بسم الله الرحمن الرحیم

مبانی و مدل‌بندی
بیزی داده‌ها

مبانی و مدل‌بندی
بیزی داده‌ها

با استفاده از برنامه‌نویسی BUGS و نرم‌افزار R

مجتبی گنجعلی
دانشگاه شهید بهشتی

تابان باغفلکی
دانشگاه تربیت مدرس

۱۳۹۶

۶۷۱
مرکز چاپ و انتشارات دانشگاه شهید بهشتی

مبانی و مدل‌بندی بیزی داده‌ها (با استفاده از برنامه‌نویسی BUGS و نرم‌افزار R)

دکتر مجتبی گنجعلی ـ دکتر تابان باغفلکی

ویراستار: دکتر محمدقاسم وحیدی اصل
حروف‌نگار و صفحه‌آرا: سمیرا دهقان
ناظر چاپ : صفر ممیزاد
چاپ اول: ۱۳۹۶
شمارگان : ۵۰۰
قیمت : ۳۰۰.۰۰۰ ریال

سرشناسه:	گنجعلی، مجتبی، ۱۳۴۳-
عنوان و نام پدیدآور:	مبانی و مدل‌بندی بیزی داده‌ها: با استفاده از برنامه‌نویسی BUGS و نرم‌افزار R/ مجتبی گنجعلی، تابان باغفلکی.
مشخصات نشر:	تهران: دانشگاه شهید بهشتی، ۱۳۹۶.
مشخصات ظاهری:	هجده، ۴۰۲ ص.
شابک:	۹۷۸ ۹۶۴ ۴۵۷ ۴۰۵ ۴
فروست:	مرکز چاپ و انتشارات شهید بهشتی، ۶۷۱.
وضعیت فهرست‌نویسی:	فیپا
موضوع:	نظریه تصمیم‌گیری آماری بیزی؛ Bayesian statistical decision theory؛ وین‌باگ؛ WinBUGS؛ آر (زبان برنامه‌نویسی کامپیوتر)؛ R (Computer program language)؛ آمار -- داده‌پردازی؛ Statistics -- Data processing
شناسه افزوده:	باغفلکی، تابان، ۱۳۶۴-
شناسه افزوده:	دانشگاه شهید بهشتی
رده‌بندی کنگره:	۱۳۹۶ ۲م۹گ/QA۲۷۹/۵
رده‌بندی دیویی:	۵۱۹/۵۴۲
شماره کتابشناسی ملی:	۵۰۰۴۱۹۹

کد ناشر ۱۰۰۱۷۳٤

www.pub.sbu.ac.ir
unipress@mail.sbu.ac.ir

تقدیم به فرزندان عزیزم علیرضا و فاطمه

مجتبی گنجعلی

تقدیم به پدر، مادر و خواهر عزیزم

تابان باغفلکی

فهرست مطالب

بخش اول: مبانی آمار بیزی

بخش دوم: مدل‌بندی بیزی داده‌ها

پیشگفتار

پیشرفت قابل‌توجه در نظریه و کاربرد روش‌های استنباط بیزی در تحلیل آماری داده‌ها باعث افزایش علاقهٔ پژوهشگران آماری به یادگیری روش‌های بیزی شده است. در انجام استنباط آماری با استفاده از روش‌های بیزی گاهی لازم است از روش‌های محاسباتی بر مبنای نمونه‌گیری استفاده شود. ازآنجاکه امروزه نرم‌افزارهای زیادی با قابلیت نمونه‌گیری از توزیع‌های آماری در دسترس عموم قرار گرفته‌اند، بسیاری از مسائل پیچیده در آمار کلاسیک با روش‌های بیزی به‌سادگی حل می‌شوند. باتوجه‌به نبود کتاب‌های استنباط بیزی به زبان فارسی در کشور، هدف از ارائهٔ این کتاب فراهم‌آوردن یک نوشتار آموزشی دربارهٔ آمار بیزی است. این کتاب می‌تواند برای پژوهشگران، دانشجویان آمار و دیگر افرادی که به تحلیل آمار کاربردی با استفاده از روش‌های بیزی علاقه‌مند هستند، مفید باشد.

برای مطالعهٔ این کتاب، داشتن اطلاعات پایه‌ای از نظریهٔ احتمال و آمار و همچنین اطلاعاتی در مورد استفاده از نرم‌افزار R مورد نیاز است. در این کتاب که شامل دو بخش (۱) مبانی آمار بیزی و (۲) مدل‌بندی بیزی داده‌ها است، ضمن بیان مفاهیم آمار بیزی، به‌دلیل کاربردی‌بودن این مفاهیم، برنامه‌های کامپیوتری لازم نیز ارائه شده‌اند. این نکته کتاب حاضر را از کتاب‌های موجود داخلی و خارجی متمایز می‌کند. ابزارهای کامپیوتری که در این راستا از آن‌ها بهرهٔ زیادی برده شده است، نرم‌افزارهای R و برنامه‌نویسی به زبان BUGS (با استفاده از نرم‌افزار WinBUGS و OpenBUGS) برای تحلیل بیزی داده‌ها است. استفاده از زبان برنامه‌نویسی BUGS، امکان معرفی روش‌های ساده و انعطاف‌پذیر برای مدل‌بندی داده‌های حتی پیچیده را در بخش دوم کتاب فراهم می‌کند. نکتهٔ قابل‌توجه این است که در این وضعیت مشخص‌سازی چگالی‌های شرطی کامل که در برخی مباحث بیزی به‌کار می‌روند، ضروری نیست. علاوه بر برنامه‌های ارائه‌شده در فصل‌های گوناگون، مروری بر نحوهٔ استفاده از زبان برنامه‌نویسی BUGS و تحلیل بیزی با استفاده از نرم‌افزار R در قالب دو پیوست داده شده‌اند که خوانندگان را در اجرای کاربردی عملیات بیزی یاری می‌دهند. از دیگر مزایای استفاده از این دو نرم‌افزار امکان فراخوانی نرم‌افزار OpenBUGS یا WinBUGS در R است، که با استفاده از بستهٔ نرم‌افزاری R2OpenBUGS یا R2WinBUGS انجام می‌شود.

دنیس لیندلی، آماردان بیزی مشهور، برای استنباط بیزی بر دو قانون بنیادی (۱) پیروی از قوانین احتمال و (۲) مدل‌بندی عدم حتمیت با استفاده از اصول احتمال تأکید کرد. بنابراین، نیاز به دانستن مفاهیم احتمال برای تحلیل بیزی داده‌ها ضروری است.

اُهاگان و فورستر (۲۰۰۴) چهار دلیل زیر را برای ضرورت استفاده از استنباط بیزی به‌جای روش‌های کلاسیک بیان می‌کنند:

۱. استنباط بیزی دارای یک پایهٔ فلسفی محکم بر مبنای قوانین احتمال و از همه مهم‌تر قضیهٔ بیز است.

۲. استنباط بیزی انعطاف‌پذیر است. چارچوب استنباط بیزی شامل ایجاد تابع درستنمایی، تعریف توزیع پیشینی و در نهایت محاسبهٔ توزیع پسینی، با شرطی‌کردن بر داده‌های مشاهده‌شده و با استفاده از قضیهٔ بیز است. این چارچوب قالب انعطاف‌پذیری را برای استنباط در مدل‌های پیچیده فراهم می‌کند.

۳. تفسیر نتایج در استنباط‌های کلاسیک واضح نیست. به‌طور مثال، تفسیر دقیق برآوردهای بازه‌ای و p-مقدار مشکل است. اغلب استفاده‌کنندگان از بازه‌های اطمینان و p-مقدار معنی این کمیت‌ها را نمی‌دانند و به‌طور تلویحی از مفهوم بیزی آن‌ها برای تفسیر نتایج خود استفاده می‌کنند. در مقابل فهم و تفسیر همتای بیزی این کمیت‌ها بسیار ساده است. بازه‌های باورمند بیزی دقیقاً همان چیزهایی هستند که اغلب مردم در مورد یک بازه فکر می‌کنند و یک آزمون فرض بیزی به‌سادگی احتمال پسینی فرض‌ها را برای پذیرش یا رد یک فرض مقایسه می‌کند.

۴. گاهی شیوهٔ گزینش توزیع‌های پیشینی به‌عنوان یک نقطهٔ ضعف استنباط بیزی قلمداد می‌شود. از طرفی قابلیت درنظرگرفتن توزیع‌های پیشینی به این معنی است که اطلاعات بیشتری را می‌توان در استنباط لحاظ کرد (مثال ۱. ۲ از فصل ۱ در مورد موسیقی‌دان و مرد مست را ببینید).

در بخش اول این کتاب توجه خاصی به مفاهیم پایه‌ای آمار بیزی به عمل آمده است. علاوه‌براین، ضمن بیان مباحث اساسی آمار بیزی شامل روش‌های گزینش توزیع‌های پیشینی در استنباط بیزی، روش‌های شبیه‌سازی مونت‌کارلویی، شبیه‌سازی به روش مونت کارلوی زنجیر مارکوفی، آزمون فرض آماری از دیدگاه بیزی و تحلیل بیزی تجربی و سلسله‌مراتبی، در قالب فصل‌های جداگانه، مدل‌های خطی نیز با فرض توزیعی نرمال از دیدگاه بیزی، مدل‌های آمیختهٔ خطی، تحلیل بیزی مدل‌های خطی تعمیم‌یافته، تحلیل مدل‌های زمان تا رخداد پیشامد

و مدل‌بندی توأم اندازه‌های طولی و داده‌های بقا از دیدگاه بیزی در بخش دوم این کتاب گنجانده شده است. فصل اول، شامل مروری اجمالی بر آمار کلاسیک (فراوانی‌گرا) و بیزی، و مقایسهٔ روش‌های بیزی و کلاسیک است. در ادامهٔ مطلب، ابتدا روش‌های گزینش توزیع‌های پیشینی در استنباط بیزی در فصل دوم بحث می‌شوند، سپس در فصل سوم به معرفی روش‌های مونت‌کارلویی برای شبیه‌سازی نمونه از توزیع پسینی می‌پردازیم. این فصل شامل روش انتگرال‌گیری مونت‌کارلویی، روش نمونه‌گیری ردـ پذیرش، روش نمونه‌گیری از نقاط مهم و روش بازنمونه‌گیری نمونه‌گیری از نقاط مهم است. در فصل چهارم به روش‌های مونت‌کارلویی زنجیر مارکوفی برای شبیه‌سازی نمونه از برداری از پارامترها پرداخته می‌شود. این فصل بحث مفصلی در استفاده از الگوریتم متروپولیسـ هستینگس را شامل می‌شود. در فصل پنجم به آزمون فرض به روش بیزی پرداخته شده که علاوه بر بحث در چگونگی آزمون بیزی فرض‌ها، عامل بیزی، pـ مقدار بیزی، بازه‌های باورمند بیزی، چگال‌ترین بازه پسینی و توزیع پسینی پیشگو در آن گنجانده شده است. مدل‌های بیزی سلسله‌مراتبی و بیزی تجربی در فصل ششم گنجانده شده است.

در بخش دوم کتاب، مدل‌های خطی با فرض توزیعی نرمال از دیدگاه بیزی عنوان فصل هفتم است. در این فصل ساختارهای مختلف ماتریس کوواریانس مانند ساختار اتورگرسیو، میانگین متحرک و ساختارهای دیگر بحث شده است. علاوه‌براین چگونگی استفاده از مدل‌های حاشیه‌ای برای مدل‌بندی داده‌های طولی نیز مورد بحث قرار گرفته است. همچنین تحلیل واریانس یک‌طرفه در این فصل بحث شده است. در فصل هشتم به بحث در تحلیل بیزی مدل‌های آمیختهٔ خطی پرداخته شده و از مدل‌های آمیختهٔ خطی برای مدل‌بندی داده‌های گروه‌بندی‌شده و داده‌های طولی استفاده شده است. در فصل نهم مدل‌های خطی تعمیم‌یافته از دیدگاه بیزی بحث شده است که در این فصل به مدل‌بندی داده‌های دودویی، دوجمله‌ای، و داده‌های شمارشی در هر دو حالت مقطعی و طولی از دیدگاه بیزی پرداخته شده است. فصل دهم بعد از مقدمه‌ای بر تحلیل بقا، به روش‌های تحلیل بیزی داده‌های بقا می‌پردازد که حاوی برخی از پژوهش‌ها و برنامه‌های نوشته‌شده توسط نویسندگان در این راستا است. بالاخره در فصل ۱۱ به مدل‌بندی توأم داده‌های طولی و داده‌های بقا از دیدگاه بیزی پرداخته می‌شود. همچنین، همان‌طورکه گفته شد، آشنایی با عملیات بیزی با استفاده از نرم‌افزار R و آشنایی با زبان برنامه‌نویسیBUGS در دو پیوست ارائه شده‌اند.

این کتاب برای همهٔ علاقه‌مندان به یادگیری تحلیل بیزی داده‌ها قابل‌استفاده است. با وجود

این فصل‌های ذکرشده در زیر برای تدریس در دوره‌های مختلف تحصیلی پیشنهاد می‌شوند. برای مثال در دورهٔ کارشناسی آمار تدریس فصل‌های ۱، ۲، ۳ و ۵ بخش اول کتاب، به شرط آنکه دانشجو درس آمار ریاضی ۲ را گذرانده باشد، توصیه می‌شود.

بخش اول

فصل‌های ۱ تا ۳	فصل ۴	فصل ۵	فصل ۶
• کارشناسی آمار • کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی	• کارشناسی ارشد آمار (گرایش آمار ریاضی) • دکتری آمار و آمار زیستی	• کارشناسی آمار • کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی	• کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی

بخش دوم

فصل ۷ و ۸	فصل ۹	فصل ۱۰	فصل ۱۱
• کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی	• دکتری آمار و آمار زیستی • دکتری همه‌گیرشناسی	• کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی	• کارشناسی ارشد آمار (همهٔ گرایش‌ها) • دکتری آمار و آمار زیستی • کارشناسی‌ارشد و دکتری همه‌گیرشناسی

از ویراستار محترم این کتاب آقای دکتر محمدقاسم وحیدی اصل کمال تقدیر و تشکر را داریم. همچنین از انتشارات دانشگاه شهید بهشتی به‌ویژه جناب آقای دکتر محمدرضا نبید، همکاران بزرگوار سرکار خانم‌ها آذرمه سنجری و خانم سمیرا دهقان برای آماده‌سازی و چاپ کتاب تشکر می‌کنیم.

به‌عنوان کلام آخر تقاضامندیم که ما را از انتقادها و پیشنهادهای خود در جهت ارتقای کیفی کتاب بهره‌مند سازید و همچنین سپاسگزار خواهیم بود اگر نظرها، پیشنهادها و سؤال‌های احتمالی خود را با ما در میان بگذارید.

مجتبی گنجعلی، عضو هیئت علمی گروه آمار دانشگاه شهید بهشتی
تابان باغفلکی، عضو هیئت علمی گروه آمار دانشگاه تربیت مدرس
پاییز ۱۳۹۶

جدول توابع چگالی استفاده‌شده

نام توزیع	پارامترها	دامنهٔ متغیر	تابع چگالی	نماد اختصاری
نرمال	μ,σ^2 $(\mu \in R, \sigma^2 > 0)$	$x \in R$	$f(x \mid \mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	$N(\mu,\sigma^2)$
نمایی	$\lambda\ (\lambda > 0)$	$x > 0$	$f(x \mid \lambda)=\lambda e^{-\lambda x}$	$\mathrm{Exp}(\lambda)$
گاما	$\alpha,\beta\ (\alpha,\beta > 0)$	$x > 0$	$f(x \mid \alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$	$\Gamma(\alpha,\beta)$
گامای معکوس	$\alpha,\beta\ (\alpha,\beta > 0)$	$x > 0$	$f(x \mid \alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}\exp\left(-\frac{\beta}{x}\right)$	$\mathrm{I}\Gamma(\alpha,\beta)$
بتا	$\alpha,\beta\ (\alpha,\beta > 0)$	$0<x<1$	$f(x \mid \alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\mathrm{Beta}(\alpha,\beta)$
کوشی	μ,τ $(\mu \in R, \tau > 0)$	$x \in R$	$f(x \mid \mu,\tau)=\frac{1}{\tau\pi\left(1+\left(\frac{x-\mu}{\tau}\right)^2\right)}$	$C(\mu,\tau)$
یکنواخت	$a,b\ (a,b \in R)$	$x \in [a,b]$	$f(x \mid a,b)=\frac{1}{b-a}I_{[a,b]}(x)$	$U(a,b)$
وایبول	$\lambda,\gamma\ (\lambda,\gamma > 0)$	$x > 0$	$f(x \mid \lambda,\gamma)=\lambda\gamma x^{\gamma-1}e^{-\lambda x^\gamma}$	$\mathrm{Weibull}(\lambda,\gamma)$
لگ نرمال	μ,σ^2 $(\mu \in R, \sigma^2 > 0)$	$x > 0$	$f(x \mid \mu,\sigma^2)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}$	$LN(\mu,\sigma^2)$
لگ لوژستیک	μ,σ^2 $(\mu \in R, \sigma^2 > 0)$	$x > 0$	$f(x \mid \mu,\sigma^2)=\frac{\exp\left(\frac{\log(x)-\mu}{\sigma}\right)}{x\sigma\left(1+\exp\left(\frac{\log(x)-\mu}{\sigma}\right)\right)^2}$	$LL(\mu,\sigma^2)$
t	$v\ (v > 0)$	$x \in R$	$f(x \mid v)=\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}\frac{1}{\sqrt{v\pi}}\left(1+\frac{x^2}{2}\right)^{-\frac{v+1}{2}}$	$\mathrm{T}(v)$
برنولی	$p\ (0 < p < 1)$	$x \in \{0,1\}$	$f(x \mid p)=p^x(1-p)^{1-x}$	$\mathrm{Ber}(p)$
دوجمله‌ای	n,p $(0 < p < 1, n = 1,2,\ldots)$	$x \in \{0,1,\ldots,n\}$	$f(x \mid n,p)=\binom{n}{x}p^x(1-p)^{n-x}$	$\mathrm{Bin}(n,p)$

ادامهٔ جدول توابع چگالی استفاده‌شده

نام توزیع	پارامترها	دامنهٔ متغیر	تابع چگالی	نماد اختصاری
پواسون	$\lambda(\lambda>0)$	$x\in\{0,1,2,\ldots\}$	$f(x\mid\lambda)=\frac{e^{-\lambda}\lambda^{x}}{x!}$	$\mathrm{Pos}(\lambda)$
دیریکله	$\alpha_i>0,$ $i=1,2,\ldots,k$	$0<\pi_i<1,$ $\sum_{i=1}^{k}\pi_i=1$	$f(\boldsymbol{\pi}\mid\boldsymbol{\alpha})=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\pi_i^{\alpha_i-1}$	Dirichlet($\boldsymbol{\alpha}$), $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_k)$
ویشارت معکوس	$\nu>p-1$ $\boldsymbol{\Psi}$ ماتریسی $p\times p$ و معین مثبت است.	$\mathbf{x}$ ماتریسی $p\times p$ و معین مثبت است.	$f(\mathbf{x}\mid\boldsymbol{\Psi},\nu)=\frac{\mid\boldsymbol{\Psi}\mid^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})}\mid\mathbf{x}\mid^{-\frac{\nu+p+1}{2}}e^{-\frac{1}{2}trace(\boldsymbol{\Psi}\mathbf{x}^{-1})}$	$IW(\boldsymbol{\Psi},\nu)$
مقدار کرانگین	$\mu\in R,\sigma>0$	$x\in R$	$f(x)=\frac{1}{\sigma}e^{-(x-\mu)/\sigma}e^{-e^{-(x-\mu)/\sigma}}$	$EV(\mu,\sigma)$

بخش اول

مبانی آمار بیزی

فصل ۱

مقدمه‌ای بر استنباط بیزی

۱.۱. مقدمه

اصطلاح "آمار بیزی" از نام ریاضی‌دان و کشیش انگلیسی تامس بیز (۱۷۰۲-۱۷۶۱) به‌دلیل ارائۀ قضیۀ بیز (هاجز و لی‌من، ۱۹۷۰ و مِندِن‌هال و همکاران، ۲۰۰۹) اقتباس شده است. در واقع، قضیۀ معروف بیز نقطۀ شروع آمار بیزی است. استنباط در روش‌های بیزی براساس این قضیه با ترکیب اطلاعات پیشینیِ غیرنمونه‌ای و داده‌های موجود صورت می‌گیرد. اطلاعات پیشینی غیرنمونه‌ای اطلاعاتی دربارۀ کمیت مورد بررسی است که غالباً از منابع دیگر گرفته می‌شود. به‌طور مثال، تجربیات قبلی در وضعیت‌های مشابه منبع خوبی از این اطلاعات پیشینی است. اطلاعات پیشینی کمک می‌کند تا با استفاده از آن‌ها یک توزیع پیشینی، که حاوی این اطلاعات است، در نظر گرفته شود. این توزیع پیشینی را می‌توان (درست مانند آن جنبه‌های آماری که یک توزیع برای کمیت تحت بررسی درنظر می‌گیرد) با استفاده از مقادیر گذشتۀ یک پارامتر به‌دست آورد.

فرض کنید یک شرکت پخش دارو قصد خرید یک مُسکن جدید را دارد. دو مورد از مهم‌ترین عواملی که بر تصمیم‌گیری شرکت اثر می‌گذارد. عبارت‌اند از: نسبت افرادی که دارو برای آن‌ها مؤثر است (θ_1) و سهمی از بازار که دارو (θ_2) به‌دست خواهد آورد. در این مثال اگرچه هر دو پارامتر θ_1 و θ_2 نامعلوم‌اند، اما می‌توان قبل از جمع‌آوری هر داده‌ای یک منبع غنی از اطلاعات دربارۀ این دو پارامتر را از داروهای مسکن مشابه به‌دست آورد. به‌عنوان مثالی دیگر، احتمال ابتلای خفیف به تالاسمی یا تالاسمی مینور فرد O (آن را با ω نشان می‌دهیم) را در نظر بگیرید. به فردی که یک ژن تالاسمی را از والدین خود به ارث می‌برد، ناقل ژن می‌گویند و اصطلاحاً گویند این فرد مبتلا به تالاسمی خفیف است. اطلاع از اینکه یکی از والدین