گزارش درس یادگیری ماشین آماری

«آزمایش کامپیوتری اول»

گردآورنده: سعید دادخواه (۹۲۳۱۰۶۶)

استاد: دکتر نیک‌آبادی

آبان 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 و در آزمایش شماره ۱۴ که باید متغیرهای تصادفی با correlation تولید می‌کردیم از کتابخانه MASS 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.

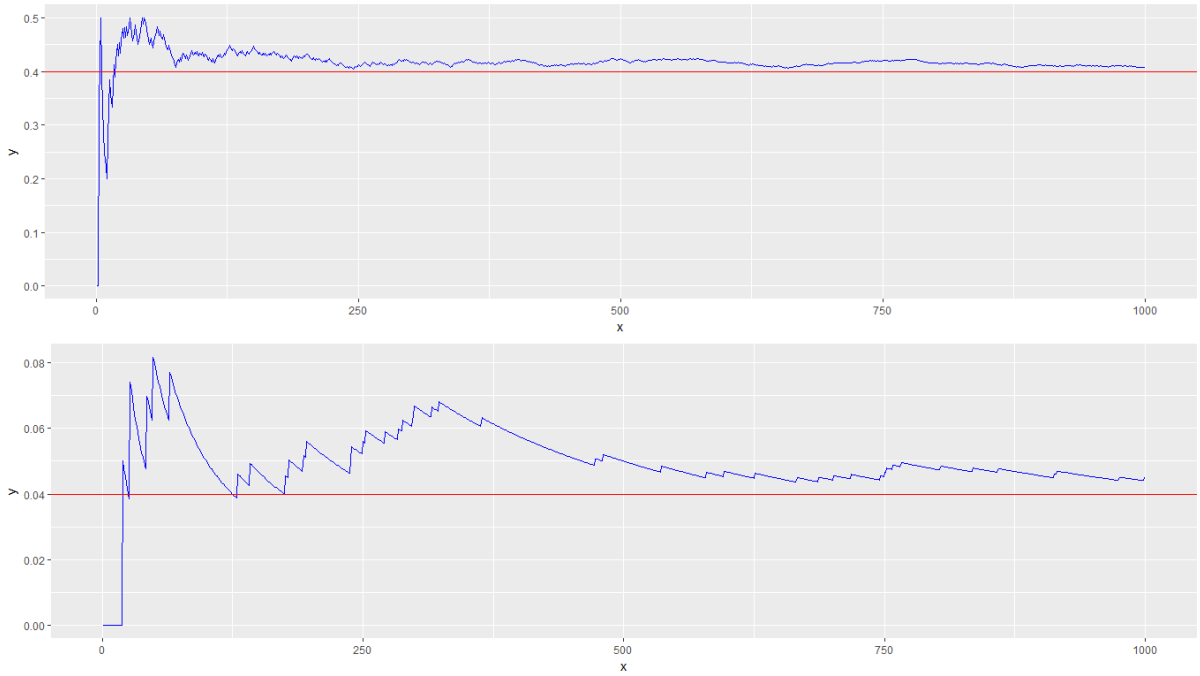
پیشنیازهای اجرای کدها

- زبان R
- کتابخانه‌های مورد نیاز
  - ggplot2
  - gridExtra
  - MASS

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

تصویر زیر نمونه‌ای از خروجی آزمایش است.

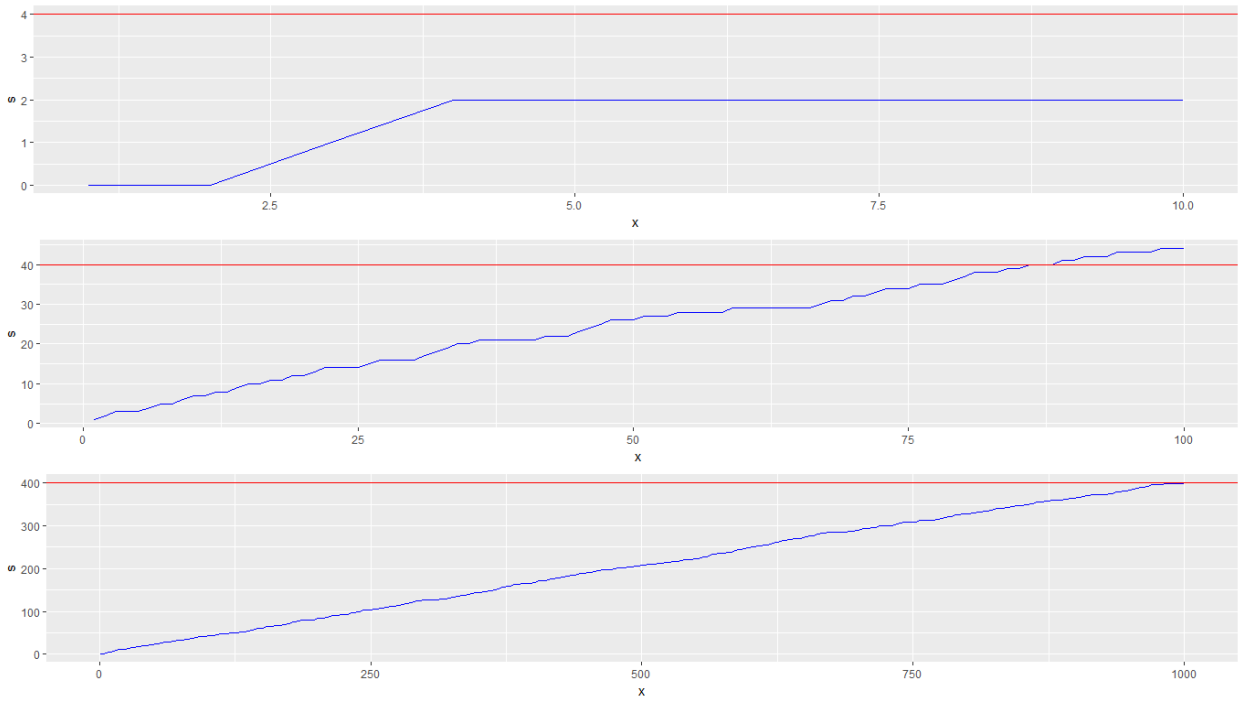


اگر آزمایش پرتاب سکه را انجام دهیم و میانگین تعداد رو آمدن را حساب کنیم به نمودارهای فوق می‌رسیم. همانطور که مشاهده می‌شود در ابتدا ممکن است فاصله میانگین با p در نظر گرفته شده زیاد باشد ولی با افزایش تعداد آزمایش‌ها میانگین نمونه‌ها به p نزدیک می‌شود. تئوری The Weak Law of Large Numbers بیان کننده همین قضیه است.

آزمایش ۲

با همان استدلال آزمایش اول با این تفاوت که طرفین را در تعداد تکرار آزمایش یعنی n ضرب کنیم می‌توان انتظار داشت توضیحات آزمایش اتفاق بیافتد یعنی مقدار X به مقدار np نزدیک شود.

تصویر زیر نمونه‌ای از اجرای آزمایش است.



آزمایش ۳

با توجه به استدلال لزوم تعویض درب انتخابی در مسئله Monty Hall و با توجه به این که اگر درب را تعویض نکنیم احتمال برابر $\frac{1}{3}$ و اگر تعویض کنیم برابر $\frac{2}{3}$ خواهد بود. خروجی زیر نمونه‌ای از اجرای آزمایش است.

```
[1] "Winning probability if we change the door: 0.66700"
[1] "Winning probability if we don't change the door: 0.33800"
```

آزمایش ۴

در قسمت a از CDF استفاده می‌شود. در قسمت b از 1 – CDF استفاده می‌شود. در قسمت c از Inverse of CDF استفاده می‌شود. در قسمت d از CDF(a) – CDF(b) استفاده می‌شود. در قسمت e با استفاده از جست‌وجوی دودویی در بازه میانگین به اضافه انحراف معیار و میانگین به اضافه سه برابر انحراف معیار و CDF مقدار مناسب را پیدا می‌کنیم.

```
[1] "P(X < 9) = 0.82711"
[1] "P(X > -3) = 0.97033"
[1] "P(X > x) = 0.05 ==> x = 11.97852"
[1] "P(0 <= X < 4) = 0.28754"
[1] "P(|X - mu| > |x|) = 0.05 ==> x = 11.979815"
```

آزمایش ۵

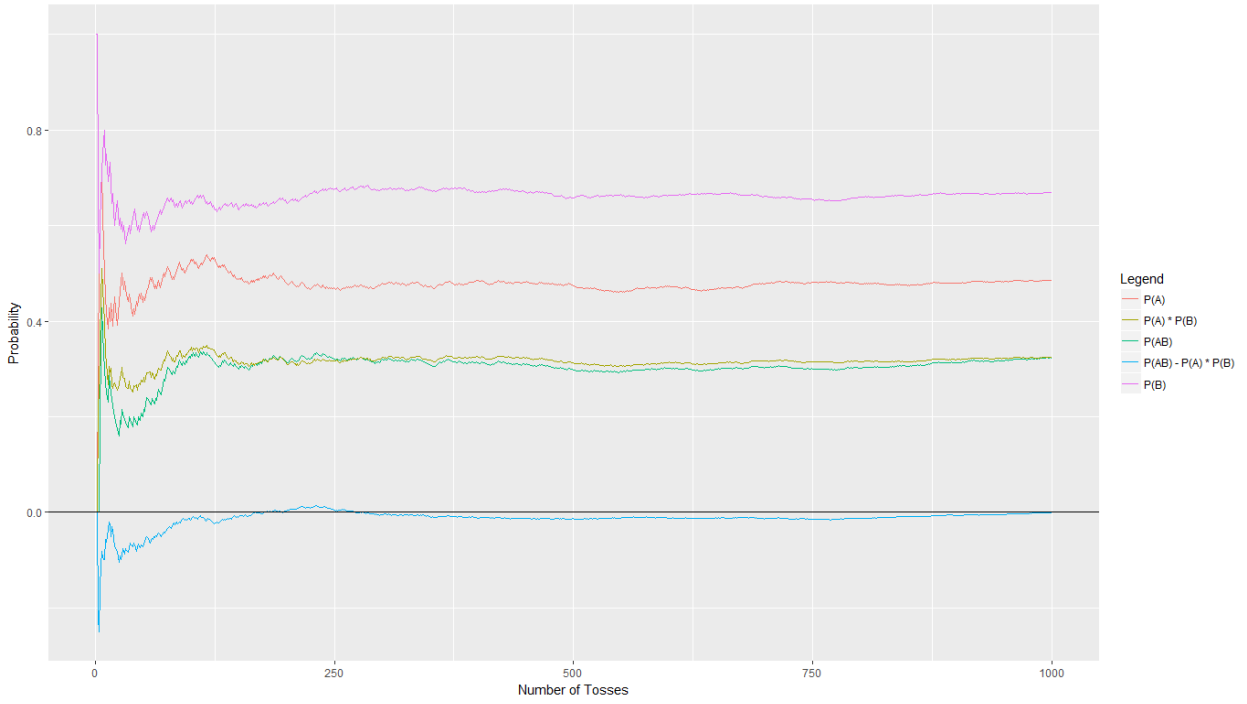
$$P(two\ or\ more\ student\ in\ January\ 1^{st}) = 1 - P(zero\ or\ one\ strudent\ in\ January\ 1^{st})$$
$$= 1 - \left[\left(\frac{364}{365}\right)^{28} + 28 \times \left(\frac{1}{365}\right) \times \left(\frac{364}{365}\right)^{27}\right]$$
$$= 1 - [0.926058745316 + 0.071235288101] = 1 - 0.997294033417$$
$$= 0.002705966583$$

برای آزمایش صدهزار بار اعدادی تصادفی از ۱ تا ۳۶۵ تولید می‌کنیم و احتمال مورد نظر را محاسبه می‌کنیم. نتیجه حاصل از انجام آزمایش به شکل زیر است.

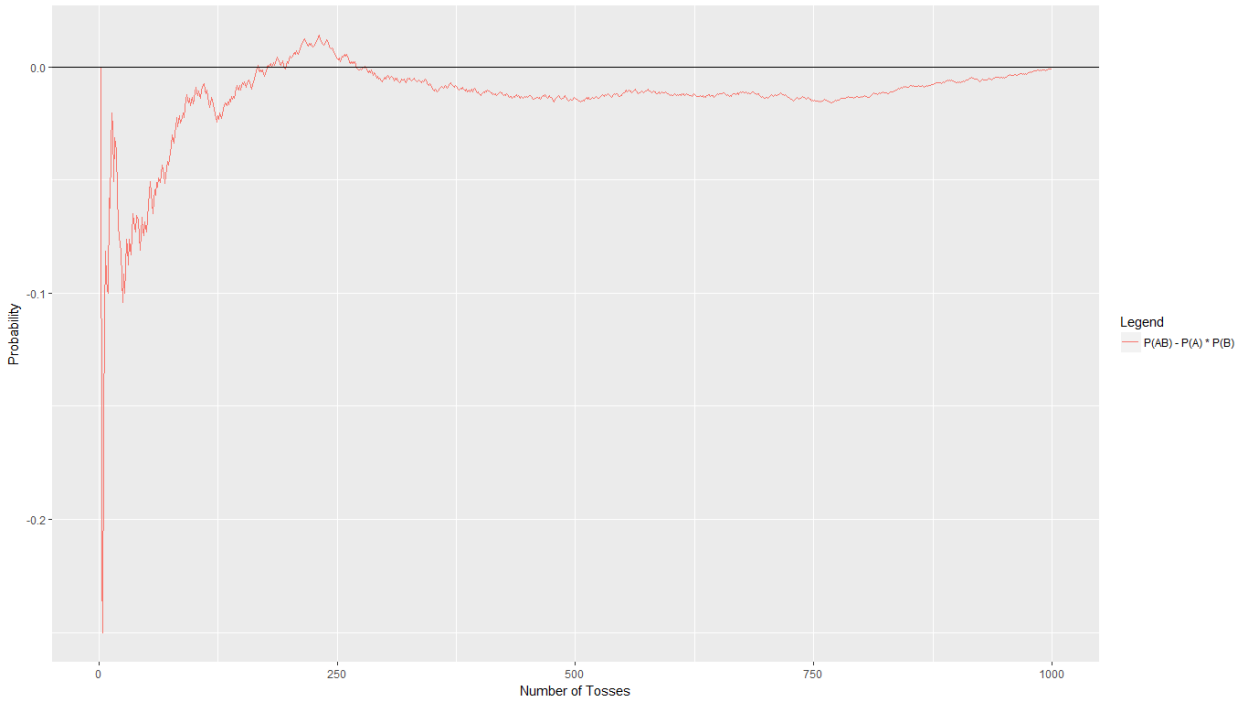
```
[1] "Total probability is 0.002770000000"
```

آزمایش ۶

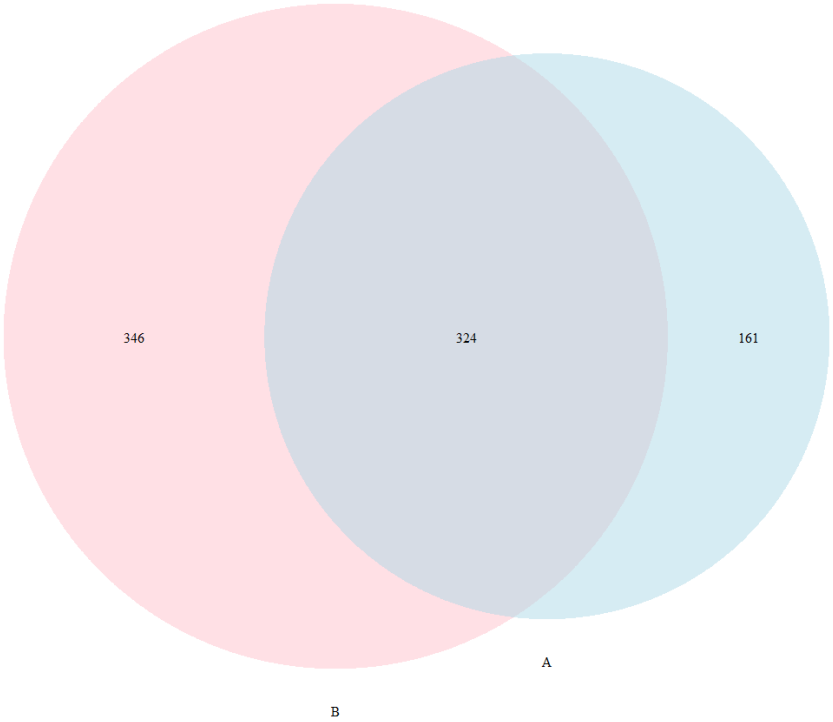
با پرتاب تاس به تعداد ۱۰۰۰ بار نتایج زیر به دست می‌آیند. همانگونه که مشاهده می‌شود با افزایش تعداد پرتاب‌ها احتمال A ضرب در احتمال B به احتمال AB نزدیک می‌شود.



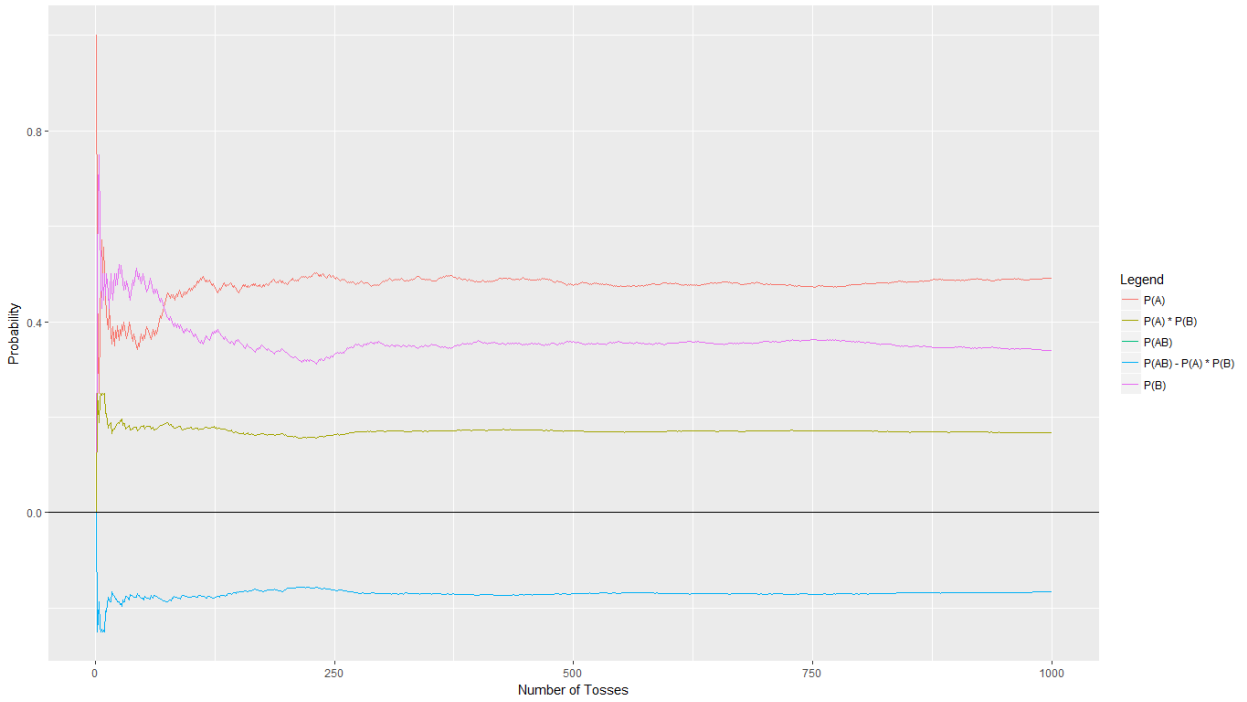
نمودار زیر نمایش دهنده فقط اختلاف P(AB) و P(A) * P(B) است تا تغییرات به شکل دقیق‌تری مشخص شوند.

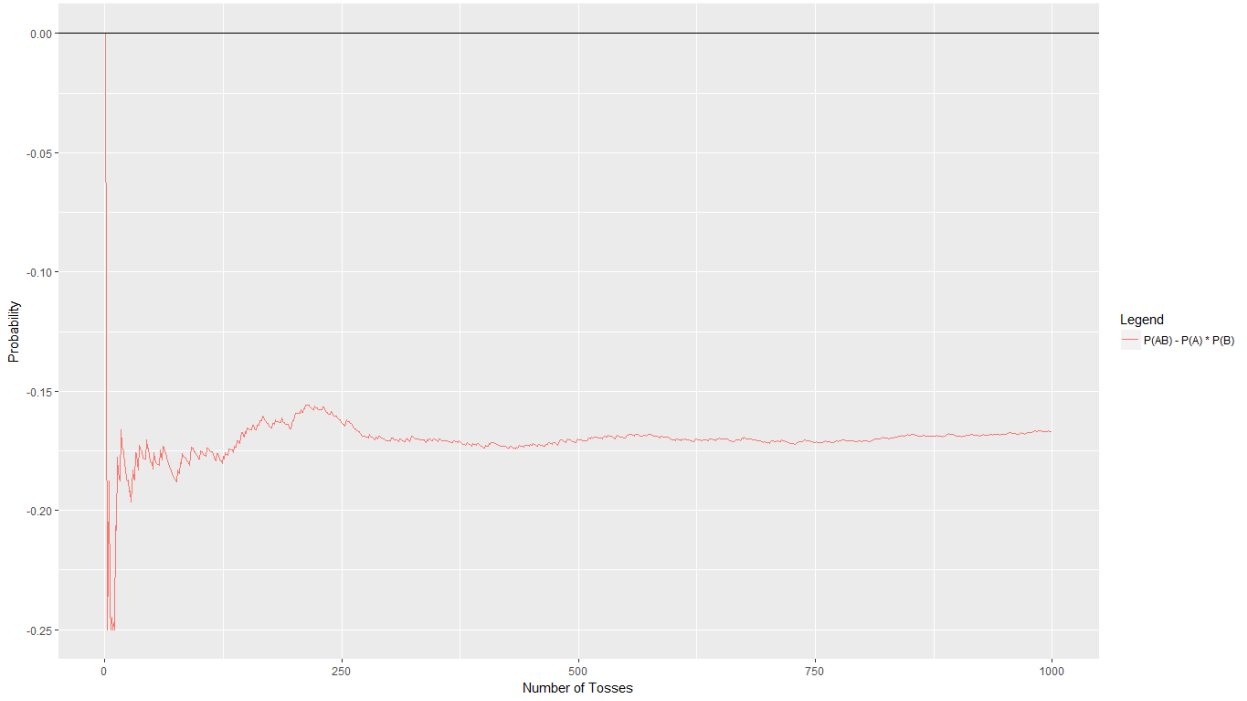


نمودار زیر نمودار ون نتیجه آزمایش را نمایش می‌دهد.



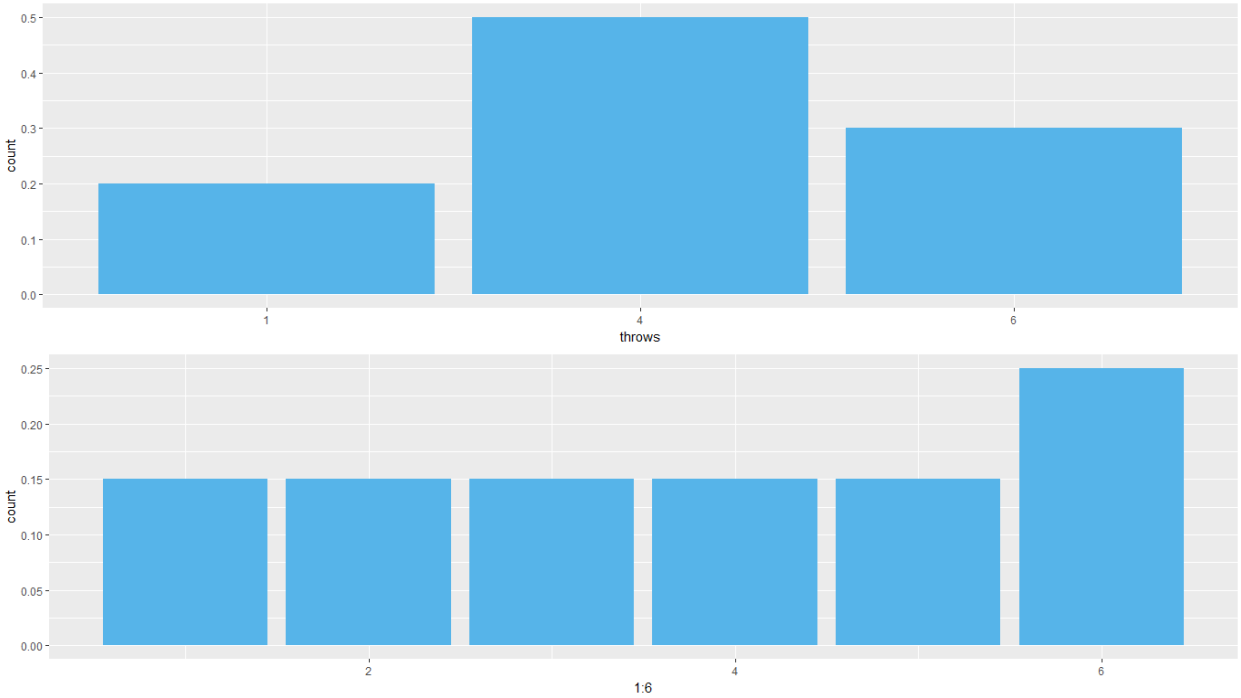
با تغییر A = {1, 2, 3} و B = {4, 6} نتایج قبل را تکرار می‌کنیم.





در نمودار ون واضح است که این دو پیشامد هیچ نقطه مشترکی ندارند.

آزمایش ۷



نمودار بالا نشان دهنده هیستوگرام پرتاب تاس و نمودار پایین نمودار جرم احتمال این تاس است. با توجه به این که فقط ده پرتاب انجام داده‌ایم می‌توانستیم انتظار داشته باشیم که این دو نمودار می‌توانند شباهتی به هم نداشته باشند که اینچنین نیز بود.

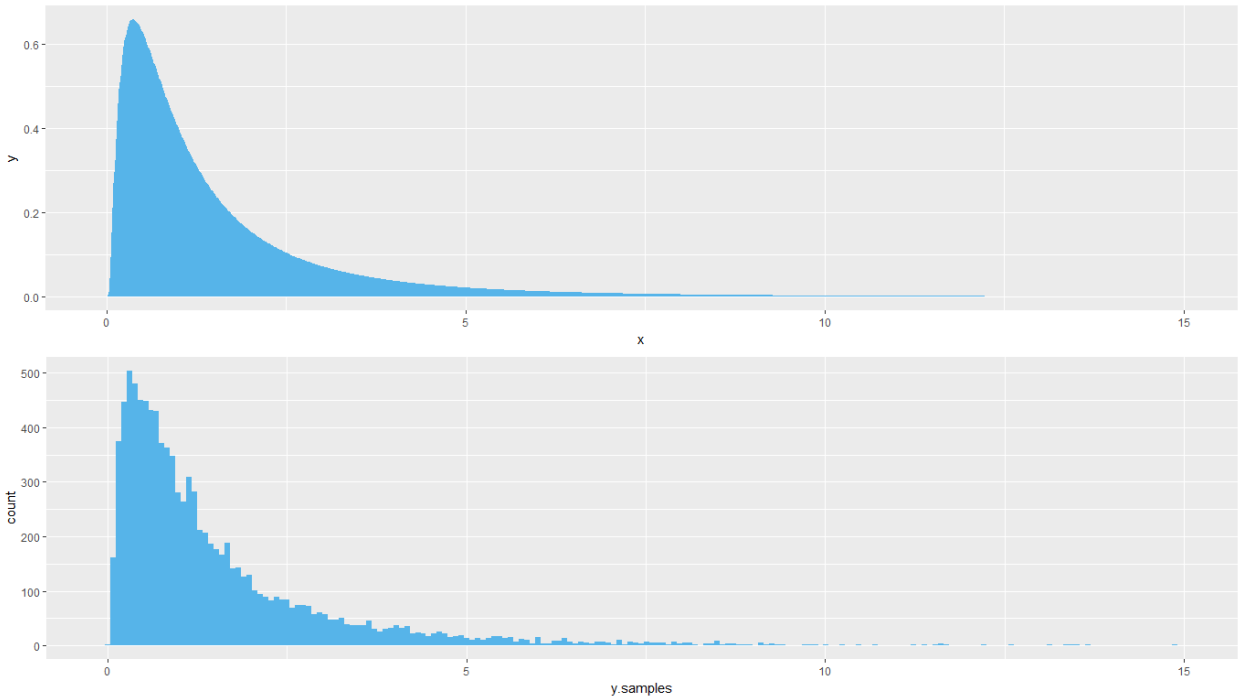
آزمایش ۸

ابتدا تابع چگالی احتمال Y را محاسبه می‌کنیم. برای yهای کمتر از صفر این مقدار همواره برابر صفر خواهد بود و برای باقی yها از طریق رابطه زیر داریم:

$$F_Y(y) = P(Y \le y) = P(e^X \le x) = P(X \le \ln(y)) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln(y) - \mu}{\sigma\sqrt{2}}\right)\right]$$

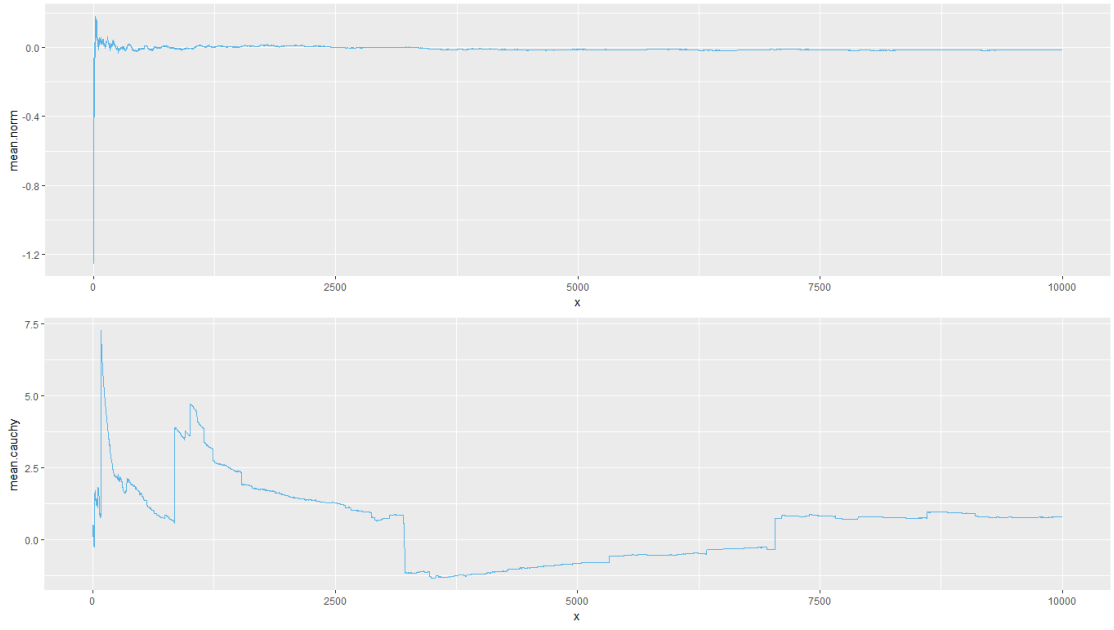
$$f_Y(y) = F_Y(y)' = \frac{1}{2}\operatorname{erf}\left(\frac{\ln(y) - \mu}{\sigma\sqrt{2}}\right)' = \frac{1}{2} \times \frac{2}{\sqrt{\pi}} e^{-\left(\frac{\ln(y)-\mu}{\sigma\sqrt{2}}\right)^2} \times \frac{1}{\sigma\sqrt{2}} \times \frac{1}{y} = \frac{1}{y\sqrt{2\pi}} e^{-\left(\frac{\ln(y)}{\sqrt{2}}\right)^2}$$
$$= \frac{1}{y\sqrt{2\pi}} e^{-\frac{\ln^2(y)}{2}}$$

با رسم تابع فوق و اجرای آزمایش مورد نظر به دو نمودار زیر می‌رسیم که همانگونه که مشاهده می‌شوند عملکرد تقریبا یکسانی دیده می‌شود.



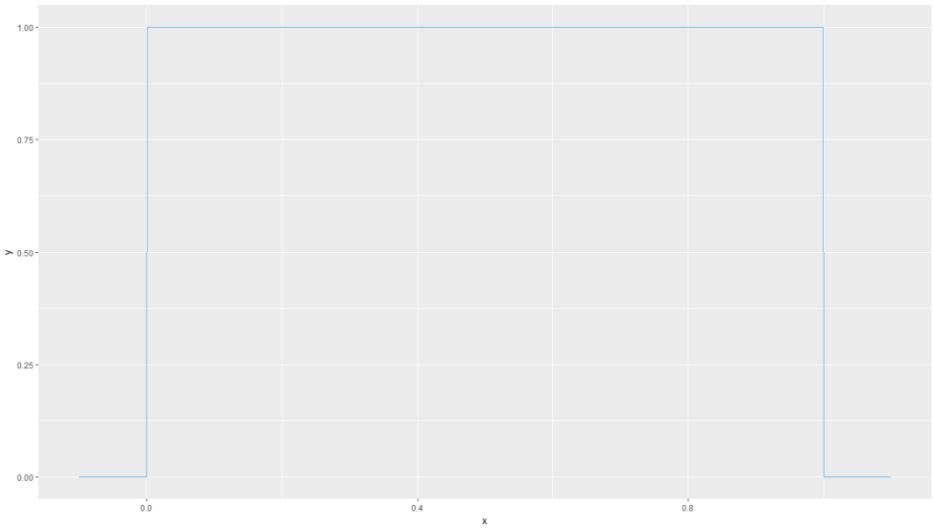
آزمایش ۹

طبق Weak Law of Large Numbers میانگین نمونه‌ها در بی‌نهایت به میانگین جامعه همگرا می‌شود که همان امید ریاضی یک متغیر تصادفی از همان جامعه است. با توجه به این که امید ریاضی برای یک متغیر تصادفی با توزیع کوشی (برخلاف یک متغیر تصادفی با توزیع نرمال) موجود نیست احتمال داده‌های بسیار بزرگ نیز وجود دارد به همین دلیل میانگین تعدادی از نمونه‌ها از چنین جامعه‌ای هیچگاه به عدد خاصی همگرا نخواهد شد.



آزمایش ۱۰

ابتدا نمودار تابع چگالی احتمال متغیر تصادفی را رسم می‌کنیم.



سپس مقادیر زیر را محاسبه می‌کنیم.

$$E(X) = \int_0^1 x \times f_X(x)dx = \int_0^1 xdx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

$$E(X^2) \int_0^1 x^2 \times f_X(x)dx = \int_0^1 x^2dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

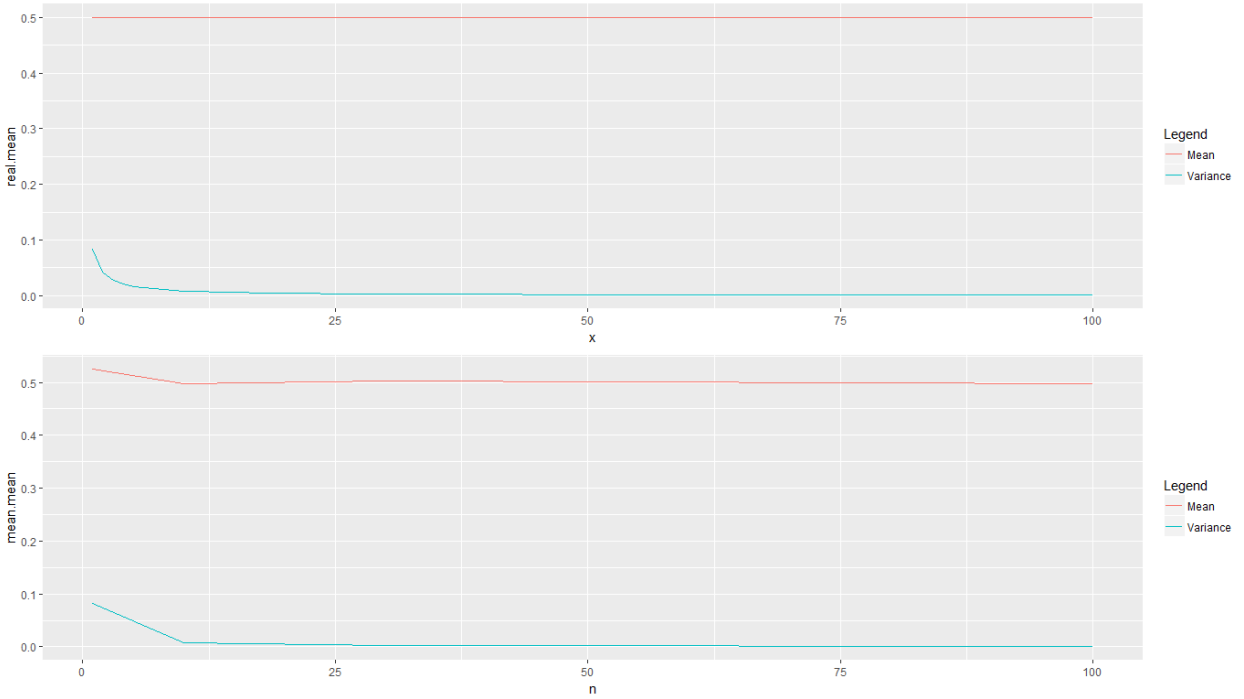
$$\sigma^2 = E(X^2) - E(X)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

روابط زیر را نیز داریم:

$$E(\bar{X}) = \mu = E(X) = \frac{1}{2}$$

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{12 \times n}$$

حال توابع فوق را به ازای هرکدام از حالت‌های ۱، ۱۰، ۳۰ و ۱۰۰ رسم می‌کنیم. سپس برای هرکدام از حالت‌ها آزمایش را ۱۰۰ بار اجرا می‌کنیم و میانگین و واریانس میانگین نمونه‌ها را رسم می‌کنیم. نتایج حاصل به شکل زیر هستند که عملکردی شبیه هم دارند.



آزمایش ۱۱

خروجی به ترتیب زیر خواهد شد.

```
[1] "Sample mean is 5.98100"
```

مشاهده می‌شود که این مقدار به ۶ نزدیک است. در کل می‌توان توزیع پواسونی با پارامتر λ را با توزیع دو جمله‌ای با تعداد n و احتمال $\frac{\lambda}{n}$ تخمین زد. تخمین زدن به این صورت است که فرض می‌کنیم عددی که از توزیع پواسون به دست می‌آید حداکثر n است. یعنی در این آزمایش فرض شده است که توزیع پواسون فقط اعداد زیر ۱۰۰ را تولید می‌کند.

آزمایش ۱۲

در این آزمایش مقدار $\sqrt{X^2 + Y^2}$ مورد نظر است. متاسفانه توزیعی برای تحلیل تئوری این سوال پیدا نکردم. توزیع کای برای چنین حالتی تعریف شده است ولی باید همه‌ی متغیرها استاندارد نرمال باشند. توزیع کای غیرمرکزی هم محدودیت انحراف معیار برابر ۱ را دارد. ظاهرا توزیعی به فرم کای غیر استاندارد باید برای تحلیل این آزمایش استفاده شود.

با تولید ۱۰۰۰ نمونه به نتیجه زیر می‌رسیم.

```
[1] "Average distance for (X, Y) where X~N(0.5, 1) and Y~N(0, 4) is 2.00852"
```

آزمایش ۱۴

با توجه به صورت سوال میانیگین متغیرها صفر است و Covariance نیز به شکل زیر محاسبه می‌شود.

$$cov(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{2} - 0 \times 0 = 0.5$$

با تولید ۱۰۰۰ نمونه به نتیجه زیر می‌رسیم که همانطور که مشاهده می‌شود به مقادیر تئوری نزدیک است.

```
[1] "The sample mean of X is 0.00621"
[1] "The sample mean of Y is -0.00259"
[1] "The covariance of X and Y is 0.49495"
```

آزمایش ۱۵

همین محدوده با توجه به نامساوی Hoeffding ۰,۰۰۰۶۷ است، چون تعداد آزمایش‌ها برابر ۱۰۰۰ تعیین شده است و نمی‌تواند برای نمایش ۰,۰۰۰۶۷ موثر باشد، برای بررسی نتایج از ۱۰۰۰۰ آزمایش استفاده خواهد شد. نتایج به شکل زیر هستند.

```
[1] "For p = 0.3 and n = 100, P = 0.0000000000 < 0.05250 = Chebyshev boundary."
[1] "For p = 0.3 and n = 200, P = 0.0000000000 < 0.02625 = Chebyshev boundary."
[1] "For p = 0.3 and n = 300, P = 0.0000000000 < 0.01750 = Chebyshev boundary."
[1] "For p = 0.3 and n = 400, P = 0.0000000000 < 0.01312 = Chebyshev boundary."
[1] "For p = 0.3 and n = 500, P = 0.0000000000 < 0.01050 = Chebyshev boundary."
[1] "For p = 0.3 and n = 600, P = 0.0000000000 < 0.00875 = Chebyshev boundary."
[1] "For p = 0.3 and n = 700, P = 0.0000000000 < 0.00750 = Chebyshev boundary."
[1] "For p = 0.3 and n = 800, P = 0.0000000000 < 0.00656 = Chebyshev boundary."
[1] "For p = 0.3 and n = 900, P = 0.0000000000 < 0.00583 = Chebyshev boundary."
[1] "For p = 0.3 and n = 1000, P = 0.0000000000 < 0.00525 = Chebyshev boundary."
[1] "For p = 0.5 and n = 100, P = 0.0000000000 < 0.06250 = Chebyshev boundary."
[1] "For p = 0.5 and n = 200, P = 0.0000000000 < 0.03125 = Chebyshev boundary."
[1] "For p = 0.5 and n = 300, P = 0.0000000000 < 0.02083 = Chebyshev boundary."
[1] "For p = 0.5 and n = 400, P = 0.0000000000 < 0.01562 = Chebyshev boundary."
[1] "For p = 0.5 and n = 500, P = 0.0000000000 < 0.01250 = Chebyshev boundary."
[1] "For p = 0.5 and n = 600, P = 0.0000000000 < 0.01042 = Chebyshev boundary."
[1] "For p = 0.5 and n = 700, P = 0.0000000000 < 0.00893 = Chebyshev boundary."
[1] "For p = 0.5 and n = 800, P = 0.0000000000 < 0.00781 = Chebyshev boundary."
[1] "For p = 0.5 and n = 900, P = 0.0000000000 < 0.00694 = Chebyshev boundary."
[1] "For p = 0.5 and n = 1000, P = 0.0000000000 < 0.00625 = Chebyshev boundary."
```

همانگونه که مشاهده می‌شود همه احتمال‌ها صفر شدند. آزمایش برای اعداد از ۱۰ تا ۱۰۰ با گام ۱۰ تکرار می‌کنیم و نتایج به ترتیب زیر می‌شوند.

```
[1] "For p = 0.3 and n = 10, P = 0.0781000000 < 0.52500 = Chebyshev boundary."
[1] "For p = 0.3 and n = 20, P = 0.0249000000 < 0.26250 = Chebyshev boundary."
[1] "For p = 0.3 and n = 30, P = 0.0078000000 < 0.17500 = Chebyshev boundary."
[1] "For p = 0.3 and n = 40, P = 0.0037000000 < 0.13125 = Chebyshev boundary."
[1] "For p = 0.3 and n = 50, P = 0.0009000000 < 0.10500 = Chebyshev boundary."
```

```
[1] "For p = 0.3 and n = 60, P = 0.0004000000 < 0.08750 = Chebyshev boundary."
[1] "For p = 0.3 and n = 70, P = 0.0001000000 < 0.07500 = Chebyshev boundary."
[1] "For p = 0.3 and n = 80, P = 0.0002000000 < 0.06562 = Chebyshev boundary."
[1] "For p = 0.3 and n = 90, P = 0.0000000000 < 0.05833 = Chebyshev boundary."
[1] "For p = 0.3 and n = 100, P = 0.0000000000 < 0.05250 = Chebyshev boundary."
[1] "For p = 0.5 and n = 10, P = 0.1121000000 < 0.62500 = Chebyshev boundary."
[1] "For p = 0.5 and n = 20, P = 0.0417000000 < 0.31250 = Chebyshev boundary."
[1] "For p = 0.5 and n = 30, P = 0.0156000000 < 0.20833 = Chebyshev boundary."
[1] "For p = 0.5 and n = 40, P = 0.0077000000 < 0.15625 = Chebyshev boundary."
[1] "For p = 0.5 and n = 50, P = 0.0037000000 < 0.12500 = Chebyshev boundary."
[1] "For p = 0.5 and n = 60, P = 0.0009000000 < 0.10417 = Chebyshev boundary."
[1] "For p = 0.5 and n = 70, P = 0.0005000000 < 0.08929 = Chebyshev boundary."
[1] "For p = 0.5 and n = 80, P = 0.0004000000 < 0.07812 = Chebyshev boundary."
[1] "For p = 0.5 and n = 90, P = 0.0002000000 < 0.06944 = Chebyshev boundary."
[1] "For p = 0.5 and n = 100, P = 0.0001000000 < 0.06250 = Chebyshev boundary."
```

آزمایش ۱۶

می‌دانیم که برای مجموع دو متغیر تصادفی دوجمله‌ای، n متغیر تولید شده برابر جمع n دو متغیر دیگر است. برای محاسبه p برای هرکدام از متغیرهای X1 + X2 و X2 + X3 از امید ریاضی آن‌ها استفاده می‌کنیم.

$$\left.\begin{array}{c} E[X1+X2]=E[X1]+E[X2]=n_1\times p_1+n_2\times p_2=1000\times 0.3+1000\times 0.5=800 \\ E[X1+X2]=(n_1+n_2)p_{12}=2000\times p_{12} \end{array}\right\} \Longrightarrow 2000\times p_{12}=800\Longrightarrow p_{12}=0.4$$

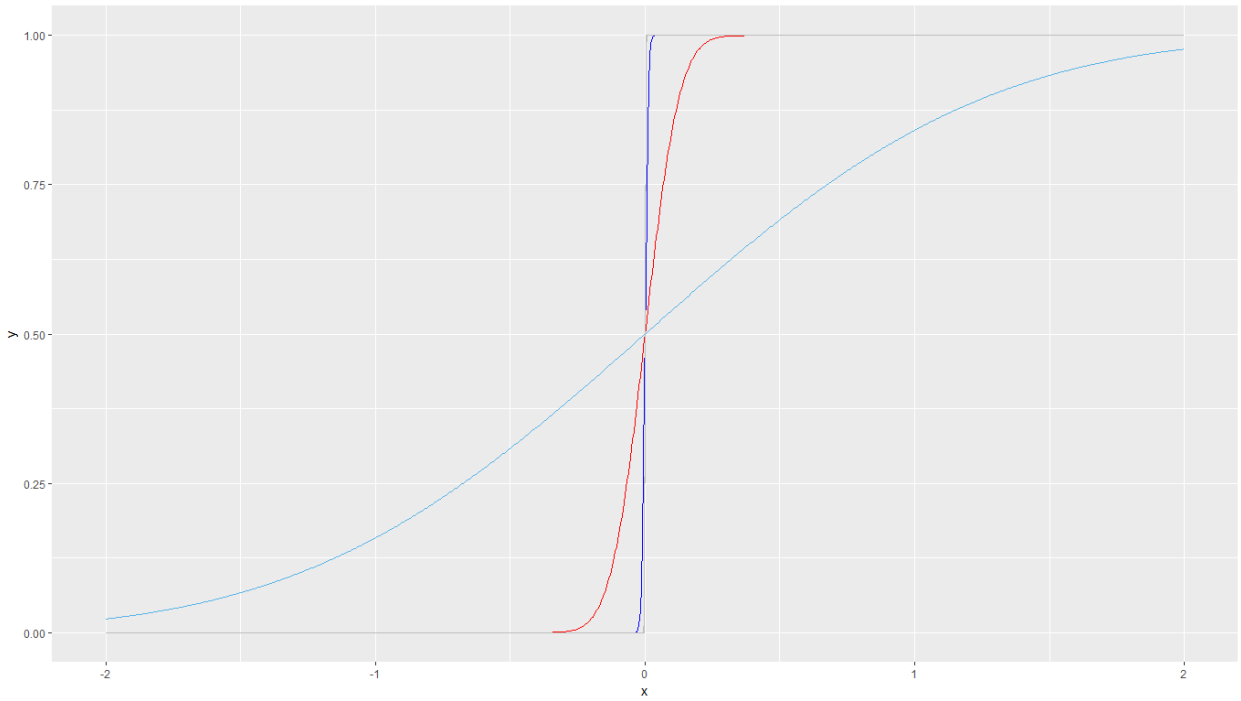
$$\left.\begin{array}{c} E[X2+X3]=E[X2]+E[X3]=n_2\times p_2+n_3\times p_3=1000\times 0.5+2000\times 0.5=1500 \\ E[X2+X3]=(n_2+n_3)p_{23}=3000\times p_{23} \end{array}\right\} \Longrightarrow 3000\times p_{23}=1500\Longrightarrow p_{23}=0.5$$

نتیجه آزمایش به شکل زیر است.

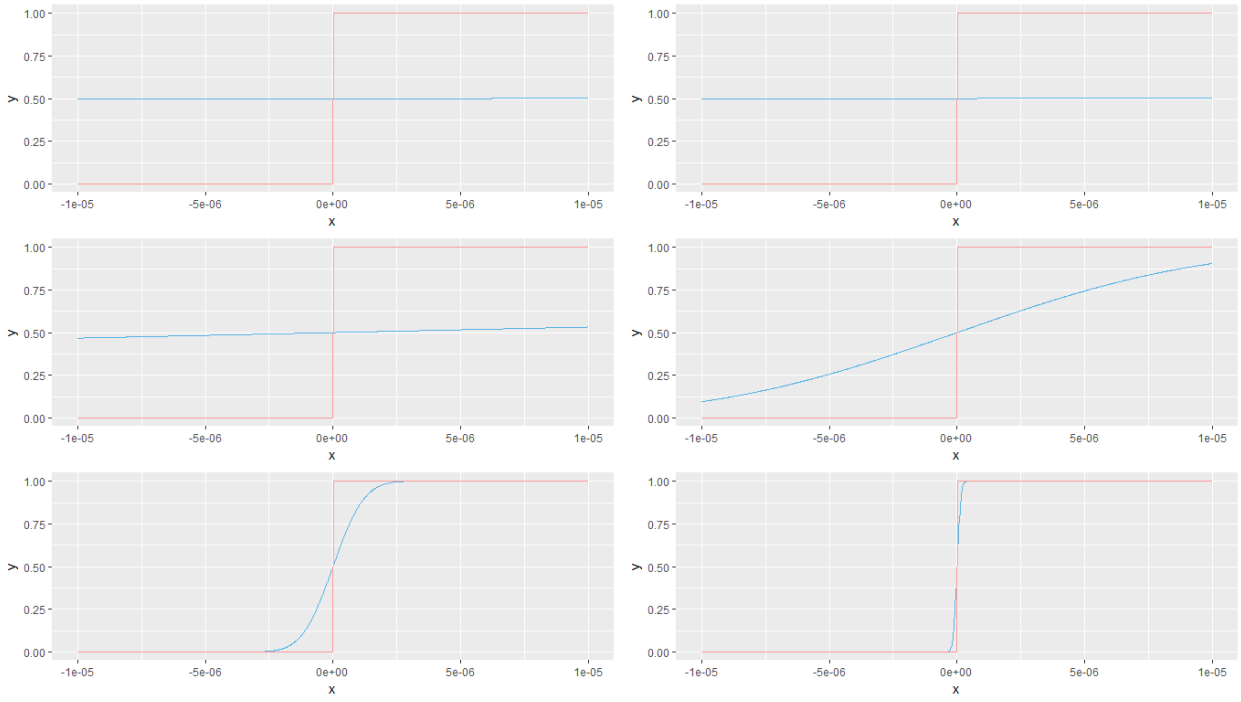
```
[1] "The sample mean of X1 + X2 for 10000 samples is 799.83370 then p is 0.4"
[1] "The sample mean of X2 + X3 for 10000 samples is 1499.98060 then p is 0.5"
```

آزمایش ۱۷

نمودار تابع توزیع تجمعی به شکل زیر درمی‌آید. طبق شکل نمودارهایی که مقدار i بیشتری دارند به محور عمودی نزدیک‌تر شده‌اند.



در تصویر زیر CDF توزیع مورد نظر را به ازای اپسیلون‌های ۰,۱، ۰,۰۱، ۰,۰۰۱، ۰,۰۰۰۱، ۰,۰۰۰۰۱ و ۰,۰۰۰۰۰۱ مشاهده می‌کنید و همانگونه که مشخص است با کاهش اپسیلون CDF توزیع به CDF صفر نزدیک می‌شود.



برای نمایش این مرحله در ابتدا i برابر یک قرار داده شده است. در هر مرحله $P(|X_i - X| > \varepsilon)$ محاسبه شده است و تا زمانی که این مقدار مخالف صفر باشد i نصف و مرحله بعد به همین ترتیب تکرار شده است. با توجه به

تعریف همگرایی در احتمال دو متغیر میانگین نمونه و $P(|X_i - X| > \varepsilon)$ نسبت به مراحل رسم شده‌اند. طبق مشاهدات میانگین نمونه با افزایش تعداد مراحل به صفر میل می‌کند و همچنین با توجه به شرط اجرای مراحل احتمال فوق نیز در نهایت صفر می‌شود. با کاهش اپسیلون فقط تعداد مراحل زیاد می‌شود ولی در نهایت شرط خاتمه رخ می‌دهد.