بررسی تجربی گرایش توزیع دوجمله‌ای به پواسن

رضا مؤمنی[۱] یدالله واقعی[۲]

چکیده

توزیع‌های دوجمله‌ای و پواسن از جمله توزیع‌هایی هستند که در آمار و احتمال کاربردهای فراوانی دارند. اگر چه این دو توزیع تعاریف و فرمولهای مختلفی دارند ولی تحت شرایط خاصی به هم نزدیک می‌شوند. به سادگی می‌توان نشان داد چنانچه در توزیع دوجمله‌ای حجم نمونه(n) به بینهایت و احتمال موفقیت (p) به صفر میل کند و np مقدار ثابتی باقی بماند، توزیع دوجمله‌ای به توزیع پواسن میل می کند و در این صورت است که می‌توان بجای توزیع دوجمله‌ای از توزیع پواسن استفاده کرد.

با توجه به اینکه در عمل غیر ممکن است حجم نمونه به بینهایت و احتمال موفقیت به صفر برسد در این مقاله به دنبال شرایط عملی هستیم که با اندکی خطا(تقریب) بتوان بجای توزیع دوجمله‌ای از توزیع پواسن استفاده نمود. برای این منظور اختلاف دو توزیع به ازای مقادیر مختلف x در قالب ملاکی که «جذر میانگین مربع اختلافات» نامیده شده برای دامنه گسترده‌ای از n و p ها محاسبه شده و در نهایت مقادیری از n و p که به ازای آنها تقریب دوجمله‌ای به پواسن مناسب است، مشخص شده‌اند.

واژه های کلیدی : پواسن، دوجمله‌ای، قضایای حدی.

۱. مقدمه

گرایش توزیع‌های احتمال بـه یکدیگر یکی از مباحث جالب در بحث نظریه احتمال است و از طرفی کار محاسبات و استنباط آماری را ساده یا ممکن می‌سازد. توزیع‌های دوجمله‌ای و پواسن دو توزیع پایه و بسیار مهم آمار هستند که معمولاً به دانشجویان همه رشته‌هایی که درس آمار و احتمال دارند آموخته می‌شوند. گرایش توزیع دوجمله‌ای به پواسن یکی از مباحث نظری و در عین حال دارای کاربرد در آمار و احتمال می باشد که سبب سادگی محاسبات می‌شود.

طبق یک تعریف ساده، اگر یک آزمایش برنولی در شرایط یکسان به طور مستقل از هم n بار تکرار شود در این صورت متغیر تصادفی X : تعداد پیروزیها در این n آزمایش، دارای توزیع دوجمله‌ای است. همچنین اگر از جامعه‌ای که شامل دو گروه (مثلاً دارای بیماری سرماخوردگی و فاقد بیماری سرماخوردگی) باشد، یک نمونه تصادفی n تایی به روش با جایگذاری گرفته شود، در این صورت تعداد افراد در هر یک از دو گروه یک متغیر تصادفی با توزیع دوجمله‌ای خواهد بود.

---

[۱] کارشناس آمار

[۲] گروه آمار، دانشگاه بیرجند

تعداد رخدادها یا موفقیتها در یک فاصله زمانی یا ناحیه مکانی را گاهی می‌توان با توزیع پواسن مدل‌سازی کرد. یکی از فرضهای پایه‌ای که توزیع پواسن بر مبنای آن ساخته می‌شود این است که در فواصل زمانی(مکانی) کوچک احتمال یک رخداد متناسب با طول فاصله باشد.همچنین تعداد رخدادها در فواصل زمانی(مکانی) مجزا باید از هم مستقل باشند. به عنوان مثال « تعداد غلطهای چاپی در یک صفحه از کتاب » ، «تعداد مشتریان اداره پست در یک روز معین» و «تعداد یک نوع میکروب در یک سانتیمتر مکعب خون» مصادیقی از توزیع پواسن هستند. وقتی در توزیع دوجمله‌ای n بزرگ باشد محاسبه احتمالهای دوجمله‌ای به دلیل وجود ضریب $n!/(x!(n-x)!)$ در تابع احتمال مشکل می‌باشد. در این صورت تحت شرایط مشخصی می‌توان بجای توزیع دوجمله‌ای از توزیع پواسن استفاده نمود. طبق یک قضیه مشهور در آمار ریاضی چنانچه در توزیع دوجمله‌ای $n \to \infty$ و $p \to ٠$ در عین حال $\lambda = np$ ثابت بماند، توزیع دوجمله‌ای به پواسن میل می‌کند. این قضیه را مستقیماً با حد گرفتن از تابع احتمال دوجمله‌ای می‌توان اثبات کرد(مثلاً، عمیدی، ۱۳۸۰، ص ۲۰۱). از روش تابع مولد گشتاور نیز می‌توان گرایش توزیع دوجمله‌ای به پواسن را نشان داد. اما در عمل بعید است که n به بینهایت و p به صفر برسد. لذا یافتن شرایط عملی که تحت این شرایط توزیع پواسن تقریب نسبتاً خوبی برای توزیع دوجمله‌ای باشد از اهمیت خاصی برخوردار است. بعضی آمارشناسان گفته‌اند که وقتی $n \geq ۲۰$ و $p \leq ٠/٠۵$ باشد توزیع پواسن تقریب خوبی برای احتمالهای دوجمله‌ای است و وقتی $n \geq ۱٠٠$ و $np < ۱٠$ تقریب عالی است (مثلاً، عمیدی، ۱۳۸۰). در این مقاله می‌خواهیم موضوع را موشکافی کرده و با مقایسه احتمالات توزیع‌های دوجمله‌ای و پواسن، مقادیر مناسب p و n را که به ازای آنها تقریب مناسب است، پیدا کنیم.

فرض کنید $f_۱(x)$ و $f_۲(x)$ به ترتیب نشان‌دهنده احتمالات توزیع‌های دوجمله‌ای با پارامترهای n و p و پواسن با پارامتر $\lambda = np$ باشد، یعنی

$$f_۱(x) = \binom{n}{x} p^x (۱-p)^{n-x} \qquad , x = ٠,۱,۲,۳,\ldots,n$$

$$f_۲(x) = \frac{e^{-np}(np)^x}{x!} \qquad , x = ٠,۱,۲,۳,\ldots,n,n+۱,\ldots .$$

به عنوان یک ملاک، جذر میانگین مربع اختلافات[۳] دو توزیع را به صورت زیر تعریف می‌کنیم.

$$rmsd = \sqrt{\frac{۱}{n+۱}\sum_{x=٠}^{n}(f_۲(x) - f_۱(x))^۲}$$

این ملاک به نوعی میانگین اختلافات دو توزیع به ازای مقادیر مختلف x را نشان می‌دهد. هر چه اختلاف دو توزیع بیشتر باشد مقدار عددی این ملاک بزرگتر خواهد بود(و برعکس). بررسی‌های مقدماتی مشخص کرد که چنانچه $rmsd < ٠/٠٠۵$ باشد می‌توان از توزیع پواسن به عنوان تقریب توزیع دوجمله‌ای استفاده نمود، زیرا در این صورت اختلاف احتمالات دو توزیع در نقاط مختلف به طور متوسط از ۰/۰۱ کمتر خواهد بود (مؤمنی، ۱۳۸۲). در این مقاله می‌خواهیم با استفاده از ملاک فوق ببینیم به ازای چه n و p هایی اختلاف دو توزیع ناچیز است، یعنی p,n چه مقادیری باشند که بتوان توزیع پواسن را به عنوان تقریبی از توزیع دوجمله‌ای مورد استفاده قرار داد.

۲. بررسی اختلاف دو توزیع به ازای p,n های مختلف

به منظور یافتن مقادیری از p,n که تقریب مناسبی را برای استفاده از توزیع پواسن بجای توزیع دوجمله‌ای فراهم می‌کنند، نخست برای مقادیر مختلف n و p :

$$(n = ۵,۱٠,۱۵,\ldots,۱٠٠; \quad p = ٠/٠۵,٠/۱,٠/۱۵,\ldots,٠/۴۵,٠/۵)$$

احتمالات توزیع دوجمله‌ای و پواسن با $\lambda = np$ محاسبه شد. سپس برای هر n و p مقدار rmsd که نشان‌دهنده میزان اختلاف دو توزیع است به دست آمد. جدول ۱ مقادیر rmsd را که با استفاده از یک برنامه SPLUS محاسبه شده است. به ازای مقادیر مختلف n و p نشان می‌دهد(مؤمنی، ۱۳۸۲).

در این جدول خطوط پله مانندی رسم شده است که مقادیر بالای پله‌ها از ۰/۰۰۵ بیشتر و زیر آنها از ۰/۰۰۵ کمتر است. مقادیری که زیر پله‌ها هستند، مقادیری از p,n را مشخص می‌کنند که تقریب توزیع پواسن برای توزیع دوجمله‌ای مناسب است یعنی به ازای این n و pها می‌توان از توزیع پواسن بجای توزیع دوجمله‌ای استفاده نمود. به عنوان

[۳] Root of mean squrae differences

مثال به ازای $p=\cdot/\cdot ۵$ و $n=۵$ مقدار rmsd برابر ۰/۰۰۴۴ می‌باشد که نشان‌دهنده تقریب خوب توزیع پواسن برای دوجمله‌ای است! برای بررسی بیشتر این موضوع در شکل۱ نمودار احتمالات توزیع دوجمله‌ای (باعلامت ٥) و پواسن (با علامت *) به ازای $n=۵$ و $p=\cdot/\cdot ۵$ و $\lambda=\cdot/۲۵$ نشان داده شده است.

با توجه به شکل۱ مشاهده می‌شود که توابع احتمال توزیع دوجمله‌ای و پواسن به ازای $n=۵$، $p=\cdot/\cdot ۵$ و $\lambda=\cdot/۲۵$ در نقاط ۵،۴،۳،۲،۱،۰ تقریباً با هم برابرند و اختلاف چندانی ندارند. لذا برای $n=۵$ هنگامیکه $p=\cdot/\cdot ۵$ یا $(p<\cdot/\cdot ۵)$ تقریب مناسب است.

بررسی سایر مقادیر rmsd زیر ستون $p=\cdot/\cdot ۵$ جدول۱ نشان می‌دهد که به ازای همه مقادیر $n=۵,۱۰,۱۵,...,۱۰۰$، مقدار rmsd کوچک است و علاوه بر این با افزایش n از ۵ به سمت ۱۰۰ مقدار rmsd از ۰/۰۰۴۴ به ۰/۰۰۰۸ کاهش می‌یابد، به عبارت دیگر با افزایش n اختلاف دو توزیع کم شده و تقریب مناسب‌تر می‌شود.

به طور خلاصه می توان چنین بیان کرد که هرگاه $p=\cdot/\cdot ۵$ (یا حتی $p<\cdot/\cdot ۵$) باشد به ازای $n\geq ۵$ تقریب دوجمله‌ای به پواسن نسبتاً خوب است. جای بسی تعجب است هنگامی‌که $p=\cdot/\cdot ۵$ (یا کمتر از آن) باشد حتی به ازای مقادیر کوچک n مثل ۵ و ۱۰ نیز تقریب مناسب است، زیرا بر طبق شرایط عملی که در اغلب کتابهای آمار و احتمال بیان شده می‌بایست $p<\cdot/\cdot ۵$ و $n\geq ۲۰$ باشد، حال آنکه نتیجه بررسی ما نشان می‌دهد که شرط $n\geq ۲۰$ خیلی قوی است و $n\geq ۵$ کفایت می‌کند.

اگر چه تئوری بیان می‌کند که باید p به سمت صفر میل کند ولی برای pهای بزرگتر از ۰/۰۵ نیز مشاهده می‌شود که با افزایش n تقریب مناسب می‌شود، به عنوان مثال با $p=\cdot/۱$ برای $n\geq ۲۰$ و با $p=\cdot/۱۵$ برای $n\geq ۳۰$ تقریب مناسب است، حتی برای $p=\cdot/۴$ که بطور معنی‌داری ار صفر دور است مشاهده می‌شود که برای $n=۱۰۰$ تقریب مناسب است. لذا مرز یکنواختی را نمی‌توان بر حسب n و p برای نزدیکی دو توزیع مشخص کرد، زیرا نزدیکی دو توزیع بستگی به تغییرات توأم n و p دارد.

۳. بررسی نمودار تراز اختلافات دو توزیع

جدول ۱ مقادیر rmsd را در یک شبکه شطرنجی(ماتریسی) نشان می‌دهند. هر مقدار داخل این جدول بیانگر میزان اختلاف دو توزیع دوجمله‌ای و پواسن به ازای n و p ی متناظر با سطر و ستون مقدار موردنظر است. با توجه به اینکه اغلب تفسیر نمودارها و شکل‌ها ساده‌تر از جداول عددی است، در شکل ۲ نمودار تراز[۴] مقادیر rmsd با SPLUS رسم شده است. در این نمودار محورهای افقی و عمودی مشخص‌کننده مقادیر n و p (متناظر با جدول۱) هستند و منحنی‌های مشخص‌شده روی آن، که اصطلاحا منحنی تراز نامیده می‌شود، مشخص‌کننده نقاطی از n و p هستند که دارای rmsd (جذر میانگین مربع اختلافات) یکسان می‌باشند و این مقدار (rmsd) در حاشیه این منحنی‌ها چاپ شده است. به عنوان مثال منحنیی که کنار آن ۰/۰۰۲ نوشته شده مشخص‌کننده مقادیر مختلف n و p است که جذر میانگین مربع اختلافات دو توزیع به ازای آنها ۰/۰۰۲ می‌باشد. روند این منحنی و سایر منحنی‌های تراز نشان می‌دهد که هر چه p افزایش می‌یابد به مقدار n بزرگتری برای رسیدن به یک rmsd ثابت (به عنوان مثال ۰/۰۰۲) نیاز داریم. مثلا هنگامی‌که p برابر ۰/۰۵ باشد حدودا n باید ۳۵ باشد تا rmsd حدود ۰/۰۰۲ باشد ولی هنگامی‌که p برابر ۰/۱ باشد حدودا n باید ۶۰ باشد تا rmsd حدود ۰/۰۰۲ باشد.

منحنی تراز در شکل ۲ از پایین به بالا مقادیر افزایشی دارند، یعنی با افزایش p مقدار rmsd نیز طبق مقادیر مشخص شده در کنار منحنی‌ها افزایش می‌یابند. به کمک این منحنی‌ها در صورت نیاز می‌توان کیفیت تقریب را بر حسب مقادیر rmsd مندرج در حاشیه منحنی‌ها به سطوح خیلی خوب، مناسب، نامناسب رده‌بندی کرد. در اینجا با توجه به مباحث قبلی، صرفا سطوح مناسب و نامناسب را مشخص می‌کنیم. با توجه به اینکه مقدار ۰/۰۰۵ در rmsd را به عنوان مرز مناسب بودن تقریب دوجمله‌ای به پواسن پذیرفته‌ایم، منحنی تراز ۰/۰۰۵ و نقاط زیر آن (که متناظر با منحنی‌های تراز کمتر از ۰/۰۰۵ هستند) مشخص‌کننده مقادیری از n و p هستند که تقریب به ازای آنها مناسب است. به عبارت دیگر زیر منحنی تراز ۰/۰۰۵ تقریب مناسب و بالای آن تقریب نامناسب می‌باشد. این منحنی مشخص می‌کند که به ازای $p>\cdot/۴$ برای هر کمتر از ۱۰۰ تقریب نامناسب است.

---

[۴] Conter plot

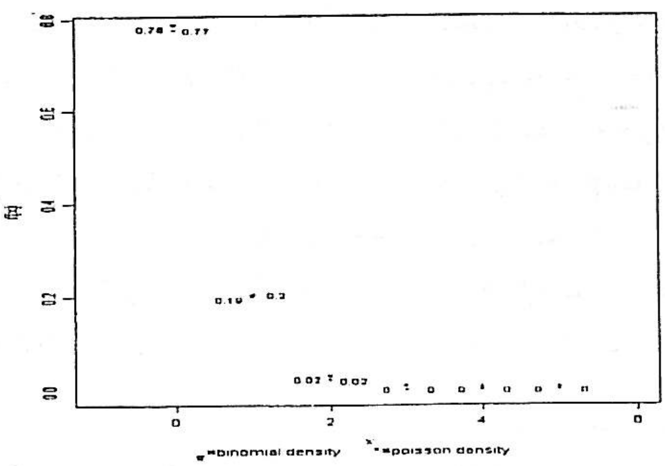
با توجه به منحنی تراز به ازای هر $\cdot/۴ > p > \cdot$ می‌توان مقدار n متناظر را که به ازای آن تقریب مناسب است مشخص نمود. برای این منظور کافی است یک خط افقی از p به منحنی تراز $rmsd = \cdot/\cdot\cdot۵$ رسم شود و از آنجا خطی به محور n عمود شود؛ مثلاً اگر $p = \cdot/۲۱$ باشد به ازای $n \geq ۴۱$ تقریب دوجمله‌ای به پواسن مناسب خواهد بود.

نتایج به دست آمده از این نمودار مطابق با نتایج بررسی جدول۱ است با این تفاوت که p بطور گسسته با تغییرات ۰/۰۵ و n با تغییرات گسسته ۵ واحدی مدرج شده است، در حالی که در شکل۲ تغییرات n و p بطور پیوسته قابل ملاحظه است. علاوه بر این به دلیل ماهیت نموداری بودن، تفسیر این نمودار راحت‌تر از جدول می‌باشد.

٤. بحث و نتیجه‌گیری

در این مقاله به منظور بررسی عملی گرایش توزیع دوجمله‌ای به پواسن، میانگین اختلاف دو توزیع به ازای مقادیر مختلف x، در قالب ملاک «جذر میانگین مربع اختلافات» یا rmsd برای دامنه وسیعی از n و pها محاسبه شد و نتایج به صورت جدول و نمودار نمایش داده شد. نتایج بررسی مقادیر rmsd نشان داد چنانچه p کمتر یا مساوی ۰/۰۵ باشد برای $n \geq ۵$ می‌توان از تقریب دوجمله‌ای به پواسن استفاده نمود. همچنین اگر p بین ۰/۰۵ تا ۰/۱ باشد برای nهای بزرگتر یا مساوی ۲۰ تقریب مناسب خواهد بود. اگر p بین ۰/۱ تا ۰/۱۵ باشد می‌بایست n بزرگتر یا مساوی ۳۰ و اگر p بین ۰/۱۵ تا ۰/۲ باشد می‌بایست n بزرگتر یا مساوی ٤۰ باشد، حتی اگر p بزرگتر از ۰/۲ (حداکثر تا ۰/٤) باشد، مقادیری از n وجود دارد که به ازای آنها تقریب توزیع دوجمله‌ای به پواسن مناسب است. بنابراین ملاحظه می‌شود که شرط p کمتر از ۰/۰۵ که در کتابهای آماری به عنوان شرط عملی برای گرایش دوجمله‌ای به پواسن ذکر شده، ضروری نیست و اگر p کمتر از ۰/۰۵ باشد شرط $n \geq ۲۰$ را می‌توان محدودتر نمود. در شکل۲ نموداری رسم شده که به ازای مقادیر مختلف n و p میزان اختلاف دو توزیع (با استفاده از ملاک rmsd) و در نتیجه میزان مناسب بودن استفاده از توزیع پواسن بجای دوجمله‌ای را نشان می‌دهد و بطور مفصل مشخص می‌کند که به ازای چه مقادیری از n و p می‌توان از توزیع پواسن به عنوان تقریب دوجمله‌ای استفاده نمود و میزان دقت تقریب نیز با استفاده از منحنی‌های تراز مشخص می‌شود که مناسب است خواننده برای مشاهده جزئیات بیشتر به آن مراجعه نماید.

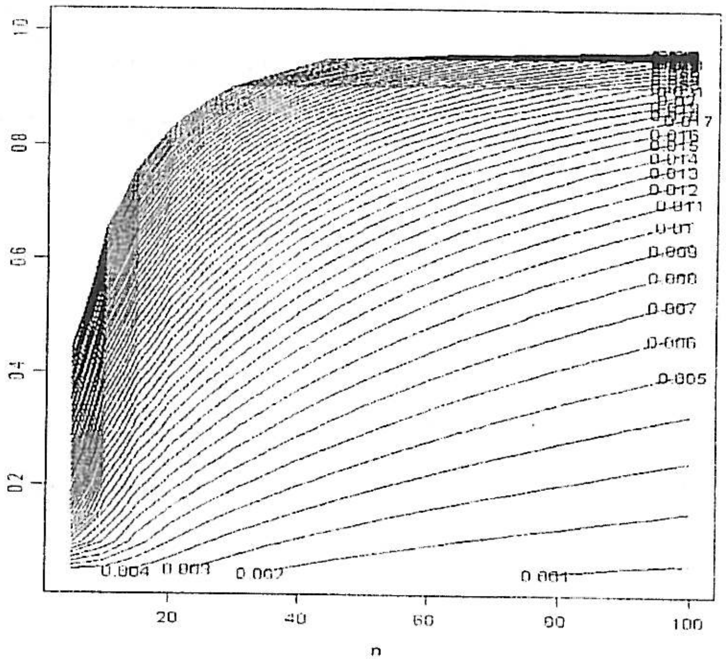
شکل ۱- نمودار احتمالات توزیع‌های دوجمله‌ای(.) و پواسن(*) برای $n = ۵$ و $p = \cdot/\cdot۵$ و $\lambda = \cdot/۲۵$



جدول (۱) مقادیر rmsd به ازای n و p های مختلف

	p= 0.05	p= 0.1	p= 0.15	p= 0.2	p= 0.25	p= 0.3	p= 0.35	p= 0.4	p= 0.45	p= 0.5
n= 5	0.0044	0.0123	0.0197	0.0258	0.0308	0.0354	0.0403	0.0455	0.0511	0.0572
n= 10	0.0043	0.0090	0.0120	0.0146	0.0174	0.0205	0.0238	0.0273	0.0311	0.0353
n= 15	0.0038	0.0064	0.0084	0.0105	0.0127	0.0151	0.0176	0.0202	0.0231	0.0263
n= 20	0.0031	0.0050	0.0067	0.0084	0.0103	0.0122	0.0142	0.0164	0.0187	0.0213
n= 25	0.0026	0.0041	0.0056	0.0071	0.0087	0.0103	0.0120	0.0139	0.0159	0.0181
n= 30	0.0023	0.0035	0.0049	0.0062	0.0076	0.0090	0.0105	0.0121	0.0139	0.0158
n= 35	0.0020	0.0031	0.0043	0.0055	0.0067	0.0080	0.0094	0.0108	0.0124	0.0141
n= 40	0.0017	0.0028	0.0039	0.0050	0.0061	0.0072	0.0085	0.0098	0.0112	0.0128
n= 45	0.0016	0.0026	0.0036	0.0046	0.0056	0.0066	0.0078	0.0090	0.0103	0.0117
n= 50	0.0014	0.0024	0.0033	0.0042	0.0051	0.0061	0.0072	0.0083	0.0095	0.0108
n= 55	0.0013	0.0022	0.0031	0.0039	0.0048	0.0057	0.0067	0.0077	0.0088	0.0101
n= 60	0.0012	0.0021	0.0029	0.0037	0.0045	0.0053	0.0063	0.0072	0.0083	0.0094
n= 65	0.0012	0.0019	0.0027	0.0035	0.0042	0.0050	0.0059	0.0068	0.0078	0.0089
n= 70	0.0011	0.0018	0.0025	0.0033	0.0040	0.0048	0.0056	0.0064	0.0074	0.0084
n= 75	0.0010	0.0017	0.0024	0.0031	0.0038	0.0045	0.0053	0.0061	0.0070	0.0080
n= 80	0.0010	0.0017	0.0023	0.0030	0.0036	0.0043	0.0050	0.0058	0.0067	0.0076
n= 85	0.0009	0.0016	0.0022	0.0028	0.0035	0.0041	0.0048	0.0056	0.0064	0.0073
n= 90	0.0009	0.0015	0.0021	0.0027	0.0033	0.0039	0.0046	0.0053	0.0061	0.0070
n= 95	0.0009	0.0014	0.0020	0.0026	0.0032	0.0038	0.0044	0.0051	0.0059	0.0067
n= 100	0.0008	0.0014	0.0019	0.0025	0.0031	0.0036	0.0043	0.0049	0.0057	0.0064

شکل (۲)



مراجع

[۱] بهبودیان، جواد؛ ۱۳۷۷، *آمار و احتمال مقدماتی*، مؤسسه چاپ و انتشارات آستان قدس رضوی.

[۲] راس، شلدون ام.؛ ۱۳۷۶، *نخستین درس احتمال*، ترجمه دکتر حسنعلی آذرنوش، ابولقاسم بزرگنیا، علی مشکانی و حسینعلی نیرومند، مؤسسه چاپ و انتشارات دانشگاه فردوسی مشهد.

[۳] رضایی، عبدالمجید؛ ۱۳۸۰، *مفاهیم آمار و احتمالات*، نشر مشهد، چاپ دوم.

[٤] فروند، جان و والپول، رانلد؛ ۱۳۷۱، آمار ریاضی، ترجمه دکتر علی عمیدی، چاپ دوم، مرکز نشر دانشگاهی.

[٥] نتر، جان و واسرمن، ویلیام؛ ۱۳۷۳، *آمار کاربردی*، ترجمه دکتر علی عمیدی، چاپ اول، مرکز نشر دانشگاهی.

[٦] مؤمنی، رضا؛ ۱۳۸۲، *بررسی تجربی گرایش توزیع دوجمله‌ای به پواسن*، پروژه کارشناسی آمار، دانشگاه بیرجند.

[۷] نعمت‌اللهی، نادر؛ ۱۳۸۱، *آمار و احتمال مهندسی*، انتشارات دلفک.