# مقایسه‌ی برآوردگرهای آنتروپی طرح‌های نمونه‌گیری

فهیمه مسیحی بیدگلی
گروه آمار، دانشکده ریاضی، دانشگاه صنعتی اصفهان

**چکیده**

یکی از روش‌های مقایسه‌ی طرح‌های نمونه‌گیری محاسبه‌ی میزان آنتروپی آن‌ها است. آنتروپی یک طرح نمونه‌گیری اندازه‌ای از تصادفی بودن و گستردگی آن را نشان می‌دهد. محاسبه‌ی آنتروپی از طریق تعریف، به‌علت زیاد بودن نمونه‌های ممکن کاری بسیار وقت‌گیر و گاهی غیر عملی است. هدف این مقاله، یافتن برآوردگرهای مناسب آنتروپی برای طرح‌های نمونه‌گیری است. استفاده از یک برآوردگر، به طرح و موقعیت آن بستگی دارد. بعضی از برآوردگرها تنها برای طرح‌هایی که دارای تابع احتمال مشخص هستند، مناسب‌اند و بعضی از آن‌ها را حتی در حالتی که تابع احتمال طرح نمونه‌گیری مشخص نباشد، می‌توان مورد استفاده قرار داد. مقایسه‌ی این برآوردگرها با استفاده از چند مثال شبیه‌سازی صورت گرفته است و نتایج نشان می‌دهد، در صورت مشخص نبودن تابع احتمال، استفاده از برآوردگر آنتروپی خام با تصحیح میزان اریبی مناسب‌تر خواهد بود.

**واژه‌های کلیدی:** آنتروپی، طرح نمونه‌گیری، نمونه‌گیری پواسون شرطی تعدیل‌شده، نمونه‌گیری پارتو.

## ۱ مقدمه

مفهوم آنتروپی در علومی که با پدیده‌های تصادفی در ارتباط هستند، کاربرد دارد و در رشته‌های گوناگون علمی معانی متفاوتی می‌یابد. از جمله معانی که برای آنتروپی به‌کار می‌رود، آشفتگی، بی‌نظمی، عدم قطعیت و میزان تصادفی بودن یک پیشامد است. شانون [۷] معیار آنتروپی را با علم آمار و احتمال پیوند داد و آنتروپی که تا قبل از آن تنها در علم فیزیک استفاده می‌شد، در سایر علوم جایگاه خود را پیدا کرد. بدین علت آنتروپی به‌طور قراردادی با آنتروپی شانون نمایش داده شد.

آنتروپی یک طرح نمونه‌گیری میزان گستردگی و تصادفی بودن یک طرح را نشان می‌دهد. هاجک [۲] برای اولین بار آنتروپی طرح‌های نمونه‌گیری را بررسی کرد و نشان داد در کلاسی از طرح‌های نمونه‌گیری با احتمال

برابر و با اندازه‌ی نمونه‌ی ثابت، طرح نمونه‌گیری تصادفی ساده بدون جایگذاری دارای حداکثر میزان آنتروپی است. او همچنین نشان داد در کلاس طرح‌های نمونه‌گیری با احتمال نابرابر و با بردار احتمال شمول مرتبه‌ی اول یکسان و با اندازه‌ی نمونه‌ی ثابت، طرح نمونه‌گیری پواسون شرطی تعدیل‌شده[۱] حداکثر آنتروپی را دارد. گرافسترم [۱] نیز آنتروپی طرح‌های نمونه‌گیری با احتمال نابرابر را مورد بررسی قرار داد و به این نتیجه رسید که چندین طرح در داشتن حداکثر آنتروپی به یکدیگر نزدیک هستند. همچنین او چندین برآوردگر آنتروپی را در موقعیت‌های مختلف معرفی کرد که بعضی از این برآوردگرها در این مقاله با یکدیگر مقایسه می‌شوند. ابتدا در بخش ۲ آنتروپی و کاربرد آن در طرح نمونه‌گیری بیان و سپس در بخش ۳ برآوردگرهای مختلف آنتروپی طرح نمونه‌گیری ارائه می‌شود. در بخش ۴ با بررسی دو جامعه و شبیه‌سازی نمونه‌های طرح‌های مختلف، برآوردگرهای آنتروپی مقایسه شده و در انتها نتایج ارائه می‌شود.

## ۲ آنتروپی

اگر $X$ متغیر تصادفی باشد که یکی از مقادیر $x_1, \dots, x_n$ را با احتمال‌های $p_1, \dots, p_n$ انتخاب کند، $-\log p_i$ نشان‌دهنده‌ی میزان تعجب حاصل از آن است که $X$ مقدار $x_i$ را اختیار کند، در نتیجه امید ریاضی میزان تعجب از اطلاع در مورد متغیر تصادفی $X$ برابر است با

$$H(X) = E[-\log Pr(X)] = -\sum_{i=1}^{n} p_i \log p_i.$$

کمیت $H(X)$ در نظریه‌ی اطلاعات، به‌عنوان آنتروپی متغیر تصادفی $X$ شناخته می‌شود. می‌توان $H(X)$ را به‌عنوان میزان عدم قطعیتی که برای مقدار $X$ وجود دارد تفسیر کرد. آنتروپی با تعریف فوق آنتروپی شانون نامیده می‌شود. محققین زیادی سعی در توسعه‌ی آنتروپی داشته‌اند. از جمله، رنی [۴] و تی‌سالیس [۹] آنتروپی‌های دیگری را معرفی کرده‌اند. در ادامه به بعضی از ویژگی‌های آنتروپی اشاره می‌شود.

۱-آنتروپی تابعی از احتمال‌های $p_1, \dots, p_n$ است و به مقادیر $x_i (i = 1, 2, \dots, n)$ بستگی ندارد.

۲- آنتروپی تابعی پیوسته است و هر تغییر بسیار اندک در مقدارهای احتمال، باعث تغییر بسیار کوچک در میزان آنتروپی می‌شود.

۳- تابع آنتروپی نسبت به هر یک از مؤلفه‌های خود متقارن است و اندازه‌ی آنتروپی با تغییر در ترتیب برآمدهای $x_i$ تغییر نخواهد کرد.

۴- اگر همه‌ی برآمدهای $x_i$ هم‌شانس باشند، اندازه‌ی آنتروپی بیشینه می‌شود و در این حالت آنتروپی با افزایش تعداد برآمدها، افزایش می‌یابد.

۵- آنتروپی توأم دو سیستم مستقل برابر با جمع آنتروپی هر یک از آن‌ها خواهد بود:

$$H(X, Y) = H(X) + H(Y).$$

در صورتی که دو سیستم مستقل نباشند، آنتروپی توأم عبارت است از

$$H(X, Y) = H(X) + H(Y|X),$$

که در آن $H(Y|X)$ متوسط عدم قطعیتی است که بعد از دریافت اطلاعات $X$، برای $Y$ باقی می‌ماند و بر اساس رابطه‌ی زیر به‌دست می‌آید.

$$H(Y|X) = \sum_{i=1}^{n} H(Y|X = x_i)\ Pr(X = x_i),$$

[۱] Adjusted Conditional Poisson

که در آن

$$H(Y|X=x_i) = -\sum_j Pr(y_j|x_i)\log Pr(y_j|x_i).$$

## ۱.۲ آنتروپی طرح نمونه‌گیری

همان‌طور که اشاره شد، آنتروپی یک طرح نمونه‌گیری اندازه‌ای از میزان تصادفی بودن طرح را نشان می‌دهد. هنگامی که طرح نمونه‌گیری آنتروپی بالایی دارد، یک میزان زیاد از عدم قطعیت یا میزان بالایی از تعجب در نمونه‌ی انتخابی وجود دارد. به‌عبارت دیگر، در این حالت میزان بالایی از تصادفی بودن وجود دارد. اگر اطلاعات اضافی در مورد جامعه و ویژگی‌های آن موجود باشد، این اطلاعات می‌تواند برای انتخاب طرح به‌کار رود. در صورت عدم دسترسی به این اطلاعات، طرح با آنتروپی بالا به‌علت تصادفی بودن زیاد، انتخاب مناسبی برای انجام نمونه‌گیری خواهد بود.

یک طرح نمونه‌گیری یک توزیع گسسته روی مجموعه‌ای از نمونه‌های ممکن $\mathbf{x}_k$ $(k \in \mathcal{Q})$ است. به $\mathcal{Q}$ تکیه‌گاه طرح گفته می‌شود و مجموعه‌ای از تمام نمونه‌های ممکن برای یک طرح نمونه‌گیری است. نمونه‌ی $\mathbf{x}_k$ می‌تواند به‌عنوان برداری از نشانگر عضویت مطرح شود. بنابراین $\mathbf{x}_k \in \{۰,۱\}^N$ که $N$ اندازه‌ی جامعه، ۱ نشان‌دهنده‌ی عضویت واحد جامعه در نمونه و ۰ عدم عضویت آن واحد است. احتمال به‌دست آمدن نمونه‌ی $\mathbf{x}_k$ با $p(\mathbf{x}_k)$ نمایش داده می‌شود و به عنوان طرح نمونه‌گیری شناخته می‌شود که دارای شرط $\sum_{k\in\mathcal{Q}} p(\mathbf{x}_k) = ۱$ است [۸]. آنتروپی طرح نمونه‌گیری به‌صورت زیر تعریف می‌شود.

$$H = -\sum_{k\in\mathcal{Q}} p(\mathbf{x}_k)\log p(\mathbf{x}_k) = -E_p[\log(p(\mathbf{x}))],$$

که در آن $\mathbf{x}$ به‌عنوان نمونه‌ی تصادفی در نظر گرفته شده است. در صورتی که جرم احتمال روی مجموعه نمونه‌های ممکن به خوبی توزیع شود، آنتروپی طرح نمونه‌گیری مقدار بالایی خواهد داشت و هنگامی که تنها تعداد کمی از نمونه‌ها احتمال زیادی داشته باشند، مقدار آنتروپی کم خواهد بود. از دیگر عواملی که در مقدار آنتروپی طرح نمونه‌گیری تأثیرگذار است، اندازه‌ی تکیه‌گاه طرح است. در صورت زیاد بودن اندازه‌ی تکیه‌گاه، آنتروپی نیز مقدار بالایی خواهد داشت.

# ۳ برآورد آنتروپی در طرح نمونه‌گیری

در صورتی که اندازه‌ی نمونه‌ی $n$ و اندازه جامعه $N$ باشد حداکثر $\binom{N}{n}$ نمونه‌ی ممکن برای یک طرح نمونه‌گیری بدون جایگذاری با اندازه‌ی نمونه‌ی $n$ وجود دارد. بنابراین در صورت بزرگ بودن اندازه‌ی جامعه، به‌علت وجود تعداد نمونه‌های بسیار زیاد محاسبه‌ی آنتروپی از طریق تعریف بسیار وقت‌گیر است. گاه نیز امکان دارد تابع احتمال طرح نمونه‌گیری مشخص نباشد، از این رو به برآورد آنتروپی از طریق شبیه‌سازی روی آورده می‌شود. در صورتی که تابع احتمال یک طرح نمونه‌گیری را نتوان مشخص کرد با استفاده از الگوریتم نمونه‌گیری، نمونه‌های طرح شبیه‌سازی تولید و تابع احتمال برآورد می‌شود. یک الگوریتم نمونه‌گیری روشی است که برای انتخاب یک نمونه صرف نظر از تابع احتمال آن به‌کار می‌رود. برای برآورد تابع احتمال بدون هیچ اطلاعاتی، زمان زیادی صرف می‌شود.

اگر $m$ تعداد نمونه‌های شبیه‌سازی شده و $m_i$ برای $i =$

$m$ ,...,۱ نشان‌دهنده‌ی تعداد تکرارهای $i$-اُمین نمونه‌ی شبیه‌سازی‌شده باشد، برآوردگر حداکثر درستنمایی احتمال نمونه‌های ممکن برابر $\frac{m_i}{m}$ بوده و یک برآوردگر خام[۲] آنتروپی به‌صورت زیر خواهد بود.

$$\hat{H}_{naive} = -\frac{1}{m}\sum_{i=1}^{m}\log(\frac{m_i}{m}).$$

می‌توان نشان داد که برآوردگر فوق دارای اریبی منفی است. هر چه تعداد نمونه‌های شبیه‌سازی بیشتر باشد، اریبی این برآوردگر کمتر و دقت برآورد آن بیشتر خواهد بود. این برآوردگر به برآوردگر حداکثر درستنمایی نیز معروف است.

اگر تعداد نمونه‌های متمایز، $q$ و برآوردگر حداکثر درستنمایی احتمال نمونه‌های ممکن $\hat{p}(\mathbf{x}_i) = \frac{m_i}{m}$ باشد، شکل دیگر نمایش برآوردگر خام آنتروپی عبارت است از

$$\hat{H}_{naive} = -\sum_{i=1}^{q} \hat{p}(\mathbf{x}_i) \log(\hat{p}(\mathbf{x}_i)).$$

محققین مختلفی در جهت رفع اریبی این برآوردگر تلاش کرده‌اند. تصحیح اریبی[۳] زیر برای برآوردگر خام، توسط میلر [۳] پیشنهاد شد.

$$\hat{H}_{BC} = -\frac{1}{m}\sum_{i=1}^{m}\log(\frac{m_i}{m}) + \frac{q}{2m}.$$

اگر تابع احتمال معلوم و برای هر نمونه‌ی ممکن قابل محاسبه باشد، می‌توان با انتخاب $m$ به اندازه‌ی کافی بزرگ و یک برآوردگر مناسب، برآورد دقیق آنتروپی طرح نمونه‌گیری را به‌دست آورد. هنگامی که اندازه‌ی جامعه و نمونه بزرگ باشد، محاسبه‌ی آنتروپی از طریق تعریف به علت بررسی تعداد زیاد نمونه‌های ممکن، بسیار وقت‌گیر است. در این وضعیت از برآوردگر $\hat{H}_{PF}$ به شرح زیر استفاده می‌شود.

اگر $\mathbf{x}_k$ برای $k = 1, 2, \dots, m$ نمونه‌ی شبیه‌سازی شده‌ی $k$-اُم و $p(\mathbf{x}_k)$، احتمال این نمونه بر اساس تابع طرح مشخص باشد، آن‌گاه برآوردگر

$$\hat{H}_{PF} = -\frac{1}{m}\sum_{k=1}^{m}\log(p(\mathbf{x}_k)),$$

برای برآورد آنتروپی طرح نمونه‌گیری به‌کار می‌رود. این برآوردگر نااریب است و برای طرح‌هایی که دارای تابع احتمال مشخص هستند، استفاده می‌شود. اگر بتوان تابع احتمال $p$ را توسط $p_0$ تقریب زد، یک برآوردگر مناسب دیگر از آنتروپی ایجاد می‌شود.

$$\hat{H}_{APF1} = -\frac{1}{m}\sum_{k=1}^{m}\log(p_0(\mathbf{x}_k)).$$

برآوردگر $\hat{H}_{APF1}$ اریب است. می‌توان با محاسبه‌ی $E_p[\frac{p(\mathbf{x})}{p_0(\mathbf{x})}]$ میزان اریبی این برآوردگر را کم کرد. برای این منظور ابتدا باید $p(\mathbf{x})$ برآورد شود. و برآوردگر آنتروپی به‌صورت زیر حاصل می‌شود:

$$\begin{aligned}\hat{H}_{APF2} &= -\frac{1}{m}\sum_{k=1}^{m}\log(p_0(\mathbf{x}_k)) \\ &- \frac{1}{2}(\frac{1}{m}\sum_{k=1}^{m}\frac{\hat{p}(\mathbf{x}_k)}{p_0(\mathbf{x}_k)} - 1).\end{aligned}$$

# ۴ شبیه‌سازی

برای مقایسه‌ی برآوردگرهای آنتروپی از مثال‌های شبیه‌سازی استفاده می‌کنیم. شبیه‌سازی با استفاده از دو جامعه‌ی معروف ترات-باندسون-میستر و سامفورد-هاجک انجام شده است. در این شبیه‌سازی از نمونه‌گیری‌های پارتو [۵، ۶] و پواسون شرطی

[۲] Naive estimator

[۳] Bias Correction

تعدیل‌شده[۲] استفاده شده است. نتایج به‌دست آمده در هر دو مورد مشابه و مستقل از جامعه و روش نمونه‌گیری است.

**مثال ۱.** جامعه‌ی ترات-باندسون-میستر با اندازه‌ی جامعه $N = ۶$، اندازه‌ی نمونه‌ی $n = ۳$ و بردار احتمال شمول (احتمال آن‌که واحد $k$-اُم در نمونه‌ی تصادفی قرار گیرد) زیر را در نظر بگیرید:

$$\pi = (\frac{۱}{۳}, \frac{۱}{۳}, \frac{۱}{۳}, \frac{۲}{۳}, \frac{۲}{۳}, \frac{۲}{۳}).$$

در این جامعه با استفاده از یک طرح نمونه‌گیری بدون جایگذاری حداکثر $\binom{۶}{۳} = ۲۰$ نمونه ایجاد می‌شود. برای تولید نمونه از نمونه‌گیری پارتو استفاده می‌شود. تابع احتمال برای نمونه‌گیری مذکور به‌صورت زیر است.

$$p_{par}(\mathbf{x}) = \prod_{i=۱}^{۶} \lambda_i^{x_i}(۱-\lambda_i)^{۱-x_i} \times \sum_{k=۱}^{۶} c_k x_k, \quad |\mathbf{x}| = ۳.$$

مقدار دقیق آنتروپی با استفاده از بررسی تمام نمونه‌های ممکن ۲/۶۹۵۵۵۸ است. شبیه‌سازی با تعداد نمونه‌های مختلف برای ۵ برآوردگر مختلف ارائه‌شده در بخش قبل صورت گرفته است. در دو برآوردگر $\hat{H}_{APF۱}$ و $\hat{H}_{APF۲}$، برای تقریب $p_۰$، تابع احتمال طرح $CP$ (پواسون شرطی) به‌کار می‌رود. به تعداد $n = ۲۰$ بار شبیه‌سازی تکرار و میانگین، انحراف معیار و فاصله اطمینان برآوردگرهای مختلف در هر بار محاسبه شده و نتایج در جدول ۱ ارائه شده است. با بررسی جدول مشاهده می‌شود، در مورد هر برآوردگر با افزایش تعداد نمونه‌ی شبیه‌سازی برآورد آنتروپی دقیق‌تر و به مقدار واقعی خود نزدیک‌تر می‌شوند. تمام برآوردگرها به‌جز برآوردگر خام آنتروپی به خوبی و تقریباً مشابه به هم عمل می‌کنند. برآوردگر خام دارای اریبی منفی است و در صورت کم بودن تعداد نمونه از مقدار واقعی بسیار کمتر برآورد می‌کند. همچنین فاصله‌ی اطمینان برای این برآوردگر در حالتی که تعداد نمونه ۵۰۰ است مقدار واقعی آنتروپی را در بر نگرفته است. اما با تصحیح میزان اریبی، این برآوردگر بسیار بهتر عمل کرده و حتی با تعداد نمونه‌ی کم برآوردی نزدیک به مقدار واقعی خود ارائه می‌دهد. در واقع در برآوردگر $\hat{H}_{BC}$ همانند برآوردگر $\hat{H}_{PF}$ با افزایش تعداد نمونه برآوردها دقیق‌تر و به مقدار واقعی نزدیک‌تر می‌شود.

همان‌طور که پیش از این توضیح داده شد برآوردگرهای $\hat{H}_{PF}$، $\hat{H}_{APF۱}$ و $\hat{H}_{APF۲}$ در حالتی به‌کار می‌رود که توزیع طرح نمونه‌گیری مشخص باشد و یا با طرح دیگری بتوان تقریب زد. اما در بسیاری از روش‌های نمونه‌گیری امکان تعیین توزیع طرح وجود ندارد که در این صورت برآوردگر $\hat{H}_{BC}$ پیشنهاد می‌شود. این برآوردگر بدون نیاز به توزیع طرح نمونه‌گیری با دقتی همانند سه برآوردگر دیگر به‌خوبی عمل می‌کند. اما برای تعداد نمونه‌ی زیاد دارای سرعت برآورد بسیار کمتری است.

**مثال ۲.** جامعه‌ی سامفورد-هاجک با اندازه‌ی جامعه $N = ۱۰$، اندازه‌ی نمونه‌ی $n = ۵$ و بردار احتمال شمول زیر در نظر بگیرید:

$$\begin{aligned}\pi &= (۰/۲, ۰/۲۵, ۰/۳۵, ۰/۴, ۰/۵۰\\ &\quad, ۰/۵, ۰/۵۵, ۰/۶۵, ۰/۷, ۰/۹).\end{aligned}$$

برای تولید نمونه از نمونه‌گیری پواسون شرطی تعدیل‌شده استفاده می‌شود. تابع احتمال برای نمونه‌گیری مذکور به‌صورت زیر است.

$$p_{Acp}(\mathbf{x}) = C_{Acp} \prod_{i=۱}^{۱۰} (Ap)_i^{x_i}(۱-(Ap)_i)^{۱-x_i}, \quad |\mathbf{x}| = ۵,$$

که $(Ap)_i$ پارامترهای اصلاح‌شده و $C_{Acp}$ ثابت نرمال‌ساز

است. در این جامعه با استفاده از نمونه‌گیری فوق مقدار دقیق آنتروپی ۴/۷۲۶۹۹۰ است. در این مثال نیز شبیه‌سازی با تعداد نمونه‌های مختلف برای ۵ برآوردگر صورت گرفته است. در دو برآوردگر $\hat{H}_{APF1}$ و $\hat{H}_{APF2}$، برای تقریب $p_0$، تابع احتمال طرح $APar$ (پارتوی تعدیل شده) به‌کار می‌رود. به تعداد $n = 10$ بار شبیه‌سازی تکرار شده است. نتایج در جدول ۲ مشاهده می‌شود. در این مثال روش نمونه‌گیری و نوع جامعه تغییر داده شد. اما باز نتایج مشابه مثال قبل حاصل می‌شود. یعنی در مورد هر برآوردگر با افزایش تعداد نمونه برآوردها به مقدار واقعی نزدیک‌تر می‌شوند و خطای برآورد مقدار کمتری را نشان می‌دهند. برآوردگر خام دارای اریبی منفی است و با تصحیح میزان اریبی دارای دقت بسیار بهتری می‌شود. برای استفاده از برآوردگر خام باید تعداد نمونه‌های شبیه‌سازی شده بسیار زیاد باشد تا این برآوردگر برآوردی نزدیک به مقدار واقعی آنتروپی داشته باشد.

برآوردگرهای آنتروپی در حالتی که طرح نمونه‌گیری دارای تابع احتمال مشخص یا تقریبی می‌باشند تقریبا مشابه هم عمل کرده و استفاده از برآوردگر خام با تصحیح میزان اریبی دارای خطای برآورد کمتر است و به‌کارگیری آن در صورت مشخص نبودن تابع احتمال طرح نمونه‌گیری پیشنهاد می‌شود.

جدول ۱: مقایسه‌ی برآوردگرهای مختلف آنتروپی در جامعه‌ی $TBM$ و با استفاده از نمونه‌گیری پارتو

| برآوردگر | تعداد نمونه | میانگین | انحراف معیار | فاصله اطمینان |
|---|---|---|---|---|
| | ۵۰۰ | ۶۸۱۰۷/۲ | ۰۲۵۷/۰ | (۶۹۲۳۳/۲،۶۶۹۸۰/۲) |
| $\hat{H}_{naive}$ | ۱۰۰۰ | ۶۸۷۲۳/۲ | ۰۲۳۶/۰ | (۶۹۷۵۷/۲،۶۷۶۸۹/۲) |
| | ۵۰۰۰ | ۶۸۹۹۲/۲ | ۰۰۸۳/۰ | (۶۹۵۰۵/۲،۶۸۴۷۹/۲) |
| | ۵۰۰ | ۶۹۹۸۹/۲ | ۰۲۲۵/۰ | (۷۰۹۷۷/۲،۶۹۰۰۰/۲) |
| $\hat{H}_{BC}$ | ۱۰۰۰ | ۶۹۹۴۲/۲ | ۰۱۶۷/۰ | (۷۰۶۷۵/۲،۶۹۲۱۰/۲) |
| | ۵۰۰۰ | ۶۹۴۵۱/۲ | ۰۰۶۴/۰ | (۶۹۷۳۴/۲،۶۹۱۶۹/۲) |
| | ۵۰۰ | ۶۹۹۷۷/۲ | ۰۲۵۲/۰ | (۷۱۰۸۲/۲،۶۸۸۷۲/۲) |
| $\hat{H}_{PF}$ | ۱۰۰۰ | ۶۹۷۱۷/۲ | ۰۱۷۴/۰ | (۷۰۴۸۲/۲،۶۸۹۵۴/۲) |
| | ۵۰۰۰ | ۶۹۵۴۲/۲ | ۰۰۶۱/۰ | (۶۹۸۱۱/۲،۶۹۲۷۳/۲) |
| | ۵۰۰ | ۶۹۹۹۱/۲ | ۰۲۶۱/۰ | (۷۱۱۳۱/۲،۶۸۸۵۱/۲) |
| $\hat{H}_{APF۱}$ | ۱۰۰۰ | ۶۹۸۷۹/۲ | ۰۱۵۸/۰ | (۷۰۸۴۵/۲،۶۸۹۱۳/۲) |
| | ۵۰۰۰ | ۶۹۶۶۰/۲ | ۰۰۶۵/۰ | (۷۰۰۴۰/۲،۶۹۲۸۰/۲) |
| | ۵۰۰ | ۶۹۹۸۳/۲ | ۰۲۳۵/۰ | (۷۱۰۶۵/۲،۶۸۹۰۱/۲) |
| $\hat{H}_{APF۲}$ | ۱۰۰۰ | ۶۹۷۵۳/۲ | ۰۱۶۹/۰ | (۷۰۵۷۰/۲،۶۸۹۳۶/۲) |
| | ۵۰۰۰ | ۶۹۵۹۴/۲ | ۰۰۵۹/۰ | (۶۹۹۰۹/۲،۶۹۲۷۹/۲) |

جدول ۲: مقایسه‌ی برآوردگرهای مختلف آنتروپی در جامعه‌ی سامفورد-هاجک و با استفاده از نمونه‌گیری پواسون شرطی تعدیل شده

| برآوردگر | تعداد نمونه | میانگین | انحراف معیار | فاصله اطمینان |
|---|---|---|---|---|
| | ۱۰۰۰ | ۶۰۹۶۵۴/۴ | ۰۲۲۵/۰ | (۶۲۳۵۹/۴،۵۹۵۷۲/۴) |
| $\hat{H}_{naive}$ | ۵۰۰۰ | ۶۸۸۶۷/۴ | ۰۱۰۹/۰ | (۶۹۵۴۲/۴،۶۸۱۹۳/۴) |
| | ۱۰۰۰۰ | ۷۱۲۳۱/۴ | ۰۱۰۷/۰ | (۷۱۸۹۴/۴،۷۰۵۶۷/۴) |
| | ۱۰۰۰ | ۷۴۲۴۸/۴ | ۰۱۲۱/۰ | (۷۴۹۹۸/۴،۷۳۴۹۹/۴) |
| $\hat{H}_{BC}$ | ۵۰۰۰ | ۷۲۷۲۲/۴ | ۰۰۸۳/۰ | (۷۳۲۳۵/۴،۷۲۲۰۸/۴) |
| | ۱۰۰۰۰ | ۷۲۳۴۷/۴ | ۰۰۷۰/۰ | (۷۲۷۸۳/۴،۷۱۹۱۲/۴) |
| | ۱۰۰۰ | ۷۴۷۲۹/۴ | ۰۱۹۶/۰ | (۷۵۹۴۳/۴،۷۳۵۱۶/۴) |
| $\hat{H}_{PF}$ | ۵۰۰۰ | ۷۳۱۹۵/۴ | ۰۱۱۹/۰ | (۷۳۹۳۲/۴،۷۲۴۵۸/۴) |
| | ۱۰۰۰۰ | ۷۲۳۳۶/۴ | ۰۰۵۸/۰ | (۷۲۶۹۶/۴،۷۱۹۷۷/۴) |
| | ۱۰۰۰ | ۷۴۶۸۹/۴ | ۰/۰ ۲۱۱ | (۷۳۵۴۹/۴ ،۷۵۸۲۹/۴ ) |
| $\hat{H}_{APF۱}$ | ۵۰۰۰ | ۷۲۹۹۶/۴ | ۰۱۰۸/۰ | (۷۲۳۹۸/۴ ،۷۳۵۹۴/۴) |
| | ۱۰۰۰۰ | ۷۲۶۱۳/۴ | ۰۰۶۵/۰ | (۷۲۳۲۱/۴ ،۷۲۹۰۵/۴) |
| | ۱۰۰۰ | ۷۴۵۳۴/۴ | ۰۲۰۵/۰ | (۷۵۵۳۱/۴،۷۳۵۳۷/۴) |
| $\hat{H}_{APF۲}$ | ۵۰۰۰ | ۷۳۰۲۴/۴ | ۰۱۲۰/۰ | (۷۳۶۵۲/۴،۷۲۳۹۶/۴) |
| | ۱۰۰۰۰ | ۷۲۵۷۰/۴ | ۰۰۶۲/۰ | (۷۲۹۴۲/۴،۷۲۱۹۸/۴) |

# ۵ بحث و نتیجه‌گیری

در صورتی که اندازه‌ی نمونه $n$ و اندازه جامعه $N$ باشد حداکثر $\binom{N}{n}$ نمونه‌ی ممکن برای یک طرح نمونه‌گیری بدون جایگذاری با اندازه‌ی نمونه‌ی $n$ وجود دارد. بنابراین در صورت بزرگ بودن اندازه‌ی جامعه و نمونه، به‌علت وجود تعداد نمونه‌های بسیار زیاد محاسبه‌ی آنتروپی از طریق تعریف بسیار وقت‌گیر است. در این صورت با استفاده از شبیه‌سازی می‌توان آنتروپی طرح نمونه‌گیری را برآورد کرد. در این مقاله برآوردگرهای مختلف آنتروپی با یکدیگر مقایسه شد. هر چه تعداد نمونه‌های شبیه‌سازی شده بیشتر باشند برآوردهای انجام شده توسط برآوردگرهای مختلف به مقدار واقعی نزدیک‌تر خواهند بود. به‌خصوص درباره‌ی برآوردگر خام باید تعداد نمونه‌های شبیه‌سازی‌شده بسیار زیاد باشد تا برآورد به مقدار واقعی نزدیک شود. این برآوردگر دارای اریبی منفی بوده و کمتر از مقدار واقعی برآورد می‌کند. با تصحیح میزان اریبی برآوردگر خام، برآورد به مقدار واقعی بسیار نزدیک‌تر می‌شود و در صورت نامشخص بودن تابع توزیع طرح نمونه‌گیری استفاده از این برآوردگر مناسب خواهد بود. در حالتی که توزیع طرح نمونه‌گیری مشخص است برآوردگر $\hat{H}_{PF}$ که دارای سرعت برآورد بالاتری نسبت به برآوردگر $\hat{H}_{BC}$ است، پیشنهاد می‌شود.

## مراجع


[1] Grafstrom, A. (2010). Entropy of unequal probability sampling designs. Statist. Methodol., 7, 84-97.

[2] Hajek, J. (1981). Sampling From a Finite Population. Marcel Dekker, New York.

[3] Miller, G. (1955). Note on the bias of information estimates. In: H. Quastler (Ed.), Information Theory in Psychology II-B, Free Press, Glencoe, IL, 95-100.

[4] Renyi, A. (1961). On measures of entropy and information. Proc, Berekeley Symposium, Statist, Probability, 1, 547-561.

[5] Rosen, B. (1997). Asymptotic theory for order sampling. J. Statist. Plann. Inference, 62, 135-158.

[6] Rosen, B. (1997). On sampling with probability proportional to size. J. Statist. Plann. Inference, 62, 159-191.

[7] Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-423, 623-656.

[8] Tille, Y. (2006). Sampling Algorithms. Springer series in statistics, Springer science, Business media, Inc., New York.

[9] Tsallis, C. (1988). Possible generalizations of Boltzmann-Gibbs statistics. Journal of Statistical Physics, 52, 479-487.