Statistical Sciences Extension
Vol. 2, No. 1, Winter 2017, (P 43-54)

گسترهٔ علوم آماری
سال دوم، شماره اول، پاییز و زمستان 1395، (ص 43-54)

مقایسهٔ مدل‌های توزیع بتا-دوجمله‌ای دومتغیرهٔ گسسته بر اساس همبستگی بین متغیرهای حاشیه‌ای

حسین پاشا زانوسی[1]، عبدالله سعادتمند[2*]

1. کارشناسی ارشد، گروه آمار، دانشگاه علوم و فنون دریایی خرمشهر
2. استادیار، گروه آمار، دانشگاه پیام نور

تاریخ دریافت: 1395/09/09 **تاریخ پذیرش**: 1395/12/17

Comparison of Discrete Bivariate Beta - Binomial Distributions Based on Correlation Between Marginal Variables

H. Pashazanoosi[1], A. Saadatmand[*2]

1. M.Sc., Department of Statistics, Khorramshahr University of Marine Science and Technology
2. Assistant Professor, Department of Statistics, Payame Noor University

Received: 2016/11/29 **Accepted:** 2017/03/07

چکیده

در این تحقیق برازش مدل‌های مختلف توزیع‌های بتا-دوجمله‌ای دومتغیرهٔ گسسته، بر اساس همبستگی بین متغیرهای حاشیه‌ای مورد مقایسه قرار می‌گیرد. این مدل‌ها شامل مدل سه پارامتری بی‌بی و وات (2011)، مدل پنج پارامتری داناهر و هاردی (2005) و مدل تعمیم‌یافتهٔ توزیع بتا-دوجمله‌ای دومتغیرهٔ کلاسیک است که الموجیمینز و همکاران (2011) معرفی کرده‌اند. نتایج حاصل از آزمون نیکویی برازش نشان می‌دهد که مدل الموجیمینز و همکاران، برای مقادیر بالای همبستگی بین متغیرهای حاشیه‌ای، برازش بهتری نسبت به مدل‌های دیگر دارد و در مقادیر پایین همبستگی بین متغیرهای حاشیه‌ای، مدل داناهر و هاردی مناسب‌تر است. نتایج با استفاده از سه مثال واقعی بررسی شده است.

واژگان کلیدی

آزمون نیکویی برازش، توزیع بتا-دوجمله‌ای دومتغیره، همبستگی بین متغیرهای حاشیه‌ای.

Abstract

The aim of this study was to compare fitting of different discrete bivariate beta-binomial distributions based on correlation between marginal variables. The models included bivariate beta-binomial distribution proposed by Bibby and Væth (2011) (with three parameters), Danaher and Hardie (2005) (with five parameters) and generalized model of the classical bivariate beta-binomial distribution proposed by Olmo - Jimenez et al. (2011). Based on results obtained from goodness of fit test, Olmo - Jimenez et al's model was found to be more appropriate than other models when correlation between marginal variables was high. Also Danaher and Hardie's model were found to be more appropriate than other models when correlation between marginal variables was low. The results were presented by using three real data set.

Keywords

Goodness of Fit Test, Bivariate Beta-Binomial Distribution, Correlation Between Marginal Variables.

* **نویسنده مسئول**: عبداله سعادتمند

Corresponding Author: abdollah.saadatmand@gmail.com

مقدمه

توزیع بتا-دوجمله‌ای برای مدل کردن تعداد موفقیت‌ها برای آزمایش‌های دوجمله‌ای در شرایطی که به خاطر وجود ناهمگونی در بین افراد، بیش پراکنش مشاهده می‌شود، به کار می‌رود (جانسون[1] و همکاران) (2005). مفهوم بیش پراکنش حداقل به یک مقاله از فیشر[2] (1925) برمی‌گردد که در آن نسبت جنسی در خانواده‌های آلمانی دارای هشت فرزند مورد بحث قرار گرفته است. فیشر نشان داد که تعداد پسرهای خانواده‌ها با استفاده از توزیع دوجمله‌ای، خوب توصیف نشده است و نتیجه گرفت تعداد خانواده‌هایی که تعداد پسر و دختر برابر ندارند، زیاد و تعداد خانواده‌هایی که تعداد پسرها و دخترها برابر یا تقریباً برابر است، کم هستند.

یکی از دلایل بیش پراکنش این است که پارامتر p در توزیع دوجمله‌ای، از یک آزمایش به آزمایش دیگر به صورت یک متغیر تصادفی پیوسته تغییر می‌کند ($0 < p < 1$). در بیشتر مقالات برای پارامتر p، توزیع بتا را پیشنهاد داده‌اند. توزیع بتا-دوجمله‌ای (یک متغیره)، نتیجهٔ چنین ساختاری است که به صورت آمیختهٔ توزیع دوجمله‌ای با پارامتر p که خود توزیع بتا با پارامترهای α و β دارد، به صورت زیر معرفی می‌شود:

$$f(x,a,b)=\int_0^1 \binom{n}{x} p^{x}(1-p)^{n-x}\frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}dp$$

$$x=0,1,\ldots,n,$$

$$=\binom{n}{x}\frac{B(a+x,b+n-x)}{B(a,b)},$$

به طوری که $B(.,.)$ تابع بتا است. توزیع بتا-دوجمله‌ای را اولین بار به صورت رسمی اسکلام[3] (1948) ارائه کرد.

توزیع‌های بتا-دوجمله‌ای دومتغیرهٔ گسسته به منظور آنالیز مجموعه‌ای از داده‌های شمارشی همبسته مطرح گردیدند. این توزیع‌ها به خاطر مشکلات محاسباتی، در عمل به‌ندرت به کار می‌روند. توزیع بتا-دوجمله‌ای دومتغیره را اولین بار ایشی و هایاکاوا[4] (1960) معرفی کرده است. این الگو را به طور مشابه اما مستقل، گلمن و سیشل[5] (1987) و همچنین آلانکو و لمنز[6] (1996) به کار گرفته‌اند. آنها این توزیع را از آمیختهٔ توزیع توأم دو متغیر دوجمله‌ای مستقل با پارامتر یکسان p که p خود از توزیع بتا پیروی می‌کند، مطابق زیر به دست آورده‌اند:

$$f(x_1,x_2)=\int_0^1 \binom{n_1}{x_1}\binom{n_2}{x_2}p^{x_1+x_2}(1-p)^{n_1+n_2-(x_1+x_2)}\frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}dp$$

$$=\binom{n_1}{x_1}\binom{n_2}{x_2}\frac{B(x_1+x_2+a,n_1+n_2+b-x_1-x_2)}{B(a,b)},$$

به طوری که $x_1=0,1,\ldots,n_1$ و $x_2=0,1,\ldots,n_2$ و $a,b>0$.

این توزیع را می‌توان به وسیلهٔ تابع فوق هندسی اپل[7] F_1 به صورت زیر نوشت:

$$f(x_1,x_2)=f_0\frac{(a)_{x_1+x_2}(-n_1)_{x_1}(-n_2)_{x_2}}{(-b-(n_1+n_2)+1)_{x_1+x_2}}\frac{1}{x_1!x_2!},$$

به طوری که $x_1=0,1,\ldots,n_1$ و $x_2=0,1,\ldots,n_2$ و f_0 ثابت نرمال‌ساز برابر است با:

$$f_0=F_1(a;-n_1;-n_2;-b-(n_1+n_2)-1;1,1)^{-1}.$$

همچنین $(a)_{x_1+x_2}$ نماد پوش هامر[8] است که از رابطهٔ زیر محاسبه می‌شود:

$$(a)_{x_1+x_2}=\frac{G(a+x_1+x_2)}{G(a)}.$$

1. Johnson
2. Fisher
3. Skellam
4. Ishii and Hayakawa
5. Gelman and sichel
6. Alanko and Lemmens
7. Appell
8. Pochhammer

این توزیع شامل دو پارامتر است و از ویژگی‌های مهم آن داشتن توزیع‌های حاشیه‌ای بتا-دوجمله‌ای و در صورتی که $n_1 = n_2$ باشد، یکسان بودن توزیع‌های حاشیه‌ای است که در این حالت اصطلاحاً می‌گویند مدل متقارن است. همچنین منحنی رگرسیون آن، خطی است و تنها اجازهٔ همبستگی مثبت بین دو متغیر را می‌دهد که این ویژگی‌ها به منظور برازش مناسب داده‌ها، یک نقطه ضعف محسوب می‌شود. مدل فوق با نام مدل کلاسیک توزیع بتا-دوجمله‌ای دومتغیره در طول این تحقیق بیان می‌شود. در ادامه مدل‌های دیگر توزیع بتا-دوجمله‌ای دومتغیره معرفی می‌شوند و برازش آنها با توجه به میزان همبستگی بین متغیرهای حاشیه‌ای مورد بررسی قرار می‌گیرد.

مدل‌های توزیع بتا-دوجمله‌ای دومتغیره

بعد از معرفی الگوی مدل کلاسیک توزیع بتا-دوجمله‌ای دومتغیره، سه مدل دیگر از این توزیع مطرح گردید که در این تحقیق، این مدل‌ها مورد بررسی قرار می‌گیرند. در سال 2005 مدلی از این توزیع را داناهر و هاردی[1] ارائه کرده‌اند که شامل پنج پارامتر است. پارامتر پنجم آن مربوط به همبستگی بین دو متغیر حاشیه‌ای است که در صورت معنادار بودن در مدل می‌ماند. مقدار همبستگی می‌تواند مثبت یا منفی باشد. در سال 2011 بی‌بی و وات[2] مدل سه پارامتری این توزیع را ارائه دادند که تنها اجازهٔ همبستگی مثبت را بین دو متغیر می‌دهد و دارای توزیع حاشیه‌ای بتا-دوجمله‌ای است. مدل دیگری را نیز در سال 2011 المو-جیمینز[3] و همکاران پیشنهاد دادند که در حقیقت مدل کلاسیک توزیع بتا-دوجمله‌ای دومتغیره را به سه یا چهار پارامتر تعمیم می‌دهد. مدل سه پارامترهٔ آن در صورت برابری $n_1 = n_2$، متقارن است. این مدل بیان صریحی برای محاسبهٔ گشتاورها از قبیل کوواریانس ندارد و در عمل تنها اجازهٔ همبستگی مثبت را می‌دهد.

مدل بی‌بی و وات

توزیع مدل بی‌بی و وات با استفاده از ساختاری مشابه آنچه اسکلام (1948) بیان کرده است، معرفی می‌شود. این کار مستلزم انتگرال‌گیری نسبت به چگالی توزیع بتای دومتغیره است. توزیع بتای دومتغیره را جونز[4] (2001) معرفی کرده است (همچنین الکین و لیو[5]، 2003). فرض کنید W_0، W_1 و W_2 متغیرهای تصادفی دو به دو مستقل باشند و $W_i \sim c^2(2n_i)$ که معادل $G(n_i, 2)$ با شرط $n_i > 0$ است، برقرار باشد؛ B_i را به صورت زیر تعریف می‌کنیم:

$$B_i = \frac{W_i}{W_i + W_0} \quad i = 1, 2.$$

در نتیجه B_i، توزیع $Beta(n_i, n_0)$ برای $i = 1, 2$ خواهد داشت. توزیع توأم B_1 و B_2، توزیع بتای دومتغیره نامیده می‌شود که چگالی آن به صورت زیر است:

$$f_B(b_1, b_2) = \frac{G(n)}{G(n_1)G(n_2)G(n_0)} \prod_{j=1}^{2} \frac{b_j^{n_j - 1}}{(1-b_j)^{n_j+1}} \left(1 + \sum_{j=1}^{2} \frac{b_j}{1-b_j}\right)^{n}$$

$$= \frac{G(n)}{G(n_1)G(n_2)G(n_0)} \frac{b_1^{n_1-1}(1-b_1)^{n_2+n_0-1} b_2^{n_2-1}(1-b_2)^{n_1+n_0-1}}{(1-b_1b_2)^{n}},$$

در صورتی که $n = n_1 + n_2 + n_0$ و $0 < b_2 < 1$ و $0 < b_1 < 1$.

حال فرض کنید $P = (p_1, p_2)$، بتای دومتغیره باشد که با پارامترهای n_0 و n_1 و n_2 توزیع شده‌اند، همچنین با مشخص بودن توزیع p، X_1 و X_2 مستقل و به صورت دوجمله‌ای توزیع شده باشند، یعنی $X_i \sim bin(n_i, p_i)$، آنگاه توزیع توأم $X = (X_1, X_2)$، توزیع سه پارامتری مدل بی‌بی و وات نامیده می‌شود و تابع جرم احتمال آن به صورت زیر خواهد بود:

1. Danaher and Hardie
2. Bibby and Væth
3. Olmo-Jimenez
4. Jones
5. Olkin and Lio

$$f_X(x_1,x_2)=\int_{[0,1]^2} f_{X|p}(x|p)f_p(p)dp$$

$$=\binom{n_1}{x_1}\binom{n_2}{x_2}\frac{\Gamma(n)}{\Gamma(n_0)\Gamma(n_1)\Gamma(n_2)}\frac{\Gamma(x_1+n_1)\Gamma(n_1-x_1+n-n_1)}{\Gamma(n_1+n)} \quad (1)$$

$$\frac{\Gamma(x_2+n_2)\Gamma(n_2-x_2+n-n_2)}{\Gamma(n_2+n)},\ {}_3F_2(n,x_1+n_1,x_2+n_2;n_1+n,n_2+n;1),$$

برای $x_2=0,1,\ldots,n_2$ و $x_1=0,1,\ldots,n_1$.

مدل داناهر و هاردی

توزیع مدل پنج پارامتری داناهر و هاردی بر اساس وجود دو همبستگی بنا شده است. نخست همبستگی که ممکن است بین دو متغیر وجود داشته باشد و دوم همبستگی که ممکن است بین آزمایش‌های متوالی در هر یک از دو متغیر به طور جداگانه وجود داشته باشد. نادیده گرفتن هر یک از این همبستگی‌ها موجب می‌شود که برآورد ضعیفی برای رفتار این گونه داده‌ها داشته باشیم.

توزیع یک متغیره شامل تعداد موفقیت در n آزمایش را در نظر بگیرید. فرض کنید Y_jها متغیر تصادفی برنولی باشند، در این صورت $X=\sum_{j=1}^{n}Y_j$ دارای توزیع دوجمله‌ای (n,p) خواهد بود. این مدل، حاوی بیش پراکنش در X که در نتیجهٔ همبستگی بین آزمایش‌های متوالی Y_j، به دست می‌آید، نیست. اگر هر فرد دارای احتمال موفقیت p باشد، طوری که p خود یک متغیر تصادفی باشد، در این صورت فرض می‌شود که $Y_j|p$ آزمایش‌های برنولی هستند، یعنی $X|p\sim bin(n,p)$. همبستگی غیر شرطی در طول مشاهدات دوتایی با این فرض که p توزیع بتا دارد، تضمین می‌گردد و این منتهی به توزیع بتا-دوجمله‌ای برای توزیع X می‌شود.

حال می‌توان حالت یک متغیره را با تعریف $X_i=Y_{i1}+Y_{i2}+\ldots+Y_{in}$، به طوری که Y_{ij} متغیر تصادفی برنولی با احتمال موفقیت p_i و $X_i=0,1,\ldots,n_i, i=1,2$ است، به حالت دومتغیره تعمیم داد. در مسائل کاربردی برای توزیع‌های حاشیه‌ای معمولاً همبستگی $(Y_{ij},Y_{ij'})$ مخالف صفر است. در این شرایط، منطقی است که فرض کنیم X_i دارای توزیع بتا-دوجمله‌ای باشد. علاوه بر وجود همبستگی درون آزمایش‌ها، ممکن است بین X_1 و X_2 نیز همبستگی وجود داشته باشد. با استفاده از روشی قابل قیاس با مدل بتا-دوجمله‌ای، می‌توان $(X_1,X_2|p_1,p_2)$ را به‌عنوان توزیع دوجمله‌ای دومتغیره بیان کرد، به طوری که متغیرهای X_1 و X_2 با شرط (p_1,p_2) مستقل هستند و همچنین $X_i|p_i\sim bin(n_i,p_i), i=1,2$. همبستگی غیر شرطی بین X_1 و X_2 از طریق توزیع دومتغیره (p_1,p_2) که اجازهٔ همبستگی بین p_1 و p_2 را می‌دهد، معرفی می‌شود. هدف ما پیدا کردن توزیع دومتغیره (p_1,p_2) است که حاشیه‌های آن توزیع بتا داشته باشد، یعنی توزیع غیر شرطی (X_1,X_2)، دارای توزیع حاشیه‌ای بتا-دوجمله‌ای باشد؛ بنابراین همبستگی درون متغیرها اصلاح می‌شود. در نتیجه نیاز به چگالی توأم (p_1,p_2) است که بتواند متغیرهایی که به صورت مثبت یا منفی همبسته هستند، پوشش دهد و توزیع‌های حاشیه‌ای بتا داشته باشد. چنین توزیع دومتغیره‌ای به وسیلهٔ لی[1] (1996) با استفاده از چارچوبی که سارمانف[2] (1966) معرفی کرده است، ارائه گردید. شکل عمومی توزیع دومتغیرهٔ سارمانف برای (p_1,p_2) با توزیع‌های حاشیه‌ای مشخص $f_1(p_1)$ و $f_2(p_2)$ برابر است با:

$$g(p_1,p_2)=f_1(p_1)f_2(p_2)\left[1+w f_1(p_1)f_2(p_2)\right],$$

به طوری که $f_i(p_i)$ تابع آمیخته نامیده می‌شود که یک تابع غیر ثابت کراندار است و رابطهٔ $\int f_i(t)f_i(t)dt=0$ برقرار است. پارامتر w

1. Lee
2. Sarmanov

همبستگی بین p_1 و p_2 را مشخص می‌کند و باید شرط $g(p_1,p_2)$ برای $1+wf_1(p_1)f_2(p_2)>0$ به‌عنوان تابع چگالی توأم به ازای هر p_1 و p_2 برقرار باشد.

لی (1996)، تابع آمیخته به شکل $f_i(p_i)=p_i-\mu_i$ با توزیع‌های حاشیه‌ای بتا ارائه داد. نتیجهٔ آن توزیع بتای دومتغیره زیر است:

(2)
$$g(p_1,p_2)=f_1(p_1|a_1,b_1)f_2(p_2|a_2,b_2)\left[1+w(p_1-\mu_1)(p_2-\mu_2)\right]$$

وقتی $w=0$ باشد، رابطهٔ فوق به حاصل‌ضرب دو بتای یک متغیره (یعنی p_1 و p_2 مستقل هستند)، تبدیل می‌شود.

توزیع غیر شرطی (X_1,X_2) به کمک آمیختهٔ (X_1,X_2) که روی (p_1,p_2) شرطی شده با توزیع بتای دومتغیرهٔ سارمانف (رابطهٔ (2)) به صورت زیر به دست می‌آید:

$$P(X_1=x_1,X_2=x_2|n_1,n_2,a_1,a_2,b_1,b_2,w)=$$

$$\int_0^1\int_0^1 P(X_1=x_1|n_1,p_1)\,P(X_2=x_2|n_2,p_2)\,g(p_1,p_2)\,dp_1dp_2$$

$$=P_{BB}(X_1=x_1|n_1,a_1,b_1)P_{BB}(X_2=x_2|n_2,a_2,b_2)$$

(3)
$$\left[1+w\frac{(x_1-n_1\mu_1)(x_2-n_2\mu_2)}{(a_1+b_1+n_1)(a_2+b_2+n_2)}\right],$$

به طوری که $P_{BB}(X_i=x_i|n_i,a_i,b_i)$ همان تابع چگالی احتمال توزیع بتا-دوجمله‌ای با پارامترهای a_i و b_i است:

$$P_{BB}(X_i=x_i|n_i,a_i,b_i)=\binom{n_i}{x_i}\frac{B(a_i+x_i,b_i+n_i-x_i)}{B(a_i,b_i)}$$

برای اینکه مطمئن شویم تابع چگالی احتمال داده شده در رابطهٔ (3) نامنفی است، w باید در شرط زیر صدق کند:

$$\max\left[\frac{-1}{\mu_1\mu_2},\frac{-1}{(1-\mu_1)(1-\mu_2)}\right]\le w$$

$$\le\min\left[\frac{-1}{\mu_1(1-\mu_2)},\frac{-1}{\mu_2(1-\mu_1)}\right]\frac{(a_1+b_1+n_1)(a_2+b_2+n_2)}{n_1n_2}$$

شرط فوق با تعیین نقاط حدی $(0,0),(0,n_2),(n_1,0)$ و (n_1,n_2) در رابطهٔ (3) حاصل می‌شود.

مدل الموجیمینز و همکاران

توزیع مدل کلاسیک بتا-دوجمله‌ای دومتغیره تولید شده به وسیلهٔ تابع فوق هندسی اپل F_1، با قید $\lambda_1=\lambda_2=1$ بوده است. بسط این توزیع شامل حذف این محدودیت و جایگزین کردن آن با این فرض است که توزیع تولید شده به وسیلهٔ تابع فوق هندسی اپل شامل دو پارامتر $\lambda_1,\lambda_2>0$ است؛ بنابراین تابع چگالی احتمال مدل الموجیمینز و همکاران، همانند آنچه برای توزیع مدل کلاسیک بتا-دوجمله‌ای دومتغیره در بخش مقدمه بیان شده، به شکل زیر خواهد شد:

(4)
$$f(x_1,x_2)=f_0\frac{(a)_{x_1+x_2}(-n_1)_{x_1}(-n_2)_{x_2}}{(-b-(n_1+n_2)+1)_{x_1+x_2}}\frac{\lambda_1^{x_1}\lambda_2^{x_2}}{x_1!x_2!},$$

به طوری که $x_1=0,1,\ldots,n_1$، $x_2=0,1,\ldots,n_2$، $(a)_{x_1+x_2}$ نماد پوش هامر و f_0 ثابت نرمال‌ساز با رابطهٔ زیر است:

$$f_0=F_1(a;-n_1;-n_2;-b-(n_1+n_2)-1;\lambda_1,\lambda_2)^{-1}.$$

می‌توان مشاهده کرد که مدل کلاسیک توزیع بتا-دوجمله‌ای دومتغیره و تعمیم‌یافتهٔ آن تنها در عامل مقیاس و عامل‌های $\lambda_1^{x_1}$ و $\lambda_2^{x_2}$ با هم تفاوت دارند. پارامترهای λ_1,λ_2 قابل تفسیر هستند. به‌عنوان مثال، وقتی $n_1 = n_2$ باشد، متفاوت بودن مقادیر دو پارامتر نشان می‌دهد که توزیع‌های حاشیه‌ای یکسان نیستند که این جمله پایدار نبودن مدل را در دو دورهٔ زمانی بیان می‌کند. در چنین مواردی، اگر $\lambda_i\ i=1,2$، کوچک‌تر یا بزرگ‌تر از یک باشد، احتمالات به ترتیب در مقادیر پایین یا بالای X_i متمرکز می‌گردند.

در مدل کلاسیک توزیع بتا-دوجمله‌ای دومتغیره، توزیع احتمالات روی محور از نقطهٔ $(0,0)$ به نقطهٔ (n_1,n_2) متمرکز می‌شوند، اما پراکندگی احتمالات در مدل تعمیم‌یافتهٔ آن ممکن است محدود به این محور نباشد؛ بنابراین پارامترهای λ_2 و λ_1 باعث جامع‌تر شدن توزیع می‌شوند.

در اینجا ثابت می‌کنیم توزیع مدل الموجیمینز و همکاران، ممکن است با در نظر گرفتن احتمالات موفقیت متفاوت در هر یک از متغیرهای حاشیه‌ای X_2 و X_1، به صورت آمیخته نیز مشاهده گردد. فرض می‌شود که $X_i \mid P_i = p_i, i=1,2$، از توزیع دوجمله‌ای با پارامترهای مستقل n_i و p_i به صورت زیر تبعیت کند:

$$p_i = \frac{\lambda_i n}{1-(1-\lambda_i)n},$$

به طوری که V توزیع بتای تعمیم‌یافته اکستونز با $0 < V < 1$ و تابع چگالی زیر باشد (گوپتا و نادارجا[1]، 2004 و فام-گیا و دونگ[2]، 1989):

$$f_V(v) = \frac{v^{a-1}(1-v)^{b-1}}{B(a,b)} \frac{(1-(1-\lambda_1)v)^{n_1}(1-(1-\lambda_2)v)^{n_2}}{F_1(a;-n_1;-n_2;a+b;1-\lambda_1,1-\lambda_2)},$$

در نتیجه، آمیختهٔ دوجمله‌ای دوبل با رابطهٔ فوق، توزیع غیر شرطی (X_1,X_2) به دست می‌آید. تابع چگالی آن عبارت است از:

$$f(x_1,x_2) = \int_0^1 f(x_1,x_2 \mid V=v) f(v)\,dv = \binom{n_1}{x_1}\binom{n_2}{x_2} \frac{1}{B(a,b)}$$

$$\times \frac{\lambda_1^{x_1}\lambda_2^{x_2}}{F_1(a;-n_1;-n_2;a+b;1-\lambda_1,1-\lambda_2)} \int_0^1 v^{a+x_1+x_2-1}(1-v)^{b+n_1+n_2-(x_1+x_2)-1}\,dv$$

$$= \binom{n_1}{x_1}\binom{n_2}{x_2} \frac{B(a+x_1+x_2, b+n_1+n_2-(x_1+x_2))}{B(a,b)} \lambda_1^{x_1}\lambda_2^{x_2} \frac{1}{F_1(a;-n_1;-n_2;a+b;1-\lambda_1,1-\lambda_2)},$$

به طوری که $a,b,\lambda_1,\lambda_2 > 0$ و $x_2 = 0,1,\ldots,n_2$ و $x_1 = 0,1,\ldots,n_1$.

با توجه به اینکه:

$$F_1(a;-n_1;-n_2;a+b;1-\lambda_1,1-\lambda_2) = \frac{F_1(a;-n_1;-n_2;-b-(n_1+n_2)-1;\lambda_1,\lambda_2)}{F_1(a;-n_1;-n_2;-b-(n_1+n_2)-1;1,1)} \quad (5)$$

تابع چگالی آن به صورت رابطهٔ (4) کاهش پیدا می‌کند.

یک حالت از توزیع مدل الموجیمینز و همکاران، این است که وقتی $\lambda_1 = \lambda_2 = \lambda$ باشد، توزیع سه پارامتری ایجاد می‌شود.

$$f(x_1,x_2) = f_0 \frac{(a)_{x_1+x_2}(-n_1)_{x_1}(-n_2)_{x_2}}{(-b-(n_1+n_2)+1)_{x_1+x_2}} \frac{\lambda^{x_1+x_2}}{x_1!x_2!},$$

که در آن:

$$f_0 = F_1(a;-n_1;-n_2;-b-(n_1+n_2)-1;\lambda,\lambda)^{-1} = {}_2F_1(a,-(n_1+n_2);-b-(n_1+n_2)+1;\lambda)^{-1}.$$

برآورد پارامترهای مدل‌های مختلف توزیع بتا-دوجمله‌ای دومتغیره

برآورد پارامترها برای مدل‌های مختلف توزیع بتا-دوجمله‌ای دومتغیره با استفاده از روش حداکثر درست‌نمایی در نرم‌افزار R صورت گرفته است. این مدل‌ها همان‌طور که قبلاً بیان شده به خاطر مشکلات محاسباتی، در عمل به‌ندرت به کار می‌روند، به‌عنوان مثال؛ مدل سه پارامتری بی‌بی و وات (رابطهٔ (1))، شامل تابع فوق هندسی تعمیم‌یافته ${}_3F_1$ است که محاسبهٔ این

1. Gupta and Nadarajah
2. Pham-Gia and Duong

تابع در شناسهٔ 1، به لحاظ عددی ناپایا است. یک روش نسبتاً پایا برای محاسبهٔ تابع فوق هندسی تعمیم‌یافته، به کارگیری شکل انتگرالی آن است که شامل تابع فوق هندسی گاوسی ${}_2F_1$ مطابق رابطهٔ زیر می‌شود:

$$
\begin{aligned}
&{}_3F_2(a_1, a_2, a_3; b_1, b_2; 1) \\
&= \frac{\Gamma(b_2)}{\Gamma(a_3)\Gamma(b_2 - a_3)} \int_0^1 x^{a_3 - 1}(1 \\
&- x)^{b_2 - a_3 - 1} {}_2F_1(a_1, a_2; b_1; x) dx,
\end{aligned}
$$

به طوری که $b_2 > a_3$.

برای اجرای تابع فوق هندسی گاوسی از تابع ویژه کتابخانهٔ gsl استفاده شده است (هانکین[1]، 2006). همچنین برای نوشتن تابع حداکثر درست‌نمایی مدل الموـجیمینز و همکاران، از شکل انتگرالی تابع اپل (رابطهٔ زیر) استفاده گردید.

$$
\begin{aligned}
&F_2(a, b_1, b_2, c; x, y) \\
&= \frac{\Gamma(c)}{\Gamma(a)\Gamma(c - a)} \int_0^1 t^{a-1}(1 - t)^{c-a-1}(1 \\
&- xt)^{-b_1}(1 - yt)^{-b_1} dt
\end{aligned}
$$

به منظور مقایسهٔ توزیع مدل‌های مختلف وقتی که متغیرهای حاشیه‌ای دارای همبستگی‌های متفاوتی هستند، نمونه‌های مختلف در دامنه‌ای از متغیرهای حاشیه‌ای مستقل ($r = 0$) تا کاملاً همبسته ($r = 1$) انتخاب شده است.

جدول 1، فراوانی مشاهده شده نمونه‌ای از سیصد فرد، وقتی $x_1 = 0,1,2$ و $x_2 = 0,1,2$ را اختیار می‌کند، به ازای مقادیر مختلف همبستگی بین متغیرهای حاشیه‌ای نشان می‌دهد (داده‌ها شبیه‌سازی شده‌اند). جدول 2، نتایج حاصل از آزمون نیکویی برازش مربوط به نمونه‌های با همبستگی‌های حاشیه‌ای مختلف را نشان می‌دهد.

همان‌طور که مشاهده می‌شود، مدل بی‌بی و وات تنها در همبستگی $r = 0.20$ برازش خوبی داشته است. مدل داناهر و هاردی در همبستگی‌های $r = 0.20$ و $r = 0.40$ توانسته مدل مناسبی باشد و مدل الموـجیمینز و همکاران نیز در همبستگی‌های $r = 0.40$، $r = 0.60$ و $r = 0.80$ توانسته الگوی مناسبی برای رفتار این گونه داده‌ها باشد. با توجه به نتایج آزمون نیکویی برازش، مدل الموـجیمینز و همکاران در همبستگی‌های بالاتر و مدل داناهر و هاردی در همبستگی‌های پایین‌تر، برازش بهتری داشته‌اند. نکتهٔ قابل توجه این است که هیچ‌کدام از مدل‌ها نتوانسته‌اند برازش مناسبی در همبستگی‌های $r = 0$ و $r = 1$؛ یعنی وقتی همبستگی کامل بین متغیرهای حاشیه‌ای وجود دارد و یا وقتی مستقل هستند، داشته باشند.

شکل 1، معیار اطلاعاتی آکائیک (AIC) را در مدل‌های مورد مطالعه به ازای همبستگی‌های حاشیه‌ای مختلف و شکل 2 نیز، معیار اطلاعاتی بیزی (BIC) را در مدل‌های مورد مطالعه به ازای همبستگی‌های حاشیه‌ای مختلف نشان می‌دهد. روند مقادیر این دو معیار از همبستگی پایین تا همبستگی بالا مشابه بوده است.

همان‌طور که در شکل‌ها مشخص است، بر اساس این دو معیار نیز، مدل الموـجیمینز در همبستگی‌های بالا و مدل داناهر و هاردی در همبستگی‌های پایین، دارای برازش بهتری بوده‌اند.

با توجه به این اشکال، ابتدا در همبستگی‌های پایین، خط مربوط به مدل الموـجیمینز و همکاران بالاتر از مدل‌های دیگر قرار گرفته و به تدریج که به همبستگی‌های بالاتر نزدیک می‌شویم، خط مربوط به این مدل، پایین‌تر قرار می‌گیرد که حاکی از برازش بهتر این مدل در همبستگی‌های بالا است. در ادامه با ارائهٔ چند مثال واقعی صحت مطالب فوق مورد بررسی قرار می‌گیرد.

مثال‌ها

اکنون توزیع مدل‌های مورد مطالعه را برای سه مثال واقعی برازش می‌کنیم. مثال اول مربوط به مصرف یک نوع نوشیدنی برای 399 نفر در دو هفتهٔ متوالی است که این داده‌ها را آلانکو و لمنز در سال 1996 با استفاده از توزیع بتا-دوجمله‌ای دومتغیره کلاسیک و الموـجیمینز و همکاران با استفاده از مدل تعمیم‌یافتهٔ آن در سال 2011، مدل کرده‌اند (جدول 3). هر دو متغیر x_1 و x_2 مقادیر 0 تا 7 را به خود می‌گیرند، به طوری که x_1، تعداد نوبت‌های مصرف

1. Hankin

در هفتهٔ اول و X_2، تعداد نوبت‌های مصرف آن در هفتهٔ دوم می‌باشد. همبستگی بین متغیرهای حاشیه‌ای در این مثال برابر 85/9 درصد است (همبستگی بالا). مثال دوم

X_2 تعداد دندان آسیاب کشیده شدهٔ فک پایین است. هر دو متغیر X_1 و X_2 مقادیر صفر تا چهار را به خود می‌گیرند (جدول 5). میزان همبستگی بین متغیرهای حاشیه‌ای در

جدول 1. فراوانی مشاهده شده نمونه‌ای از سیصد فرد، به ازای مقادیر مختلف همبستگی بین متغیرهای حاشیه‌ای

x_2

x_1	0	1	2	کل
0	160	45	17	222
1	37	10	6	53
2	10	7	8	25
کل	207	62	31	300

r=0.20

x_2

x_1	0	1	2	کل
0	90	40	20	150
1	7	8	59	74
2	61	7	8	76
کل	158	55	87	300

r=0

x_2

x_1	0	1	2	کل
0	100	44	6	150
1	36	30	8	74
2	6	20	50	76
کل	142	94	64	300

r=0.60

x_2

x_1	0	1	2	کل
0	142	42	7	191
1	40	40	9	89
2	3	10	7	20
کل	185	92	23	300

r=0.40

x_2

x_1	0	1	2	کل
0	200	0	0	200
1	0	80	0	80
2	0	0	20	20
کل	200	80	20	300

r=1

x_2

x_1	0	1	2	کل
0	205	10	4	219
1	9	14	12	35
2	3	10	33	46
کل	217	34	49	300

r=0.80

تعداد دندان آسیب‌دیدهٔ آسیاب بزرگ دوم فک بالا (X_1) و تعداد دندان آسیب‌دیدهٔ آسیاب کوچک دوم فک بالا (X_2)، مربوط به 508 کودک 15 ساله در منطقهٔ راندرز[1] در سال 1980 است که بی‌بی و وات با مدل پیشنهادی خود در سال 2011 برازش کرده‌اند (جدول 4). همبستگی بین متغیرهای حاشیه‌ای در این مثال برابر 17/4 درصد است (همبستگی پایین). مثال سوم مربوط به تعداد دندان کشیده شدهٔ آسیاب بزرگ و کوچک فک بالا و همچنین فک پایین (به جز دندان عقل) 309 نفر از دانشجویان سال اول دانشگاه علوم و فنون دریایی خرمشهر است که در سال 1395، گردآوری شده است. X_1 تعداد دندان آسیاب کشیده شدهٔ فک بالا و

این مثال برابر 48/1 درصد است (همبستگی متوسط).

جدول 6، نتایج حاصل از آزمون نیکویی برازش مربوط به مثال‌های 1، 2 و 3 را به همراه معیار اطلاعاتی آکائیک و بیزی نشان می‌دهد. همان‌طور که مشاهده می‌شود، بر مبنای این آزمون، برای میزان بالای همبستگی بین متغیرهای حاشیه‌ای (مثال 1)، تنها مدل مناسب، مربوط به توزیع المو - جیمینز و همکاران بوده است.

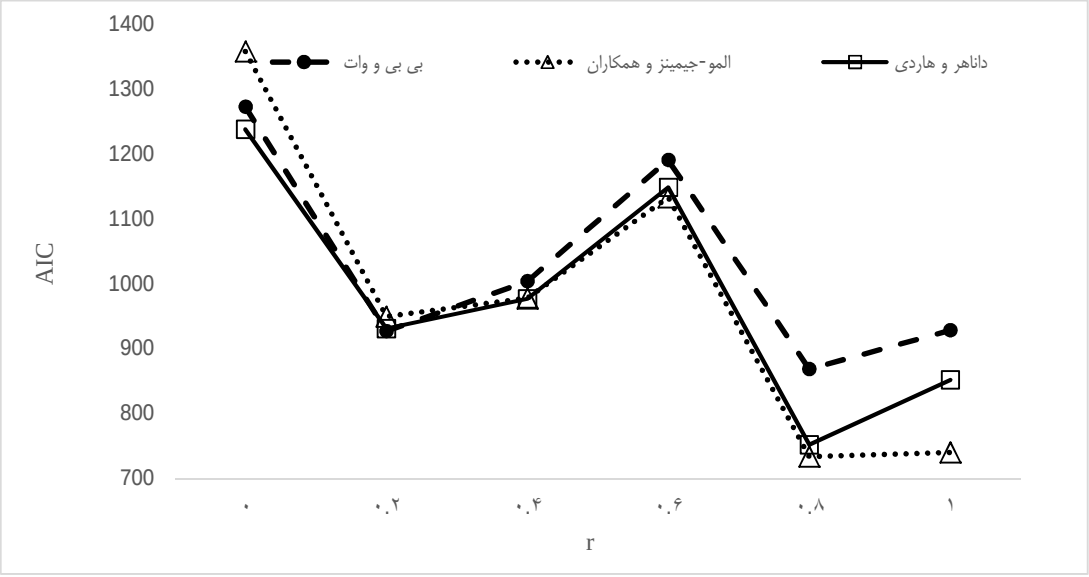
در میزان پایین همبستگی (مثال 2)، مدل‌های داناهر و هاردی و همچنین بی‌بی و وات، برازش مناسبی داشته‌اند. با توجه به اینکه معیارهای اطلاعاتی آکائیک و بیزی چیزی در مورد کیفیت مدل بیان نمی‌کنند و آزمون فرضی نیز انجام نمی‌دهند؛ بنابراین تنها برای مقایسهٔ

1. Randers

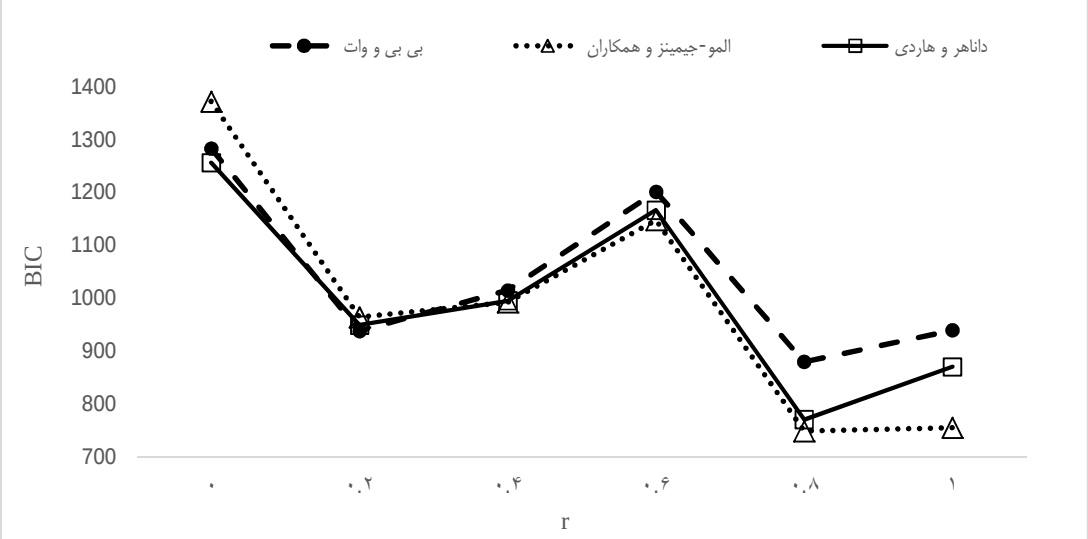
مدل‌هایی در نظر گرفته شده‌اند که از لحاظ آزمون نیکویی برازش مناسب بوده‌اند. در حقیقت این معیارها، تعادلی بین نیکویی برازش مدل و پیچیدگی مدل برقرار می‌کنند. این معیارها با در نظر گرفتن لگاریتم درست‌نمایی،

جدول 2. آزمون نیکویی برازش مربوط به نمونه‌های با همبستگی‌های حاشیه‌ای مختلف

همبستگی	مدل	لگاریتم درست‌نمایی	خی‌دو	P	مدل مناسب
$r = 0$	بی‌بی و وات	633.48	202.3	<0.001	-
	الموـجیمینز و همکاران	675.1	366.8	<0.001	-
	داناهر و هاردی	614.21	137.8	<0.001	-
$r = 0.2$	بی‌بی و وات	460.77	3.61	0.61	*
	الموـجیمینز و همکاران	471.06	23.31	<0.001	-
	داناهر و هاردی	460.73	3.58	0.31	*
$r = 0.4$	بی‌بی و وات	499.19	32.36	<0.001	-
	الموـجیمینز و همکاران	485.11	5.71	0.22	*
	داناهر و هاردی	483.68	3.43	0.33	*
$r = 0.6$	بی‌بی و وات	592.59	43.57	<0.001	-
	الموـجیمینز و همکاران	562.79	6.75	0.15	*
	داناهر و هاردی	569.46	15.22	<0.01	-
$r = 0.8$	بی‌بی و وات	431.58	206.4	<0.001	-
	الموـجیمینز و همکاران	363.27	0.21	0.9	*
	داناهر و هاردی	371.32	18.92	<0.01	-
$r = 1$	بی‌بی و وات	461.53	363	<0.001	-
	الموـجیمینز و همکاران	366.49	295.9	<0.001	-
	داناهر و هاردی	421.29	595.6	<0.001	-



شکل 1. مقایسهٔ AIC مدل‌های مورد مطالعه به ازای همبستگی‌های حاشیه‌ای مختلف



شکل 2. مقایسهٔ BIC مدل‌های مورد مطالعه به ازای همبستگی‌های حاشیه‌ای مختلف

موافق و هم‌جهت با نیکویی برازش، اما با در نظر گرفتن تعداد پارامترهای مدل، عکس نیکویی برازش عمل می‌کنند؛ زیرا افزایش تعداد پارامترها تقریباً همیشه نیکویی برازش را بهبود می‌بخشد.

در مثال 2، از بین مدل‌هایی که به لحاظ آزمون نیکویی برازش مناسب بوده‌اند، با توجه به معیار AIC،

جدول 3. فراوانی‌های مشاهده شده مصرف یک نوع نوشیدنی در دو هفتهٔ متوالی (مثال 1).

	x_2								
x_1	0	1	2	3	4	5	6	7	کل
0	26	12	6	3	0	0	0	0	47
1	13	16	15	5	4	1	0	0	54
2	3	10	15	4	7	3	1	0	43
3	0	7	10	9	9	3	1	1	40
4	0	1	6	13	9	6	3	2	40
5	0	0	1	2	13	11	9	5	41
6	0	1	0	4	3	11	9	11	39
7	0	0	1	0	4	5	20	65	95
کل	42	47	54	40	49	40	43	84	399

جدول 4. تعداد دندان آسیب‌دیدهٔ آسیاب بزرگ دوم و آسیاب کوچک دوم فک بالا (مثال 2).

	x_2			
x_1.	0	1	2	کل
0	307	91	69	467
2	3	3	6	12
کل	324	101	83	508

جدول 5. تعداد دندان کشیده شدهٔ فک بالا و فک پایین (مثال 3).

	x_2					
x_1	0	1	2	3	4	کل
0	217	33	10	1	0	261
1	16	6	6	0	0	28
2	6	2	7	2	0	17
3	0	0	2	1	0	3
4	0	0	0	0	0	0
کل	239	41	25	4	0	309

جدول 6. آزمون نیکویی برازش مربوط به مثال‌های 1، 2 و 3.

همبستگی		مدل	لگاریتم درست‌نمایی	خی‌دو	P	مدل مناسب	AIC	BIC
مثال 1	$r = 0.859$	بی‌بی و وات	1501.58	280.1	<0.001	-	3009.2	3021.1
		المو - جیمینز و همکاران	1392.30	35.24	0.41	*	2792.6	2808.6
		داناهر و هاردی	1417.40	230.4	<0.001	-	2844.8	2864.7
مثال 2	$r = 0.174$	بی‌بی و وات	622.41	4.74	0.31	*	1250.8	1263.5
		المو - جیمینز و همکاران	637.56	30.44	<0.001	-	1283.1	1300.0
		داناهر و هاردی	619.92	0.17	0.92	*	1249.8	1271.0
مثال 3	$r = 0.481$	بی‌بی و وات	394.93	16.5	0.002	-	795.9	807.1
		المو - جیمینز و همکاران	381.08	8.16	0.09	*	770.2	785.1
		داناهر و هاردی	379.75	5.36	0.15	*	769.5	788.2

مدل داناتر و هاردی و با توجه به معیار BIC، مدل بی‌بی و وات برازش بهتری را نشان داده است. این اختلاف ناشی از جریمهٔ بیشتر معیار BIC در برابر AIC به خاطر افزایش تعداد پارامترها است. نظیر این اتفاق در مثال 3 نیز مشاهده می‌شود.

نتیجه‌گیری

در این مقاله سه مدل مختلف برای توزیع بتا- دوجمله‌ای دومتغیرهٔ گسسته بررسی شد و در سه مثال چگونگی عملکرد آنها به ازای مقادیر مختلف همبستگی بین دو متغیر ارزیابی شد. نتایج حاصل از آزمون نیکویی برازش نشان می‌دهد که، مدل سه پارامتری بی‌بی و وات (2011) در مقایسه با مدل‌های دیگر عملکرد ضعیفی داشته و مناسب نیست. مدل المو - جیمینز و همکاران (2011)، برای مقادیر بالای همبستگی بین متغیرهای حاشیه‌ای، برازش بهتری نسبت به مدل‌های دیگر دارد و در مقادیر پایین همبستگی بین متغیرهای حاشیه‌ای، مدل داناهر و هاردی (2005) مناسب‌تر است.

منابع

[1] Alanko, T. & Lemmens, P.H. (1996). Response effects in consumption surveys: an application of the beta-binomial model to self-reported drinking frequencies. Journal of Official Statistics, **12**(3), 253-273.

[2] Appell, P. (1880). Sur les séries hypergéometriques de deux variables et sur des équations différentielles linéaires aux dérivées partielles. Comptes Rendus, **90**, 296-298.

[3] Bibby, B.M. & Væth, M. (2011). The two-dimensional beta binomial distribution. Statistics & Probability Letters, **81**(7), 884-891.

[4] Danaher, P.J. & Hardie, B.G.S. (2005). Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. The American Statistician, **59**(4), 282-286.

[5] Fisher, R.A. Statistical methods for research workers. Genesis Publishing Pvt Ltd 1925.

[6] Gelman, E. & Sichel, H.S. (1987). Library book circulation and the beta-binomial distribution. Journal of the American Society for Information Science, **38**(1), 5-12.

[7] Gupta, A.K. & Nadarajah, S. (Eds.). (2004). Handbook of beta distribution and its applications. CRC press.

[8] Hankin, R.K.S. (2006). Special functions in R: introducing the **gsl** package. R News **6**(4), 24-26.

[9] Ishii, G. & Hayakawa, R. (1960). On the compound binomial distribution. Annals of the Institute of Statistical Mathematics, **12**(1), 69-80.

[10] Johnson, N.L., Kemp, A.W. & Kotz, S. (2005). Univariate discrete distributions. John Wiley & Sons.

[11] Jones, M.C. (2001). Multivariate t and beta distributions associated with the multivariate F distribution. Metrika, **54**(3), 215-231.

[12] Olkin, I. & Liu, R. (2003). A bivariate beta distribution. Statistics & Probability Letters, **62**(4), 407-412.

[13] Olmo-Jiménez, M. J., Martínez-Rodríguez, A. M., Conde-Sánchez, A., & Rodríguez-Avi, J. (2001). A generalization of the bivariate Beta-Binomial distribution. Journal of Statistical Planning and Inference, **141**(7), 2303-2311.

[14] Pham-Gia, T. & Duong, Q.P. (1989). The generalized beta-and F-distributions in statistical modelling. Mathematical and Computer Modelling, **12**(12), 1613-1625.

[15] Sarmanov, O.V. (1966). Generalized normal correlation and two-dimensional Fréchet. In Soviet Mathematics. Doklady, Vol. **25**, pp. 1207-1222.

[16] Skellam, J.G. A. (1948). Probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society. Series B (Methodological), **10**(2), 257-261.

[17] Ting Lee, M.L. (1996). Properties and applications of the Sarmanov family of bivariate distributions. Communica-tions in Statistics-Theory and Methods, **25**(6), 1207-1222.