دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گزارش درس یادگیری ماشین آماری

«آزمایش کامپیوتری دوم»

گردآورنده: سعید دادخواه (۹۲۳۱۰۶۶)

استاد: دکتر نیک‌آبادی

دی ۱۳۹۶

مقدمه

تمامی آزمایش‌ها با زبان R انجام شده‌اند. برای تولید نمودارها از کتابخانه ggplot2، در آزمایش‌هایی که نیاز بود چند نمودار کنار هم رسم شود از کتابخانه gridExtra استفاده شده است. در ابتدای هر تمرین از دستور rm(ls()) برای پاکسازی متغیرهای قبلی محیط استفاده شده است. همچنین در صورتی که در آزمایش نیاز به تولید داده تصادفی بود سید برابر عدد ۹۲۳۱۰۶۶ (شماره دانشجویی گردآورنده) قرار گرفته است تا نتایج اجرای چند باره‌ی دسته‌کدها دقیقا مشابه یکدیگر و مخصوصا گزارش باشد.

پیشنیازهای اجرای کدها

- زبان R
- کتابخانه‌های مورد نیاز
  - ggplot2
  - gridExtra

دسته‌کدها به شکلی نوشته شده‌اند که هربار وجود کتابخانه‌های مورد نیاز را تست می‌کنند و در صورتی که کتابخانه مورد نظر وجود نداشت آن را دانلود و نصب می‌کنند.

آزمایش ۱

فاصله Kullback-Leibler برای توزیع‌های مختلف به شکل زیر محاسبه شده است:

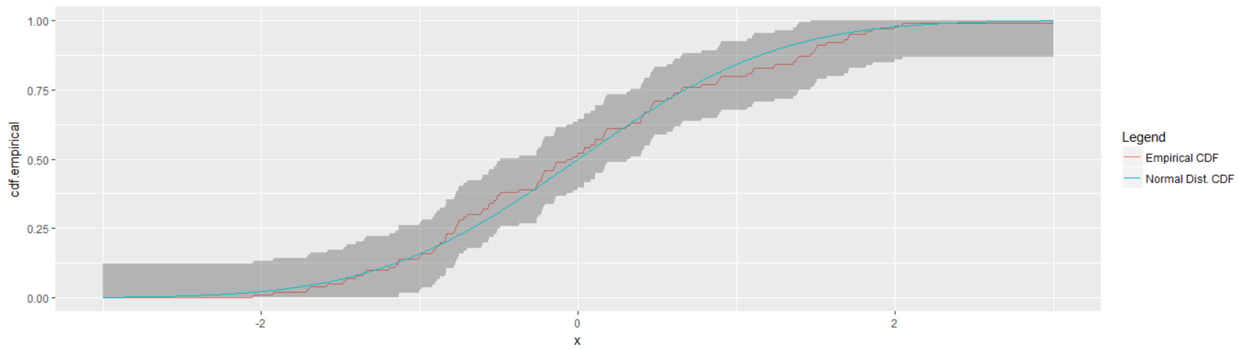
```
[1] "D(p, q1) = 7.313244"
[1] "D(q1, p) = 0.686528"
[1] "D(p, q2) = 0.862784"
[1] "D(q2, p) = 1.168750"
[1] "D(p, q3) = 7.313244"
[1] "D(q3, p) = 0.686528"
```

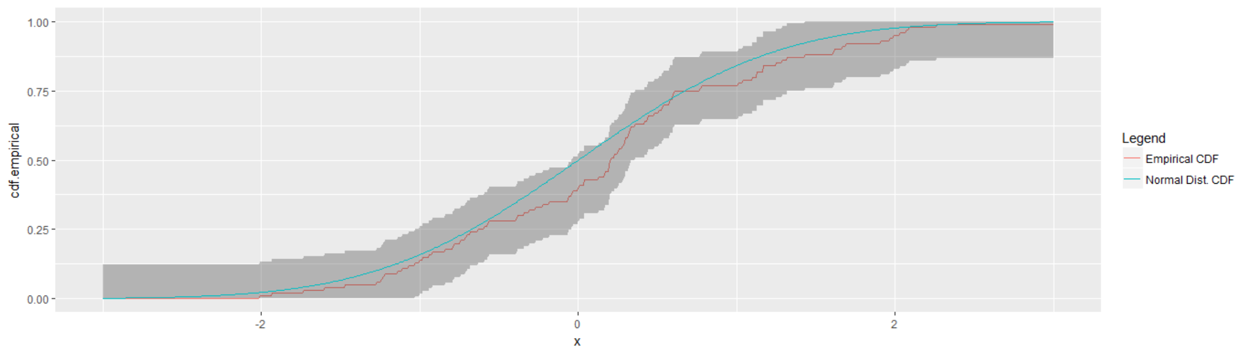
در نتیجه $q_1^*\sim Normal(5, 1.5)$ و $q_2^*\sim Normal(3,0.5)$ یا $q_2^*\sim Normal(7,0.5)$ به دست می‌آیند.

بهینه‌سازی $D_{KL}(p,q)$ باعث می‌شود که توزیع بهینه به سمتی برود که ممان‌هایش مانند ممان‌های p است. یعنی میانگین و واریانسش شبیه p خواهند بود و از طرفی بهینه‌سازی $D_{KL}(q,p)$ باعث می‌شود که توزیع بهینه به سمت یکی از مدهای p برود. پس اگر فرض کنیم که در زمینه تولید تصویر در حال بحث هستیم می‌توان گفت اگر فرض کنیم که بتوانیم تصاویر را به دو دسته تقسیم کنیم با q_1^* که مقدار $D_{KL}(p,q)$ را بهینه می‌کند تصاویری مشابه یکی از دو دسته تولید خواهد شد و با q_2^* که مقدار $D_{KL}(q,p)$ را بهینه می‌کند تصاویری ما بین دو دسته تولید خواهد شد.

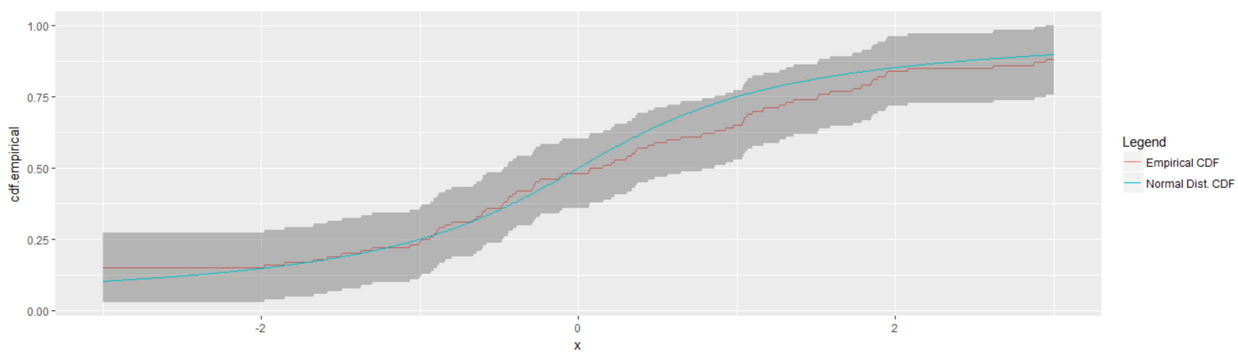
آزمایش ۲

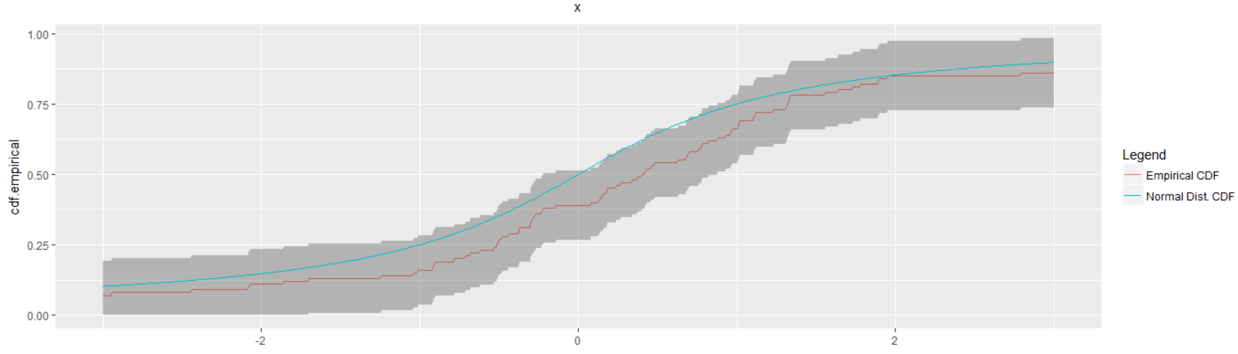
تصاویر زیر تابع توزیع تجمعی را برای دو حالت نرمال و کوشی نمایش می‌دهند. در دو نمودار اول توابع برای توزیع نرمال و در دو نمودار دوم برای توزیع کوشی رسم شده است. در هر جفت نمودار نمودار اول نمودار یکی از نمونه‌هایی است که بازه اطمینان تولید شده توسط آن شامل تابع توزیع تجمعی جامعه است و نمودار دوم نمودار یکی از مواردی است که حداقل در یک نقطه بازه اطمینان تولید شده توسط نمونه‌ها شامل تابع توزیع تجمعی نیست. در حالتی که از توزیع نرمال استفاده شده است در ۹۰,۷ درصد موارد و در حالتی که از توزیع کوشی استفاده شده است در ۹۱,۶ درصد موارد بازه اطمینان شامل تابع توزیع تجمعی بوده است. چون بازه اطمینان ۹۰ درصد تولید شده است طبق انتظار هر دو مقدار نزدیک به ۹۰ درصد هستند.





توزیع کوشی:





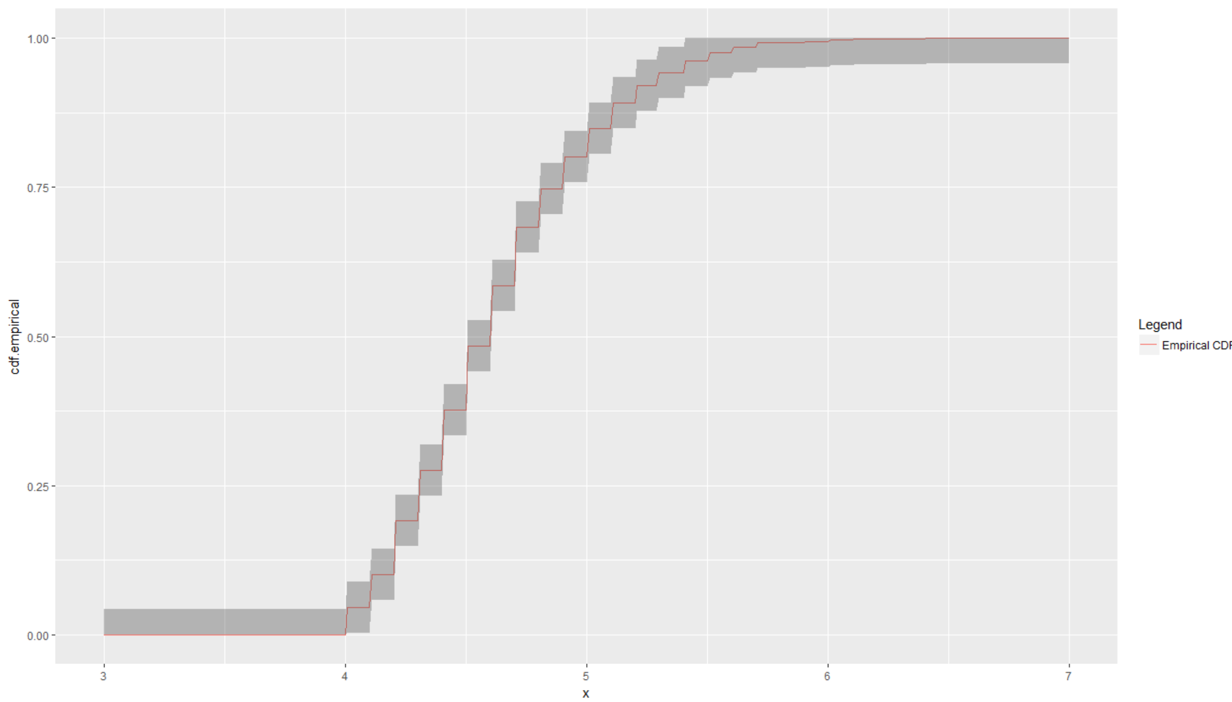
آزمایش ۳

```
[1] "Eruption mean is 3.487783 and standard error is 0.070551."
[1] "90% normal confidence interval for eruption mean is (3.371737, 3.603829)."
[1] "Eruption median is 4.000000 and standard error is 0.080501."
```

نتایج آزمایش به شکل فوق است. برای تولید خطای استاندارد از روش bootstrap استفاده شده است.

آزمایش ۴

نمودار تابع توزیع تجمعی با بازه اطمینان ۹۵ درصد به شکل زیر است. پله‌پله بودن نمودار به این دلیل است که شدت زمین‌لرزه‌ها با دقت ۰,۱ ثبت شده‌اند.



بازه‌های اطمینان حاصل شده از روش‌های مختلف برای $F(4.9) - F(4.3)$ به شکل زیر است. برای تولید خطای استاندارد از روش bootstrap استفاده شده است.

```
[1] "Normal confidence interval is (0.525037, 0.588963)."
[1] "Pivotal confidence interval is (0.525000, 0.590000)."
[1] "Percentile confidence interval is (0.524000, 0.589000)."
```

آزمایش ۵

بازه اطمینان ۹۵ درصدی از بازه اطمینان بر اساس توزیع نرمال به شکل زیر به دست می‌آید.

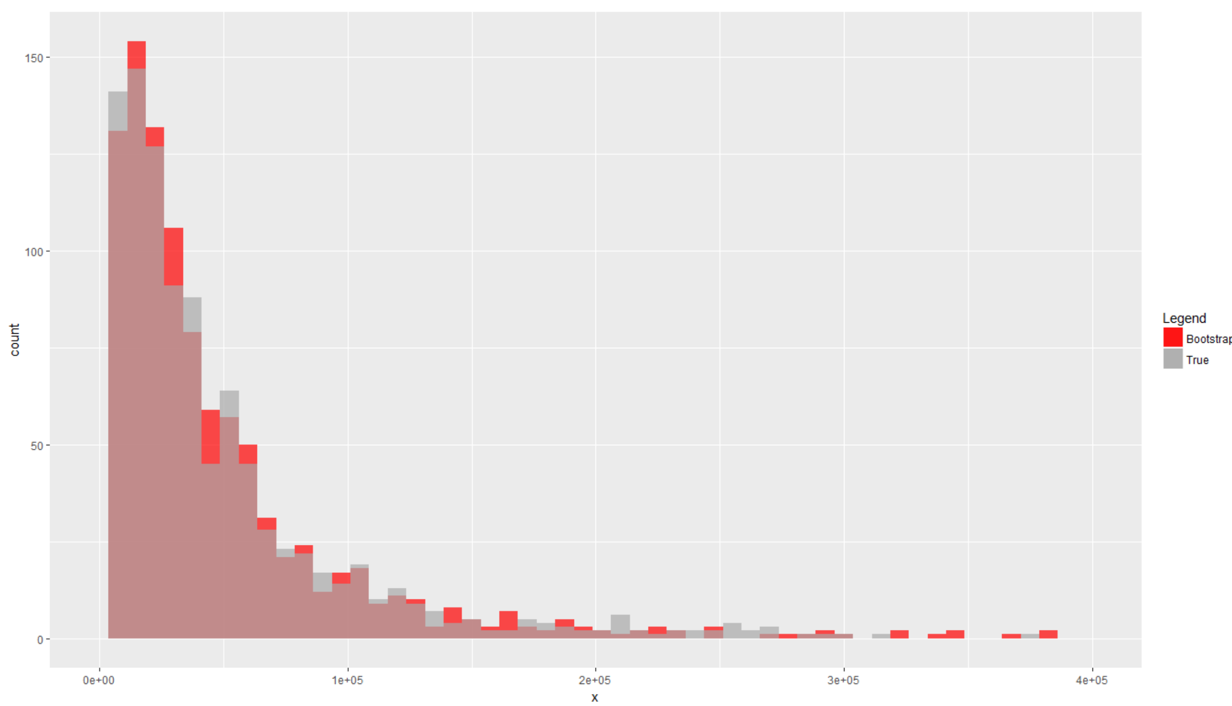
```
[1] "95% confidence interval for theta is (-94760.996938, 158422.550168)."
```

با توجه به این که این بازه اعداد منفی را نیز در بر می‌گیرد و همچنین انتظار داریم توزیعی به فرم $e^{\bar{X}}$ که Xهایش از یک توزیع نرمال به دست می‌آید، توزیع نرمال نباشد می‌توان نتیجه گرفت که این نوع بازه اطمینان مناسب نیست! به همین دلیل بازه‌های اطمینان زیر نیز گزارش می‌شود.

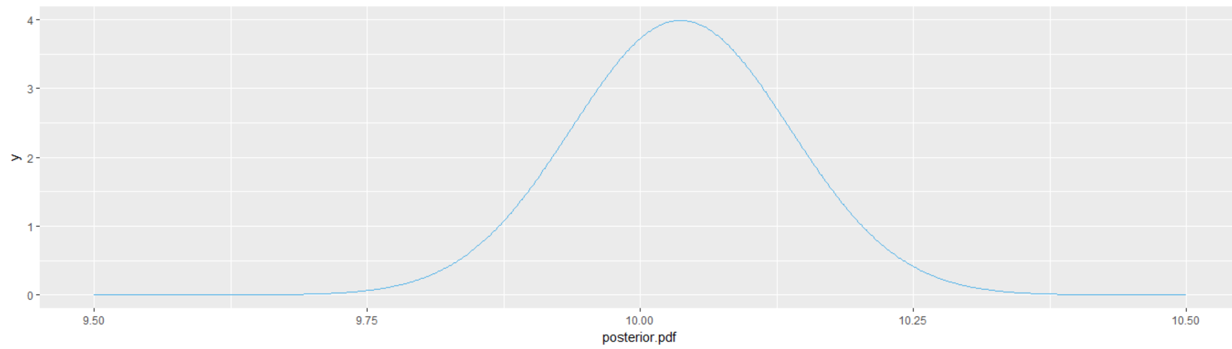
```
[1] "Pivotal confidence interval is (-137341.641570, 59085.724899)."
[1] "Percentile confidence interval is (4575.828331, 201003.194800)."
```

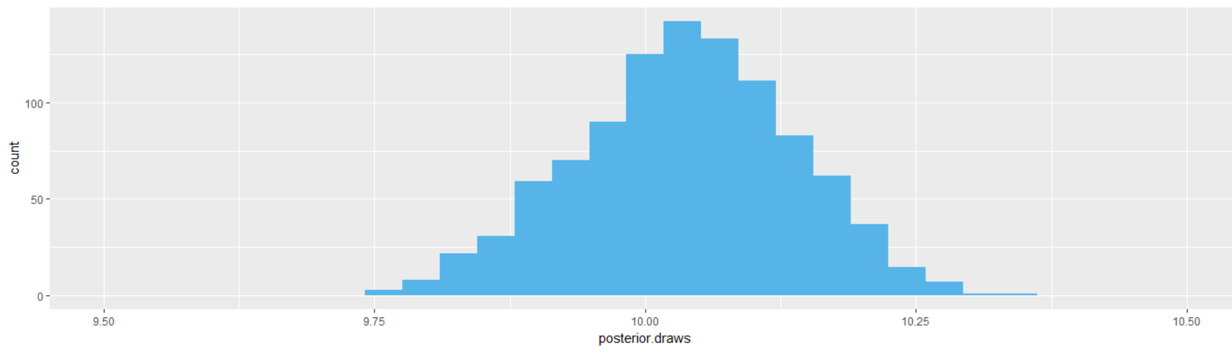
همانگونه که مشاهده می‌شود بازه اطمینان pivotal هم با استدلالی مشابه قسمت اول استدلال فوق به خوبی عمل نمی‌کند و فقط بازه اطمینان percentile است که خروجی مطابق انتظار ارائه می‌دهد.

نمودارهای خواسته شده به ترتیب زیر هستند و شباهت قابل قبولی بین آن‌ها مشاهده می‌شود.



آزمایش ۶





می‌دانیم که برای prior به شکل $f(\mu) = 1$ توزیع posterior به شکل $f(\mu|x_i) \sim \mathcal{N}\left(\bar{X}, \frac{\sigma^2}{n}\right)$ خواهد بود. نمودارهای بالا به ترتیب تابع چگالی احتمال posterior با رابطه بالا و هیستوگرام نمونه‌های تولید شده از روی آن هستند و کاملا مشهود است که رفتار مشابهی دارند.

حالا تابع چگالی احتمال $\theta = e^{\mu}$ را محاسبه می‌کنیم.

$$F_\theta(x) = P(\theta \le x) = P(e^{\mu} \le x) = P(\mu \le \ln(x)) = \int_{-\infty}^{\ln(x)} \frac{1}{\sigma\sqrt{2\pi}} \times e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

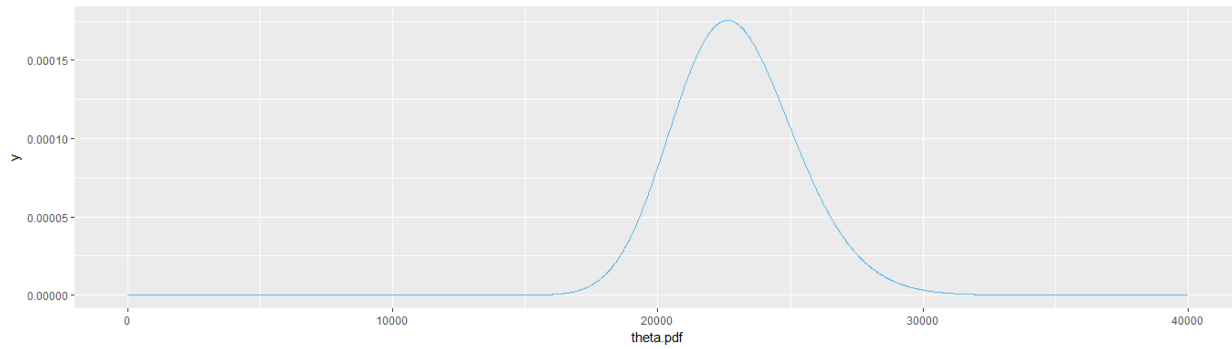
تابع فوق تابع توزیع تجمعی است. تابع چگالی احتمال مشتق تابع فوق است و به شکل زیر به دست می‌آید.

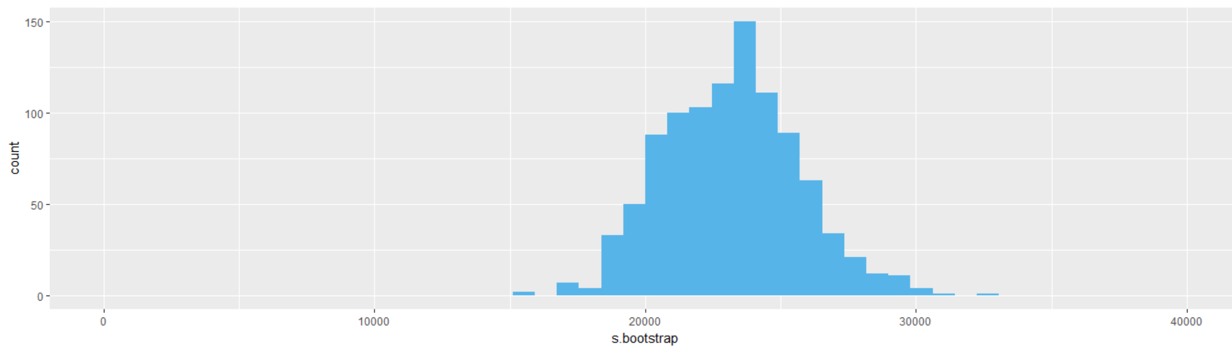
$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)} \times \frac{d\ln(x)}{dx} = \frac{1}{x\sigma\sqrt{2\pi}} \times e^{\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)}$$

و در نهایت با جایگذاری مقادیر در تابع فوق داریم:

$$f_\theta(x) = \frac{\sqrt{n}}{x\sigma\sqrt{2\pi}} \times e^{\left(-\frac{n\times(\ln(x)-\bar{X})^2}{2\sigma^2}\right)}$$

نمودارهای زیر به ترتیب تابع چگالی احتمال θ و هیستوگرام نمونه‌های تولید شده هستند.



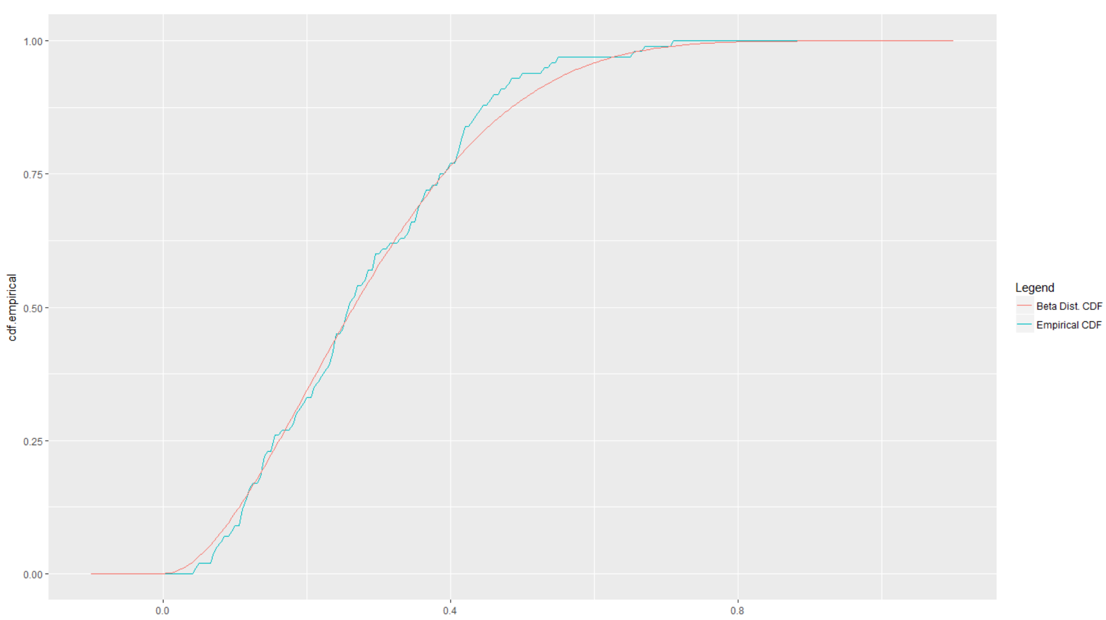


بازه‌های اطمینان به شکل زیر هستند.

```
[1] "Normal confidence interval is (18000.935402, 27742.165657)."
[1] "Pivotal confidence interval is (17431.340611, 26969.300503)."
[1] "Percentile confidence interval is (18773.800556, 28311.760447)."
```

آزمایش ۷

نمودار توزیع تجمعی تجربی به شکل زیر است.



مقادیر اصلی از روی روابط موجود برای توزیع بتا و مقادیر آلفا و بتا به دست آمده‌اند و مقادیر تخمین زده شده از روی نمونه‌ها محاسبه شده و به شکل زیر هستند:

```
[1] "Mean is 0.285714 and estimate of it is 0.279976."
[1] "Variance is 0.025510 and estimates of it are 0.021132 and 0.021345."
[1] "Skewness is 0.596285 and estimate of it is 0.556231."
```