
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان نامه ی کارشناسی ارشد رشته ی آمار گرایش آمار ریاضی

توزیع بتا – دوجمله‌ای: خواص و کاربردهای آن

استاد راهنما:

دکتر محمد بهرامی

پژوهشگر:

رامین جاهد حور

بهمن ماه ۱۳۸۹

سپاسگزاری

سپاس خدای را که مرا شوق و توان ادراک آموخت. با سپاس از کلیه اساتیدی که اندیشه امروز من، حاصل زحمات و تلاش‌های آنان است.

مراتب سپاس و تشکر خود را از استاد گرانقدر و ارجمندم جناب آقای دکتر محمد بهرامی که همواره با شکیبایی و درایت فراوان، مرا مستفیض راهنمایی‌های بی‌دریغ و ذی قیمت خود از لحاظ علمی و اخلاقی نمودند، ابراز می‌دارم. همچنین از زحمات بی‌دریغ اساتید داور جناب آقای دکتر هوشنگ طالبی و جناب آقای دکتر علی دولتی کمال تشکر را دارم.

از خانواده عزیزم مخصوصاً برادرم که تا این مرحله از زندگی همیشه حامی و پشتیبان من بوده و مادرم که صبر و ایثار را به من آموخت، نهایت تشکر و قدردانی را دارم. همچنین از دوستان و سایر عزیزان که موجد شوق فراگیری بیشتر در من می‌باشند، خالصانه و متواضعانه تشکر می‌نمایم.

رامین جاهد حور

بهمن ماه ۱۳۸۹

تقدیم به اولین معلمان و فرشتگان زندگیم

مادرم عزیز و فداکارم:

که نیمی از زجودش، ایثار و گذشت کامل است

و نیمی دیگر عشق و محبت

برادران و خواهران مهربانم:

به پاس عاطفه سرشار و گرمای امید بخش وجودشان

و به پاس محبت های بی دریغ شان که هرگز فروکش نمی کند.

و تقدیم به سرزمین مادریم

آذربایجان

چکیده

در نظریه آمار و احتمال، توزیع بتا–دوجمله‌ای خانواده‌ای از توزیع‌های احتمالی گسسته است که در بسیاری از تحقیقات کاربردی در زمینه‌هایی مانند علوم زیست شناسی، بهداشت، فیزیکی و اجتماعی مورد توجه محققان قرار گرفته است. انعطاف پذیری توزیع بتا–دوجمله‌ای باعث شده که این توزیع در برازش برخی مجموعه داده‌هایی که بیش پراکنش دارند سهم به سزایی ایفا کند، به طوری که در آمار بیزی، روش‌های بیز تجربی و آمار کلاسیک به عنوان توزیع دوجمله‌ای بیش پراکنده به صورت فراوان مورد استفاده قرار می‌گیرد. مطالعات فراوانی در مورد این توزیع انجام شده و روش‌های مختلفی برای برآورد کردن پارامترهای آن مورد بررسی قرار گرفته است. از آنجا که استفاده از روش‌های مناسب در برآورد پارامترهای این مدل دارای اهمیت ویژه‌ای می‌باشد، در این رساله، روش‌های مختلف برآوردیابی پارامترهای این مدل را مورد مطالعه قرار می‌دهیم. از جمله برآوردهای ماکسیمم درستنمایی را که به روش‌های عددی مانند روش نیوتن–رافسون به دست آمده است، از نظر کارایی مجانبی نسبی با سایر روش‌های دیگر مقایسه می‌کنیم. در بخش دیگری از رساله حاضر تحلیل بیزی مدل بتا–دوجمله‌ای را مورد بررسی قرار می‌دهیم و به مطالعه آزمون‌های مختلف آماری برای این توزیع می‌پردازیم. با توجه به نتایج آزمون‌ها، می‌بینیم که وقتی در آزمون‌های تفاوت حساسیت یا ترجیح بیش از یک منبع تغییرات وجود دارد، مدل دوجمله‌ای داده‌های دوجمله‌ای بیش پراکنده را نمی‌تواند به خوبی برازش کند، ولی مدل بتا–دوجمله‌ای یک مدل جایگزین مناسب برای این آزمون‌ها می‌باشد. در ادامه، تقریب توزیع بتا–دوجمله‌ای را توسط توزیع دوجمله‌ای و پواسن ارائه می‌دهیم.

در پایان به شرح توزیع بتا–دوجمله‌ای تعمیم یافته که به وسیله‌ی تابع فوق هندسی گاوسی تولید شده است می‌پردازیم که به صورت توزیع آمیخته‌ای از بتا تعمیم یافته و توزیع دوجمله‌ای بیان شده است. این توزیع آمیخته جدید برای برازش برخی مجموعه داده‌ها مورد استفاده قرار می‌گیرد. با ارائه دو مثال نشان می‌دهیم که این توزیع می‌تواند برازش داده‌ها بوسیله توزیع بتا–دوجمله‌ای به دست آمده است را بهبود بخشد.

واژه‌های کلیدی: بوت استرپ، بیش پراکنش، تقریب پواسن، تقریب دوجمله‌ای، توان، توزیع بتا–دوجمله‌ای، توزیع دوجمله‌ای، شبیه سازی، کارایی مجانبی نسبی، ماکسیمم درستنمایی.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

فصل سوم : آزمون‌های مختلف آماری برای توزیع بتا-دوجمله‌ای

عنوان **صفحه**

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

عنوان **صفحه**

مخفف‌ها و نمادها

AIC…… AkaikInformaition Criterion
ARE……Asymptotic Relative Efficiency
AW……..Average Width of the Interval
BBD……Beta-Binomial Distribution
CV……..Coefficient of Variation
d_{TV} ……Total Variation Distance
EDF……Empirical-Distribution-Function
GBBD……Generalization of the Beta-Binomial Distribution
GHF……Gaussian Hypergeometric Functions
GOF……Goodness-of-Fit
INCL……....Proportion of Intervals Including the True Value
K……..Kurtosis
LR……...Likelihood Ratio

MAD……Mean Absolute Deviations
MME……Method ofMomentEstimate
MLE……...Maximum Likelihood Estimate

MSE……Mean Square Error
NBD……Negative Binomial Distribution
PMF……ProbabilityMass Function
RMSE……..Root of Mean Square Errors
SD……Standard Deviation
SK……...Skewnes

فصل اول

تعاریف و مفاهیم پایه

۱ – ۱ مقدمه

در این فصل به معرفی مفاهیم و تعاریفی می‌پردازیم که در فصل‌های بعد آنها را مورد استفاده قرار خواهیم داد. در بخش دوم، ابتدا به تاریخچه تحقیق و کارهایی که در این زمینه صورت گرفته می پردازیم. در بخش سوم روش-های برآوردیابی گشتاوری و ماکسیمم درستنمایی را بیان می‌کنیم. سپس در بخش چهارم به معرفی برخی توزیع-های مهم پرداخته و بعضی از ویژگی‌ها و مشخصه‌های آنها را شرح می‌دهیم. روش‌های بوت استرپ و برآورد جانشینی و شبیه سازی مونت کارلو رادر بخش پنجم معرفی کرده و در نهایت مفاهیم پایه‌ای در آمار بیز را شرح خواهیم داد.

۱– ۲ تاریخچه تحقیق

در نظریه احتمال و آمار توزیع بتا – دوجمله‌ای خانواده‌ای از توزیع‌های احتمالی گسسته آمیخته است که در آمار بیزی، روشهای بیز تجربی و آمار کلاسیک به عنوان توزیع دوجمله‌ای بیش پراکنده[1] به طور گسترده‌ای مورد استفاده قرار می‌گیرد. این توزیع کاربردهای فراوانی در علوم اجتماعی، فیزیک و بهداشت دارد. مدل بتا-

[1] Overdispersed Binomial Distribution

دوجمله‌ای اولین بار به صورت رسمی توسط اسکلام[1](۱۹۴۸) ارائه شد، با این وجود این ایده قبلاً توسط پیرسن (۱۹۲۵)در یک آزمایش تحقیقی تئوری بیز پیشنهاد شده بود.

تنوع و انعطاف‌پذیری مدل بتا – دو جمله‌ای باعث شد که از آن در آزمون هوش (هوین[2]، ۱۹۷۹؛ لرد[3]،۱۹۶۵ ؛ ویلکاکس[4]،۱۹۸۱) ، آزمایش تاکسیکولوژی (ویلیامز[5]، ۱۹۷۵)، اپیدمیولوژی (گریفتس[6]، ۱۹۷۳)، نمایش رسانه‌ها (گرین[7]، ۱۹۷۰) و رفتار خریداران (مسی[8] و همکاران، ۱۹۷۰) استفاده شود.

ویلیامز (۱۹۷۵) داده‌های مربوط به تأثیر دارویی خاص بر روی حیوانات آزمایشگاهی را مورد بررسی قرار داد، که در این آزمایش متغیرهای پاسخ X_i تعداد توله‌هایی است که در دوره‌ی شیرخوارگی زنده مانده اند. بنابراین در این آزمایش فرض می‌کنیم که X_i متغیر تصادفی دو جمله‌ای با پارامتر P است. اسکورنیک[9] (۱۹۹۰) در طراحی یک آزمایش روی بیماران غیر بارور، جمعیتی از زنان را که به صورت منظم تخمک‌گذاری نمی‌کردند، مورد بررسی قرار داد. برای هر زن تعداد دوره‌های تخمک‌گذاری X_i از شش دوره‌ی متوالی ثبت شد که استاندارد کلینیکی درمان می‌باشد. بنابراین هر X_i یک متغیر تصادفی دوجمله‌ای با پارامتر P است. در مسئله حیوانات آزمایشگاهی میزان بقا توله‌ها در یک زایمان، از یک زایمان به دیگری تغییر می‌کند. مشابهاً در مطالعه میزان تخمک‌گذاری در پاسخ به درمان می‌توان انتظار داشت که از یک زن به زن دیگر تغییر کند. بنابراین توزیع تعداد چرخه‌های موفق درون گروه تیمار در مقایسه با مدل دوجمله‌ای بیش پراکندگی دارد، که میزان تخمک‌گذاری در هر گروه تخمین زده شده است.

در چنین مواردی، توزیع بتا – دوجمله‌ای مدل انعطاف‌پذیری برای تغییر پذیری بین نمونه فراهم می‌کند. در این مورد، فرض می‌شود که احتمال تخمک‌گذاری هر زن در یک چرخه تنها به صورت بتا باشد و توزیع تعداد چرخه‌های موفق، به شرط آن احتمال دوجمله‌ای باشد.

در مطالعه اثر زایمان روی مدل بندی دز پاسخ در تراتولوژی، کوپر[10] و همکارانش (۱۹۸۶) نتیجه گرفتند که عدم بکارگیری تأثیر زایمان می‌تواند باعث کم برآوردی واریانس مرتبط با برآوردهای پارامتر باشد، به طوری که

---

[1] Skellam
[2] Huynth
[3] Lord
[4] Wilcox
[5] Williams
[6] Griffiths
[7] Green
[8] Massy
[9] Skurnick
[10] Kupper

استفاده از درستنمایی دو جمله‌ای برای مدل‌بندی داده‌های تراتولوژی به نظر معقول نیست. آنها توزیع بتا-دوجمله‌ای را برای معرفی درجه تغییر همبستگی درون زایمان پیشنهاد کردند.

پُل[1](۱۹۸۲) در تحلیل نسبت‌های جنین تأثیر پذیر در آزمایش‌های تراتولوژیکی مشاهده کرد که مدل بتا–دوجمله‌ای نسبت به مدل دوجمله‌ای بهتر عملکرده است. پک[2](۱۹۸۶) نتیجه گرفت که بتا – دوجمله‌ای نسبت به مدل‌های معادل مانند دوجمله‌ای همبسته کوپر و هاسمن[3](۱۹۷۸) بهتر است. تارون[4] (۱۹۸۲) برای بسیاری از انواع انواع غده‌ها مشاهده کرد که نرخ کنترل عمر غده‌ها بسیار تغییر پذیرتر از آن هستند که فرض کنیم دارای توزیع دوجمله‌ای باشند، و توزیع بتا–دوجمله‌ای را برای نرخ عمر غده‌ها برازش داد. شوکرز[5] (۲۰۰۳) از توزیع بتا-دوجمله‌ای در ارزیابی کارایی دستگاه تشخیص هویت بیومتریک استفاده کرد. لین و چوو[6] (۲۰۰۷) آزمون های تابع توزیع تجمعی را برای مدل بتا-دوجمله‌ای ارائه دادند.

تکنیک‌های مختلفی برای برآورد پارامترهای توزیع بتا – دوجمله‌ای ارائه شده است. اسکلام (۱۹۴۸) از برآوردهای ماکسیمم درستنمایی (MLE) که با کمک جداول تابع دایگاما[7] به‌دست می‌آیند، استفاده کرد. گریفیس (۱۹۷۳) روشی برای به‌دست آوردن MLE توسعه داد که نیازی به استفاده از مقادیر تابع دایگاما ندارد، اما شامل فرایند تکراری برای حل معادلات غیر خطی است. نیسن-میر[8](۱۹۶۴) یک روش گرافیکی تکراری برای برای به‌دست آوردن MLE پیشنهاد کرد. ویلیامز (۱۹۸۲) روش ساده‌ای برای تغییر تحلیل لجستیک خطی استاندارد بیان کرد، که از برنامه کامپیوتری GLIM برای در نظر گرفتن تغییر پذیری دوجمله‌ای اضافی استفاده ویژه‌ای می‌کند. ویلیامز (۱۹۸۸) اریبی مرتبط با MLE را مطالعه کرد. بروکس[9] (۱۹۸۴) روش تغییر یافته‌ای از GLIM پیشنهاد کرد که روش ساده‌ای است برای به‌دست آوردن تقریب آماری آزمون نسبت درستنمایی که توزیع بتا–دوجمله‌ای دارد. علاوه بر روش‌های MLE، چتفیلد و گودهارت[10] (۱۹۷۰) روش تکراری ساده برای محاسبه برآوردها بکار بردند که بر پایه معادله‌بندی میانگین نمونه و نسبت صفرهااست. شنتون[11] (۱۹۵۰) نشان داد

[1] Paul
[2] Pack
[3] Kupper and Haseman
[4] Tarone
[5] Schuckers
[6] Lin and Chou
[7] Digamma Function
[8] Nissen-Meyer
[9] Brooks
[10] Chatfield and Goodhart
[11] Shenton

که کارایی مجانبی نسبی روش گشتاورها معمولاً از ۷۰٪ بیشتر می‌باشد. همچنین انسکومب[1] (۱۹۵۰) نشان داد که برای برازش توزیع دوجمله‌ای منفی (NB) که شکل معکوسJ (فراوانی بزرگ صفرها)دارند روش میانگین و صفرها کاراتر است.

چون توزیع NB مورد منحصری از توزیع بتا – دوجمله‌ای است، چتفیلد و گودهارت (۱۹۷۰) روش میانگین و صفرها را برای برازش توزیع بتا–دوجمله‌ای که شکل معکوسJ دارند، پیشنهاد کرد. ویکاکس (۱۹۷۹) با استفاده از تکنیک‌های مونت کارلو، تقریب نیوتن رافسون برآوردماکسیمم درستنمایی توزیع بتا–دوجمله‌ای را با چندین روش ممکن دیگر مقایسه کرد.

۱ – ۳ روش‌های برآوردیابی

۱ – ۳ – ۱ روش گشتاوری

یکی از قدیمی‌ترین روش‌های برآوردیابی، روش برآورد گشتاوری است که در سال ۱۸۹۴ توسط آماردان مشهور کارل پیرسن معرفی شده است. روش برآورد گشتاوری که از آن با عنوان برآورد MM یاد می‌کنیم، عبارت است از دستورالعملی برای به دست آوردن برآوردگری به نام "برآورد گشتاوری" که از آن به اختصار MME یاد می‌کنیم. برآورد MM مبتنی بر بکارگیری گشتاورهای جمعیتی (توزیع) و گشتاورهای نمونه ای است.

برای روشن شدن موضوع، فرض کنید $X_n,...,X_1$ یک نمونه‌ی تصادفی n تایی از توزیع $F_\theta \in \mathcal{F}$ باشد، به طوری که $\theta=(\theta_1,...,\theta_k)\in\Theta\subseteq R^k$. همچنین فرض کنید k گشتاور اول این توزیع، که به صورت توابعی از θ هستند، وجود داشته باشند. می‌دانیم که گشتاور r–ام توزیع، در صورت وجود، به صورت زیر تعریف می‌شود:

$$\mu_r = \mu_r(\theta) = E_\theta(X_1^r) \;\; ; \;\; r = 1,...,k .$$

اگر

$$M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r \;\; , \; r = 1,...,k$$

نمایانگر r–امین گشتاور نمونه‌ای بر پایه نمونه تصادفی داده شده باشد، آنگاه برآورد MM پارامترهای مجهول $\theta_1,...,\theta_k$ براساس یک ایده ساده و از تشکیل و حل k معادله زیر حاصل خواهد شد

$$\mu_r = M_r \;\; , \; r = 1,...,k . \qquad (۱-۱)$$

[1] Anscombe

معمولاً برآورد گشتاوری θ را با $\tilde{\theta}$ نمایش می‌دهند، این برآورد لزومی ندارد که منحصر به فرد باشد. بنابراین با توجه به آنچه که گفته شد، اگر $\theta=\varphi(\mu_1,...,\mu_r)$ باشد، آنگاه $\tilde{\theta}=\varphi(M_1,...,M_r)$ خواهد بود.

۱ – ۳ – ۲ روش ماکسیمم درستنمایی

این روش یکی از قدیمی‌ترین و پراهمیت‌ترین روش‌ها در نظریه برآوردهاست، که از آن به اختصار با عنوان برآورد ML یاد می‌کنیم. این روش اولین بار توسط گوس (۱۸۲۱) بکار گرفته شد و پس از آن به صورت گسترده‌تری در سال ۱۹۲۵ توسط فیشر در انتقاد از روش "برآورد گشتاوری" مورد استفاده قرار گرفت.

روش ML عبارت است از دستورالعملی برای به دست آوردن برآوردگری به نام "برآوردگر ماکسیمم درستنمایی" که از آن به اختصار MLE یاد خواهیم کرد و مبتنی بر یک تابع آماری مهم به نام "تابع درستنمایی" است. فرض کنید $X=(X_1,...,X_n)$ ، بردار n متغیر تصادفی با تابع چگالی احتمال توأم $f_\theta(x)$ ، $\theta\in\Theta\subseteq R^k$ ، باشد.

تعریف ۱–۱ تابع درستنمایی برای هر مقدار داده شده $X=x$ ، تابع درستنمایی X را تابع چگالی احتمال توأم X، یعنی $f_\theta(x)$، تعریف می‌کنیم که به صورت تابعی از θ در نظر گرفته می‌شود و آن را با نماد $L(\theta)$ نمایش می‌دهیم:

$$L(\theta)=f_\theta(x)=L(\theta;x).$$

ذکر چند نکته در ادامه ضروری به نظر می‌رسد.

نکته ۱ – ۱ تابع درستنمایی $L(\theta)$ لزوماً نسبت به θ مشتق پذیر نیست.

نکته ۱ –۲تابع درستنمایی $L(\theta)$ یک تابع چگالی احتمال نیست.

نکته ۱ – ۳اگر $X_1,...,X_n$ یک نمونه تصادفی n تایی از خانواده چگالی‌های $\{f_\theta(x):\theta\in\Theta\}$ باشد، آنگاه:

$$L(\theta)=\prod_{i=1}^{n} f_\theta(x_i)$$

تعریف ۱– ۲برآوردگر ماکسیمم درستنماییاگر $\delta(x)$برآوردگری برای θ باشد به طوری که:

$$i)\quad P_\theta(\delta(X)\in\Theta)=1 \quad ; \quad \forall\,\theta\in\Theta$$
$$ii)\quad L(\delta(x))\geq L(\theta) \quad ; \quad \forall\,\theta\in\Theta$$

آنگاه $\delta(x)$ به عنوان برآوردگر ماکسیمم درستنمایی θ تعریف می‌شود.

معمولاً برآوردماکسیمم درستنمایی θ را با $\hat{\theta}$ نشان می‌دهند و لزومی ندارد که منحصر به فرد باشد.همچنین براساس تعریف $\hat{\theta}$ ،داریم:

$$L\left(\hat{\theta}\right)=\sup_{\theta\in\Theta}L\left(\theta\right)$$

به‌طور معادل اگر تعریف کنیم: $l\left(\theta\right)=\ln L\left(\theta\right)$ ، آنگاه

$$l\left(\hat{\theta}\right)=\sup_{\theta\in\Theta}l\left(\theta\right)$$

۱ - ۴ توزیع‌های کاربردی و توابع فوق هندسی گاوسی

در این بخش به مرورتوزیع‌های بتا ،دوجمله‌ای، و توابع فوق هندسی گاوسی[1] می‌پردازیم و برخی از ویژگی‌های آنها را ارائه می کنیم.

۱ -۴-۱ توزیع بتا

اگر متغیر تصادفی X دارای تابع چگالی زیر باشد

$$f_{\alpha,\beta}(x)=\frac{1}{B(\alpha,\beta)}\,x^{\alpha-1}\,(1-x)^{\beta-1}\;;\;\;0<x<1\;\;,\;\;\alpha,\beta>0 \qquad (۱-۲)$$

آنگاه X دارای توزیع بتا با پارامترهای α و β است و آن را با نماد $X\sim Beta(\alpha,\beta)$ نشان می‌دهیم، که در (۱-۲) داریم، $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

در این توزیع $\mathrm{E(X)}=\frac{\alpha}{\alpha+\beta}$ و $Var(X)=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$است.

۱ -۴-۲ توزیع دوجمله‌ای

هرگاه X تعداد موفقیت‌ها در n بار تکرار مستقل یک آزمایش بر نولی با احتمال موفقیت P در هر آزمایش باشد $(n>0\;,\;0<p<1)$ آنگاه X را یک متغیر تصادفی دو جمله‌ای گوییم و آن را با نماد $\mathrm{Bin}\,(n,p)$ نشان می‌دهیم. تابع چگالی احتمال آن به این صورت است

$$f_X(x)=\binom{n}{x}p^x(1-p)^{n-x}\qquad,\quad x=0,1,\ldots,n$$

[1] Gaussian Hypergeometric Functions