آمار و احتمال مهندسی

قضیه حد مرکزی

(Ross 8.1-8.4)

نامساوی‌های احتمالاتی

- در بسیاری از مواقع ما شکل صحیح توزیع احتمال را نمی‌دانیم.
  - مثال: نمرات میان‌ترم
- اما از میانگین توزیع اطلاع داریم.
- ممکن است خواص و پارامترهای دیگری از این توزیع مانند واریانس، نامنفی بودن و ... را نیز بدانیم.
- نامساوی‌ها و حدهای احتمالاتی به ما کمک می‌کنند تا بتوانیم گزاره‌هایی در مورد توزیع احتمال در این حالات بیان کنیم.
  - این گزاره‌ها ممکن است در مقایسه با توزیع صحیح نادقیق باشند.

نامساوی مارکوف (Markov Inequality)

Andrey Markov
(1856-1922)

قضیه. اگر X یک متغیر تصادفی مثبت باشد (یعنی داشته باشیم $\mathrm{P}\{X<0\}=0$ یا $f_X(x)=0 : x<0$)، و $\alpha>0$ یک ثابت دلخواه باشد، داریم:

$$\mathrm{P}\{X \geq \alpha\} \leq \frac{\mathrm{E}(X)}{\alpha}$$

- مثلا اگر $\mathrm{E}(X)=1$ باشد، توزیع X هر چه باشد، حتماً داریم:
$P\{X \geq 100\} \leq 0.01$

- نامساوی مارکوف معمولا برای αهایی که نسبت به $\mathrm{E}(X)$ بزرگ باشند، استفاده می‌شود.

نامساوی مارکوف

اثبات:

$$\mathrm{E}(X)=\int_0^{+\infty} x\, f_X(x)\mathrm{d}x \geq \int_\alpha^{+\infty} x\, f_X(x)\mathrm{d}x \geq \int_\alpha^{+\infty} \alpha\, f_X(x)\mathrm{d}x$$

$$\Rightarrow \mathrm{E}(X) \geq \alpha\, \mathrm{P}\{X \geq \alpha\} \quad \Rightarrow \mathrm{P}\{X \geq \alpha\} \leq \frac{\mathrm{E}(X)}{\alpha}$$

مثال: اگر میانگین نمرات آمار ۱۲ باشد، درباره احتمال این که نمره فردی در این کلاس بالای ۱۵ باشد، چه می‌توان گفت؟

$E[X]=12 \;\;,\; P(X \geq 15)=?$

$$P(X \geq 15) \leq \frac{12}{15}=0.8$$

- احتمال نمره بالای ۱۵ گرفتن کمتر از ۸۰٪ است.

نامساوی Bienayme

- نامساوی Bienayme:

$$P\{|X-a| \geq \epsilon\} \leq \frac{E(|X-a|^n)}{\epsilon^n}$$

- این نامساوی نتیجه مستقیم نامساوی مارکوف است:

$$P\{|X-a| \geq \epsilon\} = P\{|X-a|^n \geq \epsilon^n\} \leq \frac{E(|X-a|^n)}{\epsilon^n}$$

نامساوی چبیشف (Chebychev's Inequality)

نامساوی چبیشف.

$$P\{\mu-\epsilon < X < \mu+\epsilon\} \geq 1-\frac{\sigma^2}{\epsilon^2}$$

یا به صورت دیگر:

$$P\{\mu-k\sigma < X < \mu+k\sigma\} \geq 1-\frac{1}{k^2}$$

اثبات: از نامساوی Bienayme داریم:

$$P\{|X-a| \geq \epsilon\} \leq \frac{E(|X-a|^n)}{\epsilon^n}$$

اگر بگیریم: $a=\mu$، $n=2$ و $\epsilon=k\sigma$، نتیجه می‌شود:

$$\{|X-\mu| \geq k\sigma\} \leq \frac{\sigma^2}{\epsilon^2} \Rightarrow P\{\mu-k\sigma < X < \mu+k\sigma\} \geq 1-\frac{1}{k^2}$$

مثال

- فرض کنید مقالات یک روزنامه به طور میانگین دارای ۱۰۰۰ کلمه با انحراف معیار۲۰۰ کلمه باشند. حداقل احتمال این که تعداد کلمات یک مقاله از این روزنامه بین ۶۰۰ تا ۱۴۰۰ باشد، چقدر است؟

طبق نامساوی چبیشف:

$$P\{600 < X < 1400\} = P\{|X - 1000| < 2 \times 200\} \geq 1 - \frac{1}{2^2} = 0.75$$

بنابراین حداقل ۷۵ درصد از مقالات بین ۶۰۰ تا ۱۴۰۰ کلمه خواهند داشت.

- توجه کنید که در صورتی که توزیع کلمات نرمال بود، طبق قانون تجربی ۶۸-۹۵-۹۹، حداقل ۹۵ درصد از مقالات دارای ۶۰۰ تا ۱۴۰۰ کلمه (فاصله 2σ از μ) خواهند بود.

قانون ضعیف اعداد بزرگ

قانون ضعیف اعداد بزرگ (Weak Law of Large Numbers):

- فرض کنید X_iها متغیرهای تصادفی .i.i.d با تابع توزیع انباشته F و میانگین $E[X_i] = \mu$ و واریانس $Var(X_i) = \sigma^2$ باشند.
- متوسط این متغیرهای تصادفی را به صورت $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ تعریف می‌کنیم.
- آنگاه برای هر $\epsilon > 0$ خواهیم داشت:

$$P\left(\left|\bar{X} - \mu\right| > \epsilon\right) \xrightarrow[n\to\infty]{} 0$$

- برای هر ϵ مثبت هرچند کوچک، با داشتن یک نمونه به اندازه کافی بزرگ، احتمال زیادی وجود دارد که $\bar{X}$ به μ نزدیک باشد، به عبارت دیگر در محدوده $(\mu - \epsilon, \mu + \epsilon)$ قرار گیرد.

اثبات قانون ضعیف اعداد بزرگ

- قبلا دیدیم که:

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \mu$$

$$Var(\frac{X_1 + X_2 + \cdots + X_n}{n}) = \frac{\sigma^2}{n}$$

- لذا با توجه به قضیهٔ چبیشف خواهیم داشت:

$$P\{|\bar{X} - \mu| \geq k\} \leq \frac{Var(\bar{X})}{k^2}$$

$$\Rightarrow P\{|\bar{X} - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to+\infty]{} 0$$

قانون قوی اعداد بزرگ

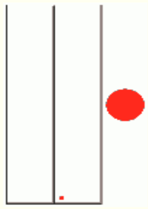
قانون قوی اعداد بزرگ (Strong Law of Large Numbers):

- فرض کنید X_iها متغیرهای تصادفی .i.i.d با تابع توزیع انباشته F و میانگین $E[X_i] = \mu$ و واریانس $Var(X_i) = \sigma^2$ باشند.
- اگر $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ متوسط این متغیرهای تصادفی باشد:

$$P\left\{\lim_{n\to\infty}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \mu\right\} = 1$$

- با میل تعداد آزمایش‌ها (یا n) به بینهایت، احتمال $\bar{X} = \mu$ برابر با یک می‌شود.

قانون اعداد بزرگ و تعبیر فرکانسی احتمال



- در ابتدای درس دیدیم که اگر در یک آزمایش تصادفی $P(A)=p$ باشد و در n بار تکرار آزمایش k بار پیشامد A اتفاق افتد، آنگاه می‌توان گفت در صورتی که n به اندازه کافی بزرگ باشد، p تقریبا برابر با k/n است.

- گزاره بالا در واقع همان قانون اعداد بزرگ برای متغیرهای تصادفی برنولی X_i (متغیرهای تصادفی شاخص پیشامد A در آزمایش i-ام) با احتمال موفقیت p :

$$\sum_{i=1}^{n} X_i = k \;\Rightarrow\; \bar{X} = \frac{k}{n}, \qquad \mu = E[X_i] = p$$

و احتمال $p = k/n$ تقریبا برابر با یک است.

مغالطه قمارباز (Gambler’s Fallacy)

- در ۱۰ پرتاب یک سکه سالم، هر ۱۰ بار شیر آمده است. احتمال این که پرتاب یازدهم هم شیر بیاید چقدر است؟

H H H H H H H H H H ?

- احتمال همچنان 0.5 است و تفاوتی با قبل ندارد.
  - به عبارت دیگر سکه حافظه ندارد.

- یک اشتباه رایج در فهم قانون اعداد بزرگ این است که تصور کنیم فرایندهای تصادفی ملزم به جبران آنچه در گذشته اتفاق افتاده هستند.
  - این مغالطه در علم آمار به مغالطه قمارباز مشهور است.

قضیه حد مرکزی (Central Limit Theorem)

- قضیه حد مرکزی: فرض کنید X_iها متغیرهای تصادفی $i.i.d.$ با تابع توزیع انباشته F و میانگین $E[X_i] = \mu$ و واریانس $Var(X_i) = \sigma^2$ باشند. آنگاه وقتی n به بینهایت میل کند:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

- شکل دقیق قضیه حد مرکزی: اگر X_iها متغیرهای تصادفی $i.i.d$ بوده و $Y = \sum_{i=1}^{n} X_i$ باشد، وقتی $n \to \infty$، به شرط محدود بودن همه گشتاورهای X_i، توزیع Y به توزیع نرمال میل می‌کند، حتی اگر X_iها نرمال نباشند.
  - قضیه حد مرکزی **لیندبرگ-لوی**: محدود بودن واریانس (گشتاور مرتبه دوم) X_iها برای نرمال بودن توزیع Y کافی است.

تقریب نرمال برای توزیع دوجمله‌ای

- **شکل جمعی قضیه حد مرکزی:** فرض کنید X_iها متغیرهای تصادفی $i.i.d.$ با میانگین $E[X_i] = \mu$ و واریانس $Var(X_i) = \sigma^2$ باشند. آنگاه وقتی n به بینهایت میل کند:

$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$

- اگر توزیع X_iها برنولی با احتمال موفقیت p باشد: $X_i \sim \mathrm{Ber}(p)$، آنگاه مجموع آنها دارای توزیع دوجمله‌ای خواهد بود:

$Y = X_1 + X_2 + \cdots + X_n \sim \mathrm{Bin}(n, p)$

$\mu = E[X_i] = p \quad , \qquad \sigma^2 = var(X_i) = p(1-p)$

$\Rightarrow Bin(n,p) \approx N(np\,, np(1-p))$

قضیه حد مرکزی

- در واقع این قضیه بیانگر خاصیتی از عملگر کانولوشن است که کانولوشن تعداد زیادی توابع مثبت به تابع گوسی میل می‌کند.
- با توجه به CLT مشخص می‌شود که چرا بسیاری از پدیده‌ها در جهان خارج توزیع تقریبا نرمال دارند.
  - بسیاری از کمیت‌ها مجموع متغیرهای تصادفی مستقل هستند: نمرات امتحان، نتایج نظرسنجی‌های انتخاباتی و ...
- اصولا هرگاه پدیده‌ای تحت تاثیر عوامل متعدد تصادفی باشد (قد یک فرد، خطا در اندازه‌گیری، ولتاژ نویز حرارتی و ...) دارای توزیع تقریبا نرمال خواهد بود.

نمونه‌برداری

- با نمونه‌برداری مکرر از یک متغیر تصادفی در یک جامعه (population)، n متغیر تصادفی *i.i.d* به دست می‌آید.
- n را اندازه نمونه (sample size) می‌گوییم.
- چون X_iها مستقل هستند، داریم:

$$f_{X_i}(x) = f_X(x)$$

$$\Rightarrow E(X_i) = E(X) = \mu$$: میانگین جامعه

$$\Rightarrow \text{var}(X_i) = \text{var}(X) = \sigma^2$$: واریانس جامعه

میانگین نمونه (Sample Mean)

- طبق تعریف، میانگین نمونه‌ها برابر است با:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

داریم:

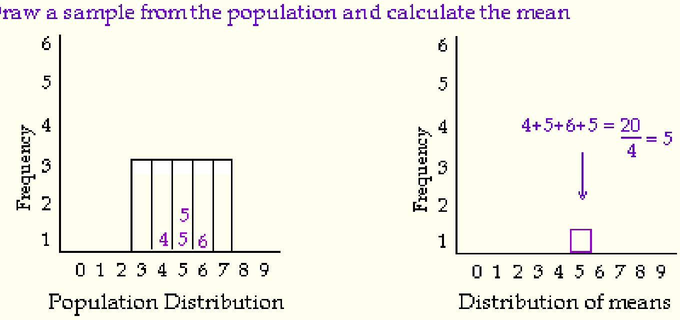
$$E(\bar{X}) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \times n\mu = \mu$$

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_{X_i}^2 = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

- پس هر چقدر n زیادتر شود، مقدار $\bar{X}$ به μ واقعی نزدیکتر خواهد بود. $\bar{X}$ میانگین نمونه است، در حالی که μ میانگین جامعه است.

مثال

Draw a sample from the population and calculate the mean



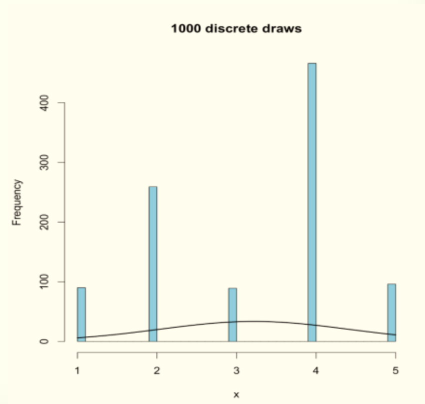
میانگین نمونه

X	**1**	**2**	**3**	**4**	**5**
$P(X)$	0.1	0.25	0.1	0.45	0.1

$E[X] = 3.2$, $\quad X_i \sim X$

Sample Size: $n = 1$

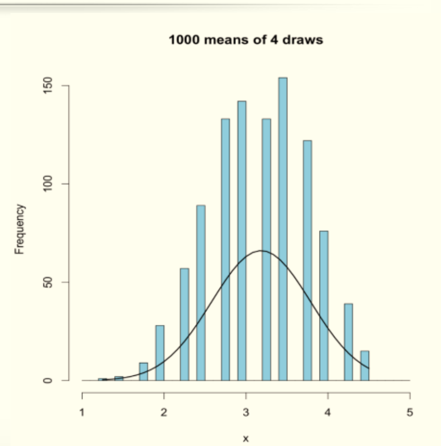
$$\bar{X} = \frac{X_1}{1}$$



میانگین نمونه

Sample Size: $n = 4$

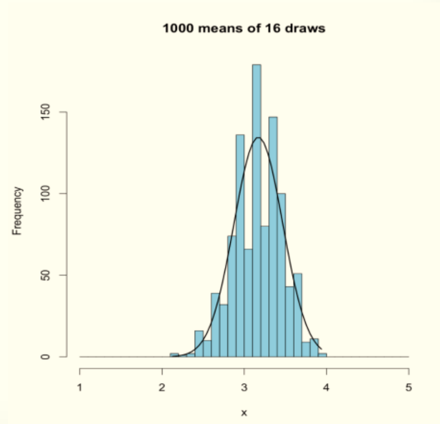
$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$



میانگین نمونه

Sample Size: $n = 16$

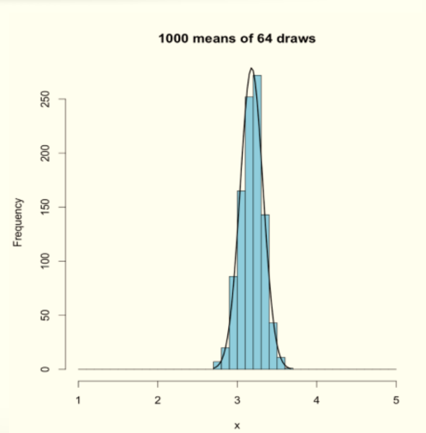
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{16}}{16}$$



میانگین نمونه

Sample Size: $n = 64$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{64}}{64}$$



به جای کل جامعه آماری مورد بحث …

می‌توان از یک نمونه بسیار کوچکتر برای کسب اطلاعات استفاده کرد

مثال ۱

- هزینه ماهیانه تلفن همراه مشترکان تهرانی دارای میانگین ۶۴ هزار تومان و انحراف معیار ۹ هزار تومان است. به طور تصادفی ۳۶ قبض تلفن همراه را انتخاب می‌کنیم. احتمال این که میانگین مبلغ هزینه این قبوض بین ۶۱ تا ۶۷ هزار تومان باشد، چقدر است؟
- از آنجایی که تعداد نمونه‌ها (۳۶) از ۳۰ بیشتر است، می‌توانیم از قضیه حد مرکزی استفاده کنیم:

$$E(\bar{X}) = 64 \ , \qquad \sigma_{\bar{X}} = \frac{9}{\sqrt{36}} = \frac{9}{6} = 1.5$$

$$P\{61 < \bar{X} < 67)\} = \mathrm{P}\{\frac{61-64}{1.5} < \frac{\bar{X}-64}{1.5} < \frac{67-64}{1.5}\}$$

$$= P\{-2 < Z < 2\} = 2G(2) - 1 = 0.9544$$

مثال ۲

- در مثال قبلی تعداد نمونه‌های انتخابی حداقل چقدر باشد تا اطمینان داشته باشیم، میانگین نمونه‌ها با احتمال بیش از ۸۴ درصد، از ۶۵ هزار تومان کمتر خواهد بود؟

$$E(\bar{X}) = 64 \ , \qquad \sigma_{\bar{X}} = \frac{9}{\sqrt{n}}$$

$$P\{\bar{X} < 65)\} = \mathrm{P}\{\frac{\bar{X}-64}{\frac{9}{\sqrt{n}}} < \frac{65-64}{\frac{9}{\sqrt{n}}}\}$$

$$= P\left\{Z < \frac{65-64}{\frac{9}{\sqrt{n}}}\right\} > 0.84$$

جدول توزیع (CDF) نرمال استاندارد

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

ادامه مثال ۲

- در نتیجه باید داشته باشیم:

$$\frac{65-64}{\frac{9}{\sqrt{n}}} \geq 1.00$$

$$\Rightarrow n \geq 9^2 = 81$$

- پس حداقل ۸۱ نمونه لازم داریم تا اطمینان داشته باشیم که میانگین آن‌ها با احتمال بیش از ۸۴ درصد کمتر از ۶۵ هزار تومان خواهد بود.

حاصلضرب متغیرهای تصادفی i.i.d.

- با استفاده از قضیه حد مرکزی دیدیم که حاصل‌جمع متغیرهای تصادفی *i.i.d.* از توزیع نرمال پیروی می‌کند.
- برای حاصلضرب متغیرهای تصادفی *i.i.d.* داریم:

$$Y = X_1 X_2 \dots X_n$$

$$\Rightarrow \log(Y) = \log(X_1 X_2 \dots X_n) = \log(X_1) + \log(X_2) + \cdots + \log(X_n)$$

- تعریف می‌کنیم: $Y_i = \log(X_i)$، از آنجایی که متغیرهای تصادفی X_i دارای توزیع یکسان هستند و مستقل از یکدیگر می‌باشند، متغیرهای تصادفی Y_i نیز دارای همین ویژگی‌ها خواهند بود:

$$\log(Y) = Y_1 + Y_2 + \ \dots + Y_n$$

و Y_iها *i.i.d.* هستند، در نتیجه طبق قضیه حد مرکزی $\log(Y)$ دارای توزیع نرمال است.

قانون بنفورد (Benford’s Law)

- متغیر تصادفی $\log(Y)$ دارای توزیع نرمال است، در نتیجه Y دارای توزیع لگاریتمی نرمال (Log-Normal) خواهد بود:

$$f_Y(y) = \frac{1}{y\,\sigma\sqrt{2\pi}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}$$

- با استفاده از رابطه بالا می‌توان نشان داد که متغیرهای تصادفی که از حاصلضرب مقادیر تصادفی مستقل از هم تشکیل شده باشند، در قانون بنفورد صدق می‌کنند:

قانون بنفورد: رقم آغازین اعداد موجود در بسیاری از دنباله‌ها در طبیعت (به ویژه دنباله‌های ضربی) از توزیع احتمال زیر پیروی می‌کنند:

$$P(X = d) = \log_{10}\left(1 + \frac{1}{d}\right) \ : \ \ d = 1, 2, \dots, 9$$

قانون بنفورد

- بسیاری از دنباله‌های ریاضی از قانون بنفورد پیروی می‌کنند:
  - دنباله توان‌های ۲
  - دنباله فیبوناچی
  - دنباله فاکتوریل

d	P(d)	Relative size of P(d)
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

- کاربردهای قانون بنفورد:
  - اقتصاد کلان
  - تشخیص تقلب در حسابرسی و فرار مالیاتی
  - تشخیص تقلب در انتخابات
  - تشخیص تقلب علمی و عددسازی