روشی گرافیکی برای انتخاب بهینه کسر نمونه

اکرم فخاری اسفریزی[۱] ، ناهید سنجری فارسی‌پور[۲]

چکیده:

ضریب دمی — وایبل، θ، به عنوان ضریب تغییرات منظم معکوس تابع انباشتگی نرخ شکست تعریف می‌شود. برآورد این پارامتر دم توزیع بر اساس آماره‌های ترتیبی بزرگ به‌دست می آید. بیرلنت و همکاران [۱] یک برآورد بر اساس k_n تای بالایی آماره‌های ترتیبی معرفی کردند. بنابراین مساله مهم در برآورد کردن این است که چه تعدادی از نمونه در برآورد استفاده شود. ما با استفاده از روش گرافیکی روشی را برای انتخاب کسر نمونه پیشنهاد می‌کنیم.

واژه‌های کلیدی: ضریب دمی — وایبل، آماره‌های ترتیبی، نرمال بودن مجانبی.

۱ مقدمه

فرض کنید $X_1, X_2, ..., X_n$ یک دنباله از متغیرهای تصادفی مستقل و هم توزیع با تابع توزیع F باشند. ما مساله برآورد ضریب دمی — وایبل[۳] را بررسی می‌کنیم وقتی که دم توزیع در روابط زیر صدق کند:

$$1 - F(x) = e^{-H(x)} \quad x \geq x_0 \geq 0 \qquad (1)$$

$$H^{\leftarrow}(t) = \inf\{x : H(x) \geq t\} = t^{\theta} l(t)$$

(x_0 نقطه آستانه توزیع است) به‌طوری که l یک تابع با آهنگ تدریجی باشد یعنی اگر برای $\lambda > 0$ وقتی $x \to \infty$ داشته باشیم:

$$\frac{l(\lambda, x)}{l(x)} \to 1$$

توزیعی که در روابط بالا صادق است جزُ خانواده توزیع دمی — وایبل می‌باشد و پارامتر θ را ضریب دمی ـوایبل گویند. تابع $H^{\leftarrow}$ یک تابع با آهنگ تغییر منظم با شاخص θ در بی‌نهایت گفته می‌شود و می‌نویسیم $H^{\leftarrow} \in R_{\theta}$ (در این‌جاعلامت $\leftarrow$ نشان دهنده معکوس تابع است.) در (۱) تابع l می‌تواند یک تابع ثابت باشد که در این حالت پارامتر θ متناسب با پارامتر شکل توزیع وایبل می‌شود[۳] ، توزیع‌هایی هم با l غیر ثابت وجود دارند مثل نرمال، گاما، وایبل توسیع یافته[۴] که در (۱) صدق می‌کنند. چنین توزیع‌هایی در بیمه و مطالعات قابلیت اعتماد کاربرد دارند.

توزیع نرمال $N(\mu, \sigma^2)$

$$H^{\leftarrow}(x) = x^{\frac{1}{2}} l(x)$$

$$l(x) = \sqrt{2}\sigma - \frac{\sigma}{2^{\frac{3}{2}}} \frac{log x}{x} + O(1/x)$$

$$\theta = \frac{1}{2}$$

[۱] گروه آمار، دانشگاه شیراز
[۲] گروه آمار، دانشگاه شیراز
[۳] Weibull Tail-Coefficent
[۴] Extended Weibull

توزیع گاما $\Gamma(\beta,\alpha)$

$$\begin{aligned} f(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \\ H^{\leftarrow}(x) &= xl(x) \\ l(x) &= \frac{1}{\beta}+\frac{\alpha-1}{\beta}\frac{logx}{x}+O(1/x) \\ \theta &= 1 \end{aligned}$$

توزیع وایبل $W(\alpha,\lambda)$

$$\begin{aligned} f(x) &= \frac{\alpha}{\lambda^\alpha}x^{\alpha-1}e^{\frac{-x^\alpha}{\lambda^\alpha}} \\ H^{\leftarrow}(x) &= \lambda x^{\frac{1}{\alpha}} \\ l(x) &= \lambda \\ \theta &= \frac{1}{\alpha} \end{aligned}$$

در اینجا اگر برآورد پارامتر θ را بخواهیم در واقع باید برآورد پارامتر شکل توزیع وایبل را پیدا کنیم.

توزیع وایبل توسیع یافته

$$\begin{aligned} EW(\alpha,\beta) &\quad ;\alpha\in(0,1),\beta\in R \\ F(x) &= 1-r(x)\exp(-x^\alpha)\ \ r\in R_\beta \\ H^{\leftarrow}(x) &= x^{\frac{1}{\alpha}}l(x) \\ l(x) &= 1+\frac{\beta}{\alpha^2}\frac{\log x}{x}+O(1/x) \\ \theta &= \frac{1}{\alpha} \end{aligned}$$

چون پارامتر θ، پارامتر دم توزیع است پس برای برآورد آن از بخشی از نمونه استفاده می‌شود. یکی از مهمترین پژوهش‌ها برای برآورد θ توسط برد [۲] صورت گرفته است که برآوردگری براساس مقادیر رکورد پیشنهاد کرد. روش دیگر برآورد از k_n تای بالایی آماره‌های ترتیبی استفاده می‌کند یعنی برآورد براساس $x_{n-k_n+1,n}\le\ldots\le X_{n,n}$ که k_n یک دنباله از اعداد صحیح است به‌طوری‌که $1\le k_n\le n$ باشد. به این ترتیب برای برآورد θ، کسری از نمونه مورد استفاده قرار می‌گیرد. انتخاب بهینه برای این کسر از نمونه[۵] مساله مورد بررسی در این مقاله است.

برآوردگر گرارد بر اساس همین روش به‌صورت زیر می‌باشد[۴]:

$$\hat\theta_n^G=\frac{\sum_{i=1}^{k_n-1}(\log(X_{n-i+1,n})-\log(X_{n-k_n+1,n}))}{\sum_{i=1}^{k_n-1}(\log_2(\frac{n}{i}-\log_2(\frac{n}{k_n}))}$$

که $\log_2(t)=\log(\log(t))\ ;t>1$

در [۱] برای به‌دست آوردن برآوردگر θ از تابع $e(x)=E(X-x|X>x)$ یعنی تابع میانگین باقی‌مانده عمر استفاده شده است[۱]. تحت (۱) داریم:

$$(\log(x))\frac{e(H^{\leftarrow}(\log(x)))}{H^{\leftarrow}(\log(x))}\to\theta\ \ as\ \ x\to\infty$$

و برآوردگر برای θ به فرم زیر خواهد بود:

$$\begin{aligned} \hat\theta_n^{BBTV} &= \frac{\log(\frac{n}{k_n})}{X_{n-k_n+1,n}}\frac{1}{k_n-1} \\ &\times \sum_{i=1}^{k_n-1}(X_{n-i+1,n}-X_{n-k_n+1,n}) \end{aligned}$$

برآوردگرهای دیگری برای θ توسط برانیاتوسکی [۳]، کلاپلبرگ و ویلاسنور[۶] ارائه شده‌اند.

۲ روش گرافیکی انتخاب k_n

کلید اصلی حل مساله برآورد θ در انتخاب صحیح k_n است، چرا که اگر k_n خیلی بزرگ انتخاب شود اریبی

[۵] Sample Fraction

$\hat{\theta}_n$ بزرگ می‌شود و اگر k_n خیلی کوچک انتخاب شود، واریانس $\hat{\theta}_n$ خیلی بزرگ می‌شود.

یکی از روش‌ها برای انتخاب k_n بهینه می‌تواند با استفاده از نمودار چندک، بر اساس برآوردگر گرارد باشد. این روش بر اساس تئوری زیر می‌باشد:

$$q(t) = (۱ - F)^{\leftarrow}(t) = H^{\leftarrow}(\log(\frac{۱}{t}))$$

تحت (۱)

$$\begin{aligned}\log(q(t)) - \log(q(s)) &= \theta(\log_۲(\frac{۱}{s}) \\ &- \log_۲(\frac{۱}{t}) + \log\left(\frac{l(\log(\frac{۱}{t}))}{l(\log(\frac{۱}{s}))}\right)\end{aligned}$$

چون l یک تابع با آهنگ تغییر تدریجی است پس

$$\log(q(t)) - \log(q(s)) \approx \theta(\log_۲(\frac{۱}{t}) - \log_۲(\frac{۱}{s})$$

این دقیقاً مثل این است که معادله خطی با شیب θ نوشته‌ایم که این رابطه حداقل برای tهای کوچک خطی خواهد بود.

در این‌جا اگر $t = \frac{i}{n}$ را در نظر بگیریم با توجه به این‌که $\hat{q}(\frac{i}{n}) = X_{n-i+۱,n}$ را به‌عنوان برآورد ناپارامتری از چندک‌ها در نظر می‌گیریم. پس باید $(\log_۲(\frac{n}{i}), X_{n-i+۱,n})$ یک رابطه خطی تقریبی با شیب θ داشته باشند که در [۴] آمده است.

در این روش گرافیکی، روشی که برای انتخاب k_n بهینه انتخاب می‌شود بدین صورت است که k_n را طوری انتخاب می‌کنیم تا نمودار خطی شود. البته اشکال این روش این است که یک انتخاب صریح به ما نمی‌دهد و انتخاب شهودی است. در نمودار مربوط به توزیع وایبل این نمودار به صورت کاملاً خطی است دلیل این است که چون l در این توزیع یک تابع ثابت است پس (۱.۲) به صورت تقریبی نخواهد بود و یک رابطه دقیق است.

در این روش گرافیکی، از دیدگاهی دیگر، می‌توان یک برآورد برای θ به‌دست آورد. با توجه به این‌که می‌دانیم حداقل برای i های بزرگ نمودار خطی خواهد شد، پس می‌توان شیبی را که برای iهای بزرگ وجود دارد را به‌عنوان برآوردی برای θ به‌حساب آورد(شکل ۱).

لازم به ذکر است در نمودارها $|N|$ منظور قدرمطلق مقادیر شبیه‌سازی شده از توزیع نرمال است.

روش گرافیکی دیگر با استفاده از نمودار چندک، بر اساس برآوردگر بیرلنت و همکاران [۱] می‌باشد که به‌صورت زیر است:

$$\begin{aligned}\hat{\theta}_n^{BBTV} &= \frac{\log(\frac{n}{k_n})}{X_{n-k_n+۱,n}}\frac{۱}{k_n - ۱} \\ &\times \sum_{i=۱}^{k_n-۱}(X_{n-i+۱,n} - X_{n-k_n+۱,n})\end{aligned}$$

با تعریف کردن $v_{m,n}$

$$v_{m,n} = \frac{X_{n-m}}{\frac{۱}{m}\sum_{i=۱}^{m}(X_{n-i+۱,n} - X_{n-m,n})}$$

خواهیم داشت:

$$\hat{\theta}_n^{BBTV} = \left(\frac{۱}{v_{k_n-۱,n}}\right) / \left(\frac{۱}{\log(\frac{n}{k_n})}\right)$$

اگر نمودار نقاط $(\frac{۱}{\log(\frac{n}{k_n})}, \frac{۱}{v_{k_n-۱,n}})$ رسم شود، باز هم جایی‌که نمودار شروع به خطی شدن می‌کند را انتخاب مناسبی برای k_n در نظر می‌گیریم.

از دیدگاهی دیگر این نمودار دارای شیب θ می‌باشد. با در نظر گرفتن این مطلب می‌توان برآوردی برای θ به دست آورد.

۳ روش تطبیقی برای انتخاب k_n

در یک روش دیگر برای به‌دست آوردن k_n بهینه، دیابلت و همکاران [۴] میانگین مربعات خطای مجانبی مربوط به $\hat{\theta}_n^{ML}(k_n)$ را به‌صورت

$$AMSE(\hat{\theta}_n^{Ml}(k_n)) = \frac{\theta^2}{k_n} + \left(b(\log\left(\frac{n}{k_n}\right)\frac{1}{k_n}\sum_{i=1}^{k_n}(\frac{\log(\frac{n}{i})}{\log\left(\frac{n}{k_n}\right)})^n\right)^2 \quad (2)$$

ارایه کردند که $\hat{\theta}_n^{Ml}(k_n)$ برآورد حداکثر درستنمایی θ است [۳].

در [۴] به‌منظور به‌دست آوردن توزیع حدی برآوردگر معرفی شده، شرط دومی برای توزیع‌های دمی — وایبل مطرح کرده است: یک $\eta \geq 0$ و یک تابع $b(x)$ وجود داشته باشند، به‌طوری که وقتی $x \to \infty$ آن‌گاه $b(x) \to 0$. همچنین برای هر $1 < A < \infty$ به‌طوری که برای هر $\lambda \in [1, A]$

$$\log(\frac{l(\lambda x)}{l(x)}) \sim b(x)k_\eta(\lambda) \quad (3)$$
$$k_\eta(\lambda) = \int_1^\lambda u^{\eta-1}du = \frac{1}{\eta}(\lambda^\eta - 1)$$

(منظور از $\sim$ هم‌ارزی می‌باشد) پارامتر $\eta \geq 0$ نرخ همگرایی $\frac{l(\lambda x)}{l(x)}$ به ۱ را نشان می‌دهد. هر چه η به صفر نزدیکتر باشد نرخ همگرایی آهسته‌تر است و نیز داریم که $|b| \in R_\eta$

برآورد η بسیار پیچیده است به همین علت مقدار کانونی ۱— برای آن در نظر می‌گیریم. مقدار η را در رابطه (۲) برابر ۱— قرار دهیم این کار باعث سادگی محاسبات می‌شود. همچنین در رابطه (۲) برآوردهای حداقل مربعات θ و B را قرار دهیم و به دنبال k_n ای می‌گردیم که $AMSE(\hat{\theta}_n^{Ml}(k_n))$ را حداقل کند.

$$\widehat{AMSE}(\hat{\theta}_n^{Ml}(k_n)) = \frac{(\hat{\theta}_n^{LS}(k_n))^2}{k_n} + \left(\hat{b}(\log(\frac{n}{k_n})\frac{1}{k_n}\sum_{j=1}^{k_n}(\frac{\log(\frac{n}{j})}{\log(\frac{n}{k_n})})^{-1}\right)^2$$

که این رابطه برآوردی از رابطه (۲) می‌باشد.

پس برآوردی که برای k_n به دست می‌آید به‌صورت زیر می‌باشد:

$$\hat{k}_n = arg\ min_{k_n}\ \ \widehat{AMSE}(\hat{\theta}_n^{Ml}(k_n))$$

استفاده از روش فوق در مطالعات شبیه‌سازی به‌شرح زیر است:

از توزیع‌های $W(4,4), |N|(0,1), \Gamma(4,1), \Gamma(0/25,1)$ و $W(0/25, 0/25)$ به تعداد ۱۰۰ بار نمونه ۵۰۰ تایی شبیه‌سازی شده است. پس $N = 100$ و $n = 500$

$$(x_{n,i})_{i=1,\ldots,N}$$

روی هر نمونه $x_{n,i}$، برآورد $\hat{\theta}_n^{Ml}(k_n)$ را به‌عنوان تابعی از k_n به‌دست می‌آید که این برآوردها به‌ازای $k_n = 2, \ldots, 350$ می‌باشد.

به‌منظور یافتن انتخاب مناسب برای k_n مقدار زیر را محاسبه کرده

$$\hat{k}_n = argmin_{k_n \in [2,350]}\widehat{AMSE}(\hat{\theta}_{n,i}^{Ml}(k_n)) \ \ ; i = 1, \ldots, N$$

سپس

$$\mu(\hat{k}_n) = \frac{1}{N}\sum_{i=1}^{N}\hat{k}_{n,i}$$
$$\sigma(\hat{k}_n) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{k}_{n,i} - \mu(\hat{k}_n))^2}$$
$$K_n^{opt} = argmin_{k_n \in [2,350]}AMSE(\hat{\theta}_n^{Ml}(k_n))$$

برآوردهای θ را بر اساس $\hat{K}_{n,i}$ به‌دست می‌آوریم:

$$\mu(\hat{\theta}_n^{ML}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{n,i}^{ML}(\hat{k}_{n,i})$$

$$\sigma(\hat{\theta}_n^{ML}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{n,i}^{ML}(\hat{k}_{n,i}) - \mu(\hat{\theta}_n^{Ml}))^2}$$

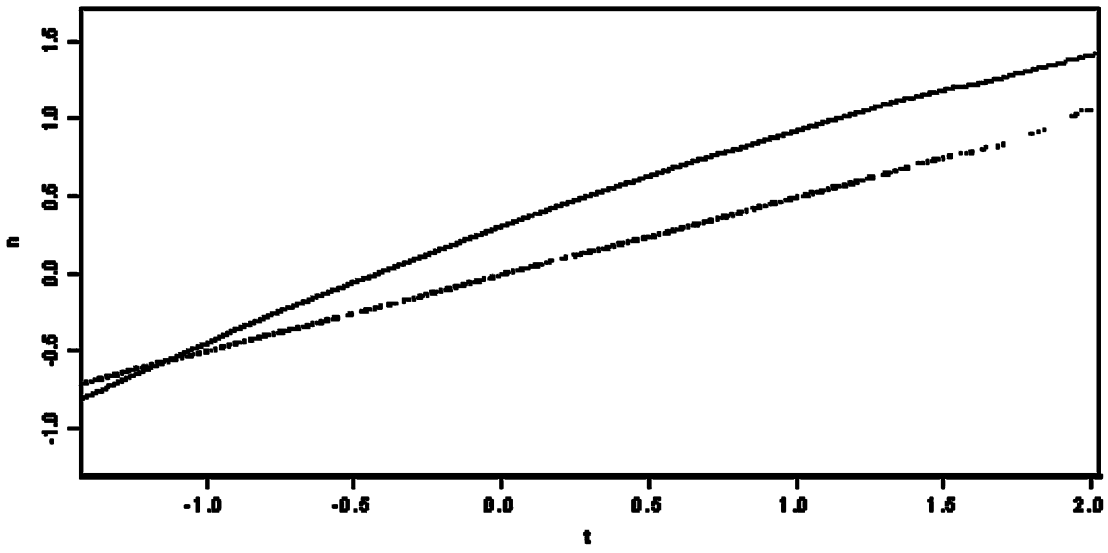
همچنین کمیت R_n^2 یعنی نسبت میانگین توان دوم خطای تجربی به حداقل میانگین توان دوم می‌تواند به‌عنوان یک معیار به‌منظور انتخاب بهینه k_n استفاده گردد.

$$R_n^2 = \frac{\sum_{i=1}^{N}(\hat{\theta}_{n,i}^{ML}(\hat{k}_{n,i}) - \theta)^2}{min_{k_n\in[2,350]}\ \sum_{i=1}^{N}(\hat{\theta}_n^{Ml}(k_n) - \theta)^2}$$

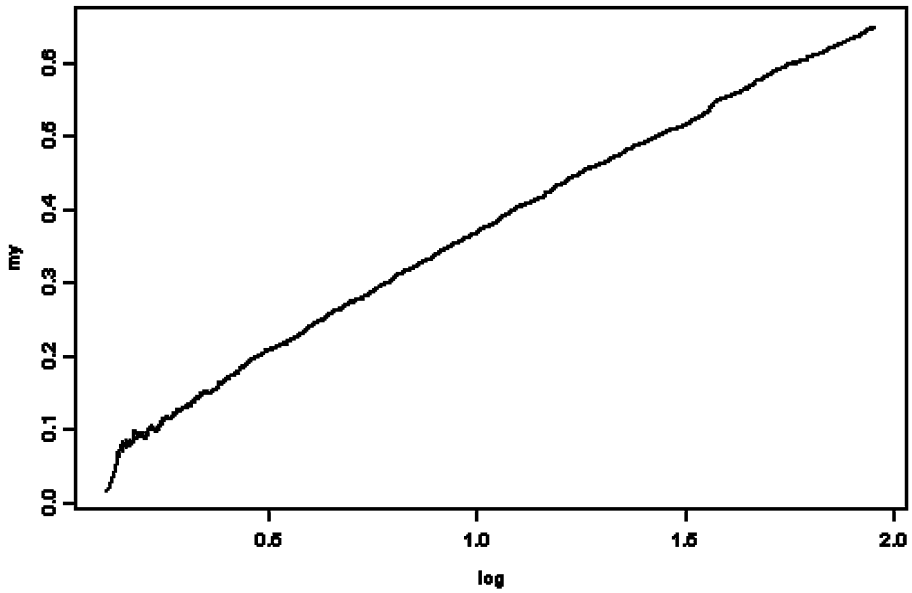
با استفاده از شبیه‌سازی، روش فوق در به‌دست آوردن k_n استفاده می‌شود که نتایج را در جدول (۱) قرار داده‌ایم.

همان‌طورکه پیداست نتایج بهتری برای برآورد $\hat{\theta}_n$ به‌دست می‌آید. برای $\Gamma(0/25, 1)$ مقدار $k_n^{opt} = 186$ شده که نشان می‌دهد برای برآورد θ از آماره ترتیبی ۱۸۶ تا ۵۰۰ استفاده شده است.

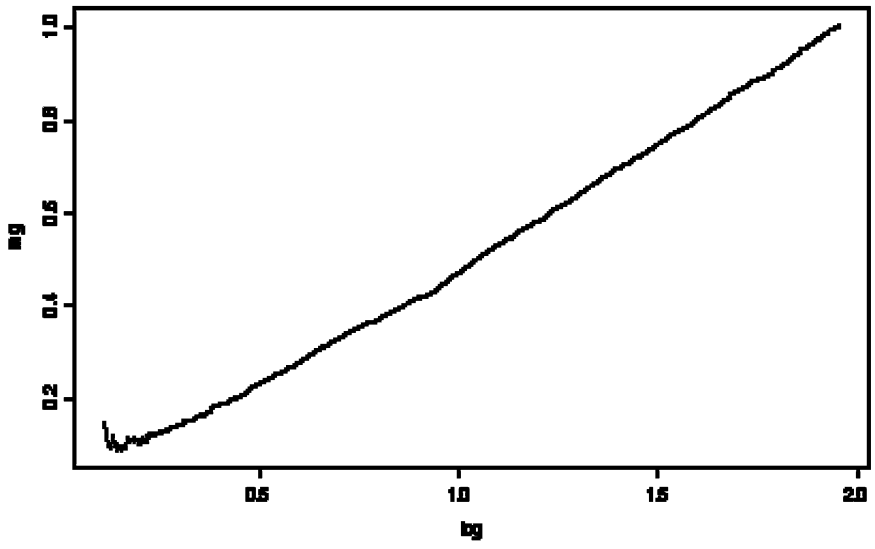
مقدار ایده‌آل برای R_n برابر ۱ می‌باشد. در جدول R_n معمولاً نزدیک ۱ است بجز برای توزیع وایبل، که در مورد این توزیع مقادیر بزرگ R_n و همچنین مقادیر بزرگ $\mu(\hat{(k)}_n)$ نشان می‌دهد که k_n بهینه بزرگتر از ۳۵۰ انتخاب می‌شود.



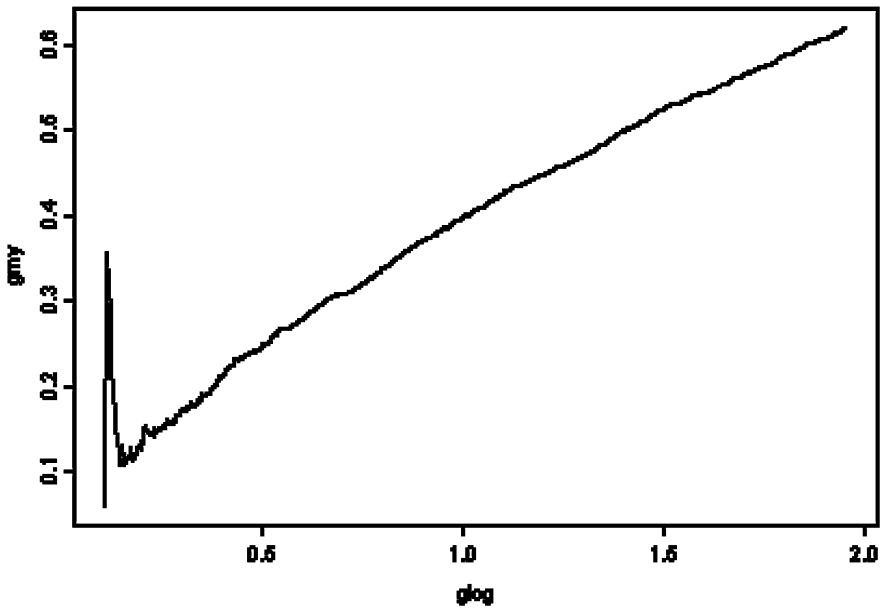
شکل ۱. نمودار چندک مربوط به توزیع $W(2, 1)$ و توزیع $|N|(1, 1)$
شبیه‌سازی به تعداد ۱۰۰۰۰ نمونه تصادفی



شکل ۲. نمودار مربوط به توزیع $W(۲, ۱)$ با ۱۰۰۰۰ نمونه



شکل ۳. نمودار مربوط به توزیع $|N|(۲, ۱)$ با ۱۰۰۰۰ نمونه



شکل ۴. نمودار مربوط به توزیع $\Gamma(۴, ۱)$ با ۱۰۰۰۰ نمونه

جدول ۱. نتایج شبیه‌سازی برای به دست آوردن انتخاب بهینه k_n .

توزیع	θ	η	$\mu(\hat{k}_n)$	$\sigma(\hat{k}_n)$	$\mu(\hat{\theta}_n^{ML})$	$\sigma(\hat{\theta}^{ML})$	R_n	k_n^{opt}
$\Gamma(۰/۲۵, ۱)$	۱	-۱	۱۰۵/۵	۶۲/۲	۱/۶۶۷	۰/۲۹۴	۱/۲۶	۱۸۶
$\Gamma(۴, ۱)$	۱	-۱	۲۲۲/۷	۸۲/۱	۰ /۰۵۴۸	۰/۰۵۱	۱/۱۳	۱۸۴
$\|N\|(۰, ۱)$	۰/۵	-۱	۲۴۶/۶	۸۱/۱	۰/۶۷۹	۰/۰۱۰۹	۱/۲۱	۱۸۹
$W(۲۵, ۲۵)$	۴	$-\infty$	۳۰۵/۸	۵۹/۰	۴/۰۱۶	۰/۲۶۵	۱/۶۲	۳۵۰
$W(۴, ۴)$	۰/۲۵	$-\infty$	۳۱۰/۴	۵۰/۹	۰/۲۴۹	۰/۰۱۳	۱/۴۳	۳۵۰

مراجع

[1] Beirlant, J., Broniatowski, M., Teugels, J. L. and Vynckier, P. (1995), The mean residual

[2] Berred, m. (1991), Record values and the estimation of the weibull tail-coefficient, *comptes-rendus de l'Academie des sciencest. Serie I*, 312, 943-946.

[3] Broniatowski, M. (1993). On The Estimation of Weibull Tail Coefficient, *J.Statist. Plan. Inference*, 35, 349-366.

[4] Diebolt, J., Gardes, S. L., Girard, S. and Guillo, A. (2006), Bias- Reduced Estimators of the Weibull Tail-Coeficient, Test-girad.pdf. 10 October 2006.

[5] Girard, S. (2004), A Hill Type Estimate of the Weibull Tail-Coeficient, *Commun. Statist. Theor. Meth*, 33(2), 205-234.

[6] Kluppelberg, C. and Villasenor, J. A. (1993), Estimation of distribution tails- a semiparametric approach,*deutschen gesellschaft fur versicherungs mathematik XXI(2)*, 213-235.