شرح داده شده‌اند.

توجه به مقادیر رکوردها نیز به‌طور جدی از ۵۰ سال پیش شروع شده است، یعنی زمانی که «چاندلر» اولین بار در سال ۱۹۵۲ آماره‌های رکوردی را فرمول‌بندی کرد. گلیک (۱۹۷۸) به‌طور جامع رکوردها را