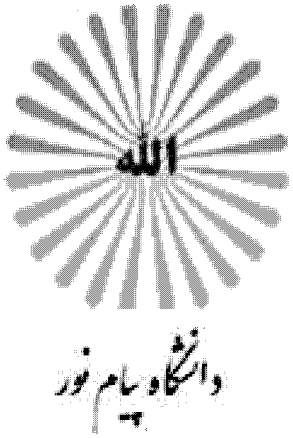


دانشکده علوم پایه

پایان نامه کارشناسی ارشد آمار ریاضی

مدل‌بندی آماری با استفاده از الگوی نرمال چند متغیره آمیخته

توسط:

امیر انشکانی

استاد راهنما:

دکتر پرویز نصیری

استاد مشاور:

دکتر علی شادرخ

تابستان ۱۳۹۱

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مرکز تهران شرق

شماره: .....
تاریخ: .....
پیوست: .....

دانشگاه پیام نور
دانشگاه پیام نور استان تهران
اللهم عجل لولیک الفرج و العافیه و النصر

صورتجلسه دفاع از پایان نامه کارشناسی ارشد

جلسه دفاع ازپایان نامه کارشناسی ارشد آقای امیر انشکانی

دانشجوی رشته آمار ریاضی به شماره دانشجویی: ۸۷۰۰۰۰۲۶۰

تحت عنوان:

مدل بندی آماری با استفاده از الگوی نرمال چند متغیره آمیخته

جلسه دفاع با حضور داوران نامبرده ذیل در روزسه شنبه مورخ: ۹۱/۰۴/۲۷ ساعت : ۱۲-۱۱درمحل تهران شرق برگزار شد. و پس از بررسی پایان‌نامه مذکور با نمره به عدد.....۱۸٫۹..... به حروف .....هیجده و نه دهم..... و با درجه ارزشیابی.....عالی.....مورد قبول واقع شد ☒ نشد ☐

ردیف	نام و نام خانوادگی	هیات داوران	مرتبه دانشگاهي	دانشگاه/ موسسه	امضاء
۱	دکترپرویز نصیری	استاد راهنما	دانشیار	پیام نور	
۲	دکترعلي شادرخ	استاد مشاور	استادیار	پیام نور	
۳	دکتراحسان جمالي	استاد داور	استادیار	سازمان سنجش آموزش کشور	
۴	دکترمسعود یارمحمدي	نماینده علمي گروه/ نماینده تحصیلات تکمیلي	دانشیار	پیام نور	

تهران ، خیابان کریمخان زند ،خیابان استاد نجات الهی ، خیابان شهید فلاح پور ، پلاک ۲۷ مرکز تهران شرق
تلفن : ۸۸۹۱۳۴۷۵
دورنگار: ۸۸۹۴۸۹۸۴
Tshargh.Tpnu.ac.ir
Tshargh@Tpnu.ac.ir

اگر شایسته تقدیم باشد

به روح بلند پرواز پدرم

که مشوق من بوده، هست و خواهد بود.

قدردانی

به نام یگانه پروردگار بزرگ

شکر ایزد منان را که به من توانایی تفکر و عمل کردن آموخت. برخود لازم می‌دانم که پس از پایان دوره کارشناسی ارشد از تمامی کسانی که در طول این دوره در کنار من بوده‌اند، به خصوص اعضای خانواده‌ام، تقدیر و تشکر نموده و درود و صلوات نثار روح پدر مرحومم کنم.

همچنین از آقای دکتر پرویز نصیری که زحمت راهنمایی این پایان نامه و جناب آقای دکتر علی شادرخ که مشاوره بنده را قبول نمودند، به خاطر تمامی زحماتشان تشکر می‌کنم.

امید که این تحقیق سهمی اندک در اعتلای علمی کشور داشته باشد.

امیر انشکانی

تهران

تابستان ۱۳۹۱

چکیده

در بسیاری از مسائل کاربردی تشخیص مدل مناسب برای توزیع صفت جامعه مورد بررسی از اهمیت ویژه‌ای برخوردار است. کاربرد و استفاده از توزیع‌های آمیخته به عنوان یکی از روش‌های داده‌کاوی سابقه بسیار طولانی در اغلب زمینه‌های علمی از جمله پزشکی، کشاورزی، هواشناسی، بازاریابی، مدیریت و غیره دارد. مزیت استفاده از اینگونه توزیع‌ها زمانی که داده‌ها دارای پراکندگی زیاد و حتی گمشده باشند، بدلیل انعطاف پذیری فوق‌العاده آنهابه خوبی معلوم است. این نوع مسائل آماری اولین بار مورد توجه فلر قرار گرفت و سپس در مقالات متعددی روی این موضوع بحث و بررسی شد. توزیع‌های آمیخته را می توان برپایه بسیاری از توزیع‌های شناخته شده مثل یکنواخت، نمایی، نرمال، پواسن، دوجمله‌ای و غیره مدل بندی کرد. هدف از هر مدل‌بندی آماری برآورد پارامترهاست، برای برآورد پارامترها روشهای گوناگونی از جمله روش گشتاورها، حداقل مربعات خطا، روش ماکسیمم درستنمایی، روش مونت کارلو، الگوریتم EM و الگوریتم $MCMC$ وجود دارد.

واژه‌های کلیدی : نرمال چند متغیره آمیخته، شناسایی پذیری، هم واریانس، ناهم واریانس، تابع درستنمایی

KeyWords: Mixture of multivariate normal, Identifiability, Homoscedastic, Heteroscedasticity, Likelihood function

فهرست مندرجات

لیست اشکال

لیست اشکال

لیست جداول

لیست جداول

مقدمه

الگوی نرمال بدلیل برخورداری از برخی ویژگی‌های ذاتی و سهولتی که معمولاً در استنباط فراهم می‌آورد، از جایگاه ویژه‌ای در استنباط آماری برخوردار است. در سال‌های اخیر تلاش‌های گسترده‌ای برای توسعه الگوهایی جامعتر از نرمال که علاوه بر ویژگی‌های مطلوب الگوی نرمال قابلیت بیشتری در مدلبندی داده‌های متنوع داشته باشند، صورت پذیرفته است. اما با وجود موفقیت آمیز بودن بخشی از این تلاش‌ها الگوی نرمال هنوز هم بدلیل پشتوانه استنباطی قابل توجهی که در ادبیات آماری دارد، دارای جایگاه ممتازی است. از طرف دیگر موارد بسیار زیادی وجود دارد که مشاهدات متعلق به چند الگوی یکسان اما با پارامترهای متفاوت هستند، چنین الگوهایی را آمیخته گویند.

تحقیق و بررسی بر روی مدل‌های خاص آمیخته چند سالی است که در بین آماردانان رواج پیدا کرده است، از این رو پیدا کردن رابطه اینگونه توزیع‌ها با سایر موضوعات علمی از اهمیت ویژه‌ای برخوردار است. این در حالی است که اولین بار کارل پیرسن در سال ۱۸۹٤ اینگونه توزیع‌ها را معرفی نمود. بدنبال وی فنگ و مک‌کلاچ در سال ۱۹۹۲ شناسایی‌پذیری مدل‌های آمیخته را مطرح کردند. مک لاکلان و کریشنان (۱۹۹٤) مطالعات خود را بر الگوریتم EM

متمرکز نمودند و مگنوس و بولدا (۲۰۰۸) با ادغام الگوریتم EM و روش بوت استرپ، روشی را ابداع کردند که در نهایت منجر به افزایش سرعت همگرایی در داده‌ها و کاهش زمان و هزینه انجام محاسبات شد.

هدف اصلی در این پایان نامه بررسی مدل‌های چند متغیره نرمال آمیخته می‌باشد. در این تحقیق و در فصل اول مدل‌های نرمال یک متغیره و چند متغیره مورد مطالعه قرار گرفته‌اند. فصل دوم مربوط به مدل‌های آمیخته و مسائل مربوط به آن است. در فصل سوم مدل‌های نرمال و نرمال چند متغیره آمیخته به همراه ویژگی‌های آن‌ها و در فصل چهارم آزمون‌های شناخته شده روی مدل‌های آمیخته بررسی شده است. در فصل پنجم و پایانی نیز مثالی عددی و شبیه سازی اینگونه توزیع‌ها ارائه شده است.

فصل ۱

توزیع نرمال و تعمیم چند متغیره آن

۱.۱ مقدمه

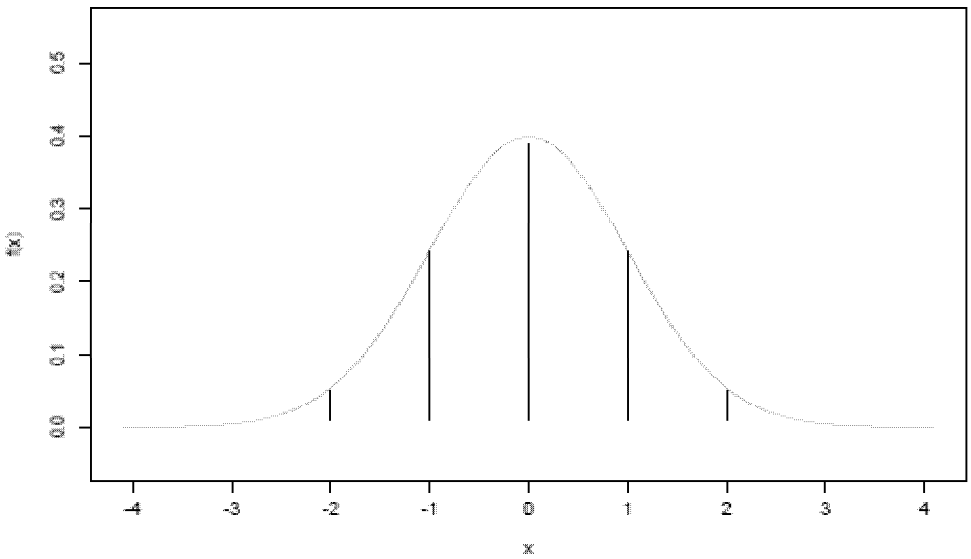
بدلیل اینکه در این پایان نامه از توزیع نرمال و ویژگی‌های آن زیاد استفاده می‌نماییم در این فصل به اجمال برخی از ویژگی‌های این توزیع مهم گردآوری شده است. در بخش دوم این فصل توزیع نرمال یک متغیره و در بخش سوم توزیع نرمال چند متغیره معرفی می‌گردد.

۲.۱ توزیع نرمال یک متغیره

تعریف ۱.۲.۱ متغیر تصادفی $-\infty < X < \infty$ دارای توزیع نرمال با میانگین $-\infty < \mu < \infty$ و واریانس $\sigma^۲ > ۰$ است ($X \sim N(\mu, \sigma^۲)$)، اگر تابع چگالی آن بصورت زیر تعریف شود.

$$f(x) = \frac{۱}{\sqrt{۲\pi}\sigma} \exp[-\frac{۱}{۲}(\frac{x-\mu}{\sigma})^۲], \qquad (۱.۲.۱)$$

همچنین نمودار تابع چگالی نرمال در شکل (۱.۵.۵) نشان داده شده است. برخی از



شکل ۱.۲.۱: تابع چگالی توزیع نرمال $X \sim N(\mu, \sigma^۲)$

ویژگی‌های این توزیع در زیر فهرست شده است.

۱. برای یک فاصله غیر تهی (a, b)، $f(x) \geq ۰$ و

$$P(X \in (a, b)) = area_{(a,b)}.$$

که در آن $X \in (-\infty, +\infty)$.

۲. تابع چگالی f حول μ متقارن است، به عبارت دیگر $f(\mu + x) = f(\mu - x)$.

۳.

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= ۰/۶۸ \\ P(\mu - ۲\sigma \leq X \leq \mu + ۲\sigma) &= ۰/۹۵ \\ P(\mu - ۳\sigma \leq X \leq \mu + ۳\sigma) &= ۰/۹۹. \end{aligned}$$

۴. احتمال کوچکتر یا مساوی بودن متغیر تصادفی X با b یا سطح زیر منحنی نرمال از رابطه زیر بدست می آید.

$$F_X(b) = P(X \leq b) = \int_{-\infty}^{b} \frac{۱}{\sqrt{۲\pi}\sigma} \exp[-\frac{۱}{۲}(\frac{x-\mu}{\sigma})^۲]\, dx.$$

۵. اگر $X \sim N(\mu, \sigma^۲)$

$$Z = \frac{X - \mu}{\sigma} \sim N(۰, ۱),$$

که Z را متغیر تصادفی نرمال استاندارد می‌نامند.

۶. اگر $x_۱, x_۲, \cdots, x_n$ نمونه‌ای تصادفی به حجم n از توزیع $N(\mu, \sigma^۲)$ باشد. $\bar{X}$(میانگین نمونه) دارای توزیع نرمال با میانگین μ و واریانس $\frac{\sigma^۲}{n}$ است($\bar{X} \sim N(\mu, \frac{\sigma^۲}{n})$).

نکته ۱ اگر $x_1, x_2, \cdots, x_n$ نمونه‌ای تصادفی به حجم n از توزیع $N(\mu, \sigma^2)$ باشد. به طوری که $-\infty < X < \infty$، $-\infty < \mu < \infty$ و $\sigma^2 > 0$ با استفاده از قضیه دسته بندی نیمن–فیشر می‌توان اثبات نمود $(\Sigma_{i=1}^n x_i, \Sigma_{i=1}^n x_i^2)$ آماره بسنده مینیمال توام و کامل برای (μ, σ^2) است.

نکته ۲ برای یک نمونه تصادفی به حجم n از توزیع $N(\mu, \sigma^2)$ که $-\infty < X < \infty$، $-\infty < \mu < \infty$ و $\sigma^2 > 0$، $\bar{X}$ و $\frac{1}{n}\Sigma_{i=1}^n (x_1 - \bar{x})^2$ برآوردگرهای ماکسیمم درستنمایی μ و σ^2 هستند.

نکته ۳ برای یک نمونه تصادفی به حجم n از توزیع $N(\mu, \sigma^2)$ که $-\infty < X < \infty$، $-\infty < \mu < \infty$ و $\sigma^2 > 0$، $\bar{X}$ و $\frac{1}{n}\Sigma_{i=1}^n (x_1 - \bar{x})^2$ به ترتیب برآوردگرهای نااریب با کمترین واریانس برای μ و σ^2 هستند.

۱.۲.۱ توزیع‌های نمونه‌گیری

توزیع‌های نمونه‌گیری توزیع‌هایی هستند که براساس یک نمونه معمولاً n تایی از توزیع نرمال بدست می آیند. اینگونه توزیع‌ها با خواص جالب و ویژه خود در حالت مجانبی در مطالعات شبیه‌سازی کاربرد وسیعی دارند. توزیع‌های نمونه‌گیری بسیار زیادی وجود دارند اما مهمترین آن‌ها توزیع‌های χ^2، T و F هستند. که در زیر به اجمال به برخی از ویژگی‌های آنها اشاره می‌شود.

تعریف ۲.۲.۱ در توزیع‌های نمونه‌گیری پارامتری به نام درجه آزادی وجود دارد، درجه آزادی تعداد متغیرهای تصادفی مستقل اینگونه توزیع‌ها است.

توزیع کای دو: ۱. اگر X داری توزیع نرمال استاندارد باشد، متغیر تصادفی $X^۲$ دارای توزیع کای دو با یک درجه آزادی است.($X \sim \chi^۲_{(۱)}$)

۲. در آمار از تابع چگالی زیر با عنوان توزیع کای دو با n درجه آزادی استفاده می‌شود.

$$f(x) = \frac{۱}{۲^{\frac{n}{۲}}\Gamma(\frac{n}{۲})}\, x^{\frac{n-۲}{۲}} e^{-\frac{x}{۲}} \; ; \; x > ۰.$$

۳. امید ریاضی و واریانس متغیر تصادفی $\chi^۲$ با درجه آزادی ν به ترتیب ν و ۲ν است.

۴. اگر $X_i \sim N(۰, ۱), i = ۱, ۲, \cdots, n$ و مستقل از هم باشند، آنگاه $Y = \Sigma_{i=۱}^{n} X_i^۲$ دارای توزیع کای دو با n درجه آزادی است.

۵. اگر $X_i \sim \chi^۲_{(i)}, i = ۱, ۲, \cdots, n$ و مستقل از هم باشند، آنگاه $Y = \Sigma_{i=۱}^{n} X_i$ دارای توزیع کای دو با $Y = \Sigma_{i=۱}^{n} n_i$ درجه آزادی است.

۶. اگر $\bar{X}$ و $S^۲$ میانگین و واریانس نمونه تصادفی به حجم n و مستقل از هم باشند، متغیر تصادفی $\frac{(n-۱)S^۲}{\sigma^۲} = \frac{\Sigma_{i=۱}^{n}(X_i-\bar{X})^۲}{\sigma^۲}$ دارای توزیع کای دو با $(n-۱)$ درجه آزادی است.

توزیع T: ۱. اگر Y, Z متغیرهای‌تصادفی مستقلی باشند، که Y دارای توزیع کای دو با ν درجه آزادی و Z داری توزیع نرمال استاندارد باشند، آنگاه چگالی توزیع $T = \frac{Z}{\sqrt{\frac{Y}{\nu}}}$ دارای توزیع t با ν درجه آزادی است.