

# آمار و احتمال همراه با پایتون (۱)

نجمه مدنی

# فهرست مطالب

- درس اول
  - آشنایی با پایتون و spyder
  - آشنایی با مفاهیم پایه احتمال
  - شبیه سازی
- درس سوم: تولید متغیرهای تصادفی
  - متغیرهای تصادفی پیوسته
  - متغیرهای تصادفی گسسته
  - متغیرهای تصادفی توام
  - محاسبه کواریانس و همبستگی
  - شبیه سازی
- درس چهارم: بررسی قضایا و نامساوی های کاربردی احتمال
  - قضیه دموآور-لاپلاس
  - قضیه حد مرکزی
  - نامساوی مارکوف و چبی شف
  - شبیه سازی
- درس پنجم: آمار
  - تخمین پارامتر
  - آزمون فرضیه
  - شبیه سازی

# درس اول

- مروری بر پایتون و Spyder
- مروری بر مفاهیم پایه احتمال
- شبیه سازی

## مروری بر پایتون

- برنامه نویسی با آن ساده است.
- پکیج های متنوعی دارد و استفاده از آنها رایگان است.
- یک زبان تفسیر شده و سطح بالاست.
- نسبت به C/C++ کندتر است.
- نیازی به گذاشتن سمی کالن در انتهای سینتکس ها نداریم.

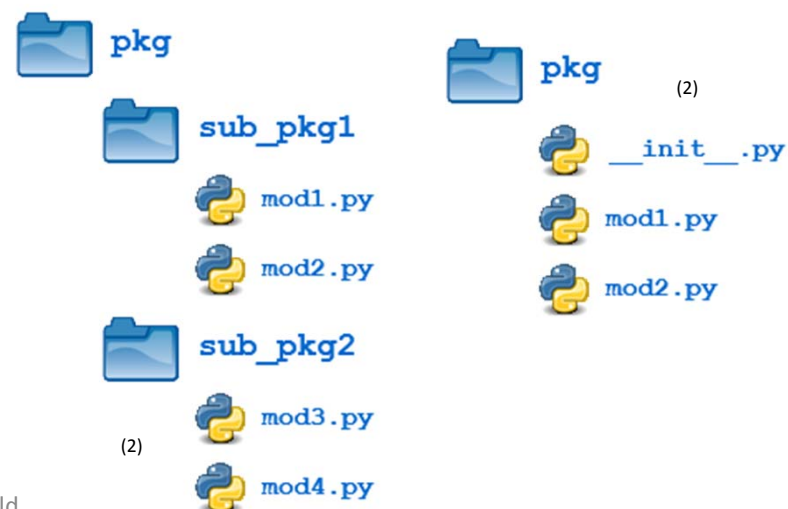
# مروری بر پایتون

• ماژول:

فایلی شامل توابع و تعاریف است.

• برای استفاده از توابع و کلاس ها لازم است ماژول ها / پکیج های حاوی آن ها را وارد برنامه کنیم:

```
• project
  • — package1
    • module1.py
    • module2.py
  • — package2
    • init__.py
    • module3.py
    • module4.py
    • subpackage1
      • module5.py (1)
```



1:<https://realpython.com/absolute-vs-relative-python-imports/>

2:<http://www.pybloggers.com/2018/04/python-modules-and-packages-an-introduction>

# مروری بر پایتون

- وارد کردن ماژول ها به برنامه

```
import module
```

```
    module.func()
```

```
import module as mod
```

```
    mod.func()
```

```
from package import module
```

```
    module.func()
```

# مروری بر پایتون

- همراه با نصب پایتون ماژول های استاندارد `math, random, os` هم نصب می شوند.
- ماژول های مورد نیاز را با دستورهای `conda/ pip` نصب می کنیم.
- پکیج ها/ کتابخانه های مهم پایتون برای محاسبات عددی و علمی:

Scipy ■

Numpy ■

Matplotlib ■

## مروری بر پایتون

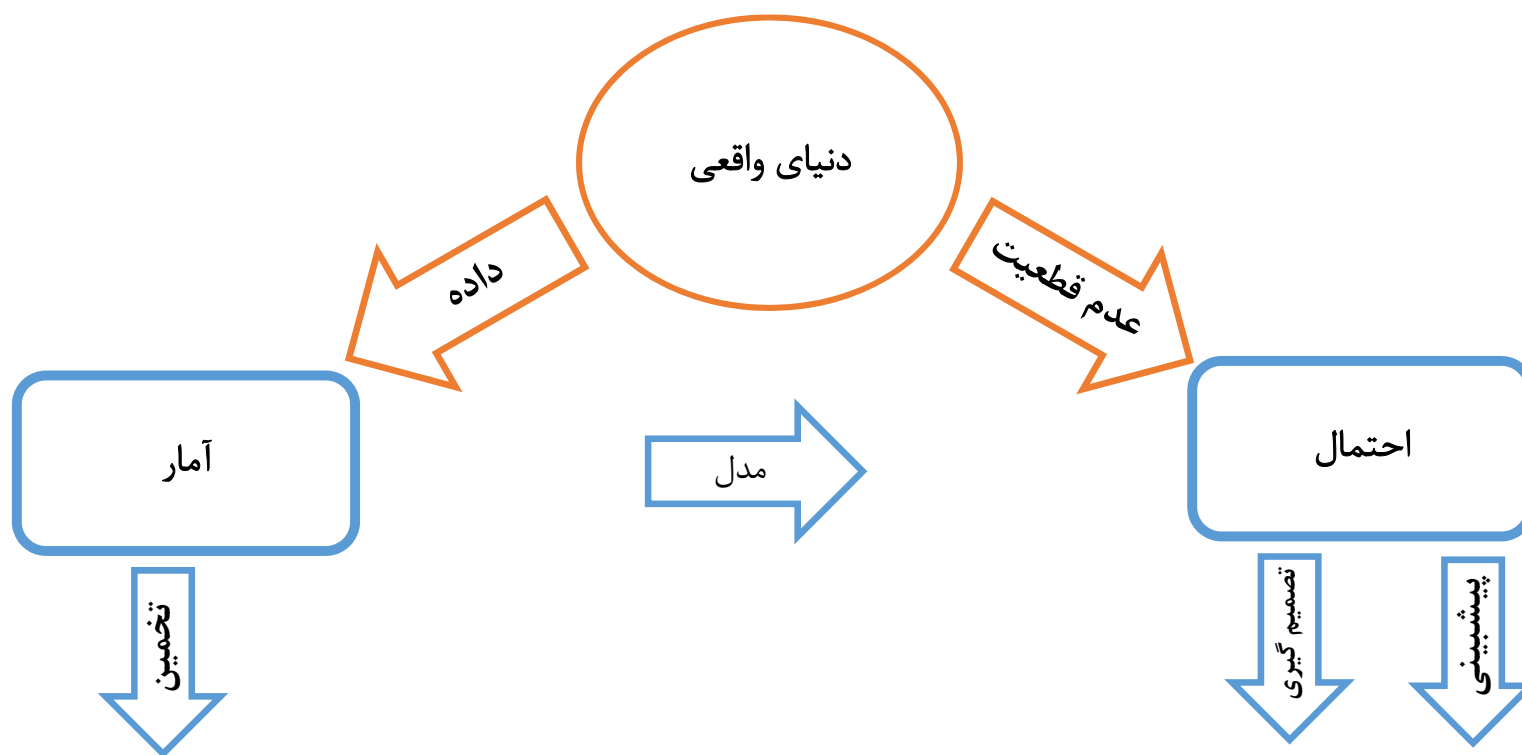
- دو نوع متغیر عددی وجود دارد: `int` , `float`.
- متغیر `char` وجود ندارد.
- رشته ها را با استفاده از “ ” یا ‘ ’ مقدار می دهیم.
- متغیرها : نیازی به تعیین نوع متغیر نیست.
- یک متغیر می تواند در طی برنامه نوع های مختلف داشته باشد.



# نصب پایتون

- Python [www.python.org](http://www.python.org)
- Anaconda [www.anaconda.com](http://www.anaconda.com)
- IDE: مجموعه ای از ابزارهای مناسب برای ویرایش، دیباگ و اجرای برنامه....
- Spyder: مناسب برای برنامه نویسی محاسباتی و علمی.

# آمار و احتمال



# احتمال: مفاهیم پایه

- آزمایش تصادفی (Random experiment)

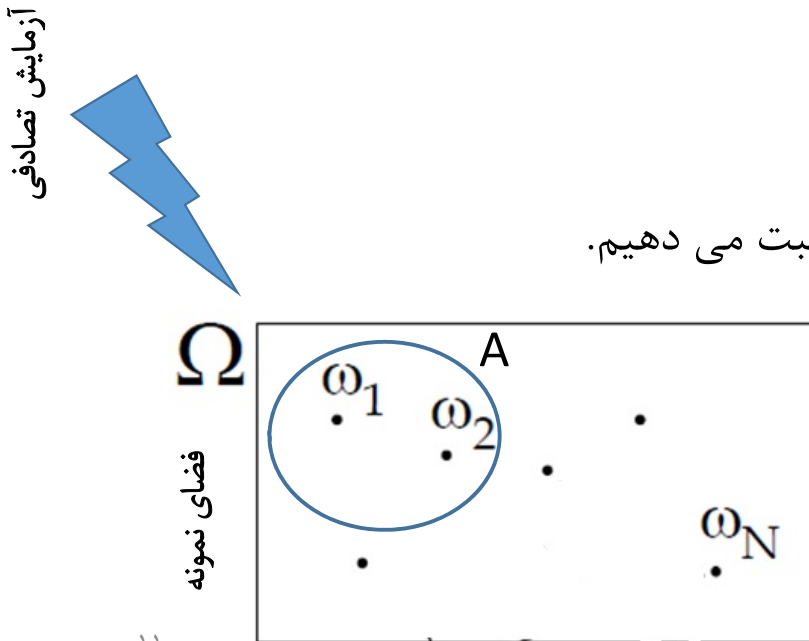
- آزمایش (پدیده) که نتیجه آن را با قطعیت نمی توان مشخص کرد: پرتاب سکه/تاس

- فضای نمونه (Sample space)

- مجموعه تمام نتایج ممکن آزمایش (پدیده) تصادفی .

- واقعه (Event)

- مجموعه ای از نتایج آزمایش تصادفی که به آن یک احتمال نسبت می دهیم.



# تعریف احتمال

- تعریف کلاسیک
- تعریف **تعریف فرکانس نسبی**
- احتمال با استفاده از اصول موضوعه

# تعریف فرکانس نسبی

•  $n$  بار آزمایش تصادفی را انجام می دهیم،  $n(A)$  بار واقعه مورد نظر اتفاق می افتد

$$P(A) \simeq \lim_{n \rightarrow \infty} \frac{n(A)}{n}$$

امکان آزمایش تصادفی به دفعات وجود داشته باشد.

## شبیه سازی

- شبیه سازی پرتاب یک سکه سالم
- محاسبه احتمال شیر آمدن در پرتاب یک سکه سالم

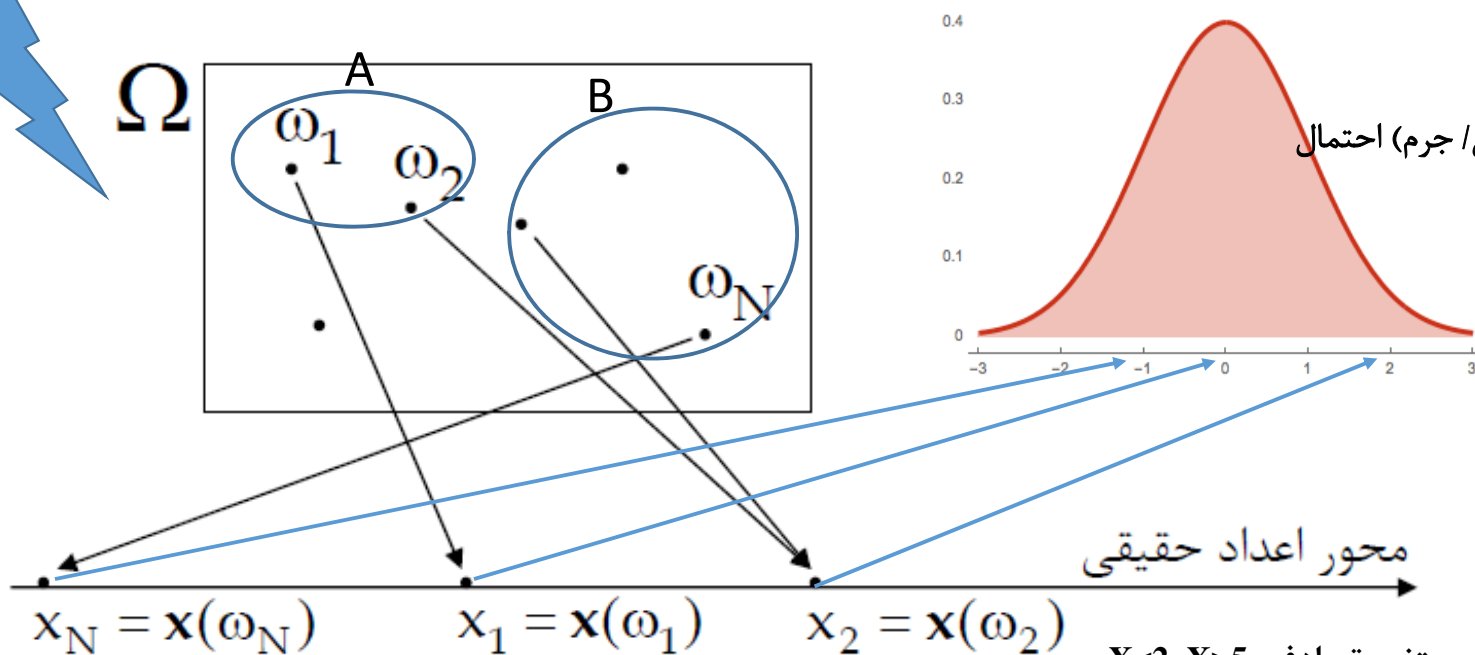
## درس دوم:

- متغیرهای تصادفی
- متغیرهای تصادفی پیوسته
- متغیرهای تصادفی گسسته
- شبیه سازی

# متغیر تصادفی

- متغیر تصادفی: تابعی از فضای نمونه به اعداد حقیقی

آزمایش تصادفی



تعریف وقایع بر حسب متغیر تصادفی  $X < 2, X > 5$



## خواص متغیر تصادفی

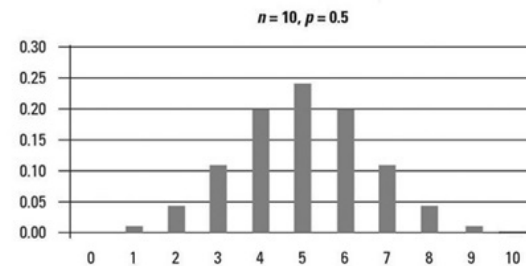
- **سمبل**: با حروف بزرگ نمایش داده می شود  $X$  و مقادیری که به خود می گیرد با حروف کوچک  $x$
- **تکیه گاه**: (support): مجموعه مقادیری که متغیر تصادفی به خود می گیرد
- **تابع توزیع**: pmf یا pdf
- **تابع توزیع تجمعی** (cumulative distributed function)
- **امید ریاضی**: میانگین وزن دار
- **انحراف معیار**: پراکندگی مقادیری که متغیر تصادفی می گیرد نسبت به امید ریاضی
- **مد**: محتمل ترین مقداری که متغیر تصادفی اختیار می کند
- **میانه**، **آنتروپی** .....

## متغیر تصادفی گسسته

- مقادیری که متغیر تصادفی می تواند به خود بگیرد قابل شمارش است.
- مثال: تعداد خط ها در  $n$  بار پرتاب سکه، نتیجه پرتاب یک تاس، تعداد ایمیل های دریافتی در هر روز
- تابع جرم احتمال (probability mass function: pmf) مقادیری که متغیر تصادفی می گیرد و احتمال متناظر آن ها را مشخص می کند.

$$p_X(x_k) = P(X = x_k), k = 1, 2, 3, \dots$$

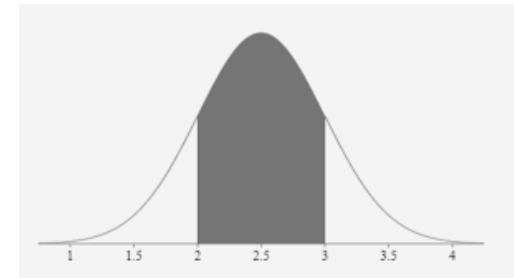
$$\sum p_X(x_k) = 1, \forall x_k \quad p_X(x_k) \geq 0$$



## متغیرهای تصادفی پیوسته

- متغیر تصادفی می تواند هر مقداری در یک بازه از اعداد حقیقی را بگیرد.
- مثال: مدت زمانی که یک مشتری در یک فروشگاه در صف منتظر می ماند.
- در متغیرهای تصادفی پیوسته در مورد احتمال بازه ای از اعداد حقیقی صحبت می کنیم
- با استفاده از تابع چگالی احتمال (PDF) این احتمالات را حساب می کنیم

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(u) du$$



## تابع توزیع تجمعی

- هم برای متغیرهای تصادفی پیوسته و هم برای متغیرهای تصادفی گسسته تعریف می شود

$$F_X(x) = P(X \leq x)$$

- متغیرهای تصادفی پیوسته:

$$F_X(x) = \int_{-\infty}^x f_X(u) du$$

- متغیرهای تصادفی گسسته:

$$F_X(x) = \sum_{x_k \leq x} p_X(x_k)$$

# امید ریاضی-واریانس

گسسته

$$E(X) = \sum x_k p_X(x_k)$$

$$\sigma_X^2 = \sum (x_k - E(X))^2 p_X(x_k)$$

پیوسته

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx$$

# توزیع نرمال

- بسیاری از پدیده ای طبیعی با این متغیر تصادفی مدل می شوند. قد، وزن، خطاهای اندازه گیری. فشار خون،.....
- مجموع تعدادی زیادی متغیر تصادفی یک متغیر تصادفی نرمال
- نویز را غالباً با توزیع نرمال مدل می کنیم.

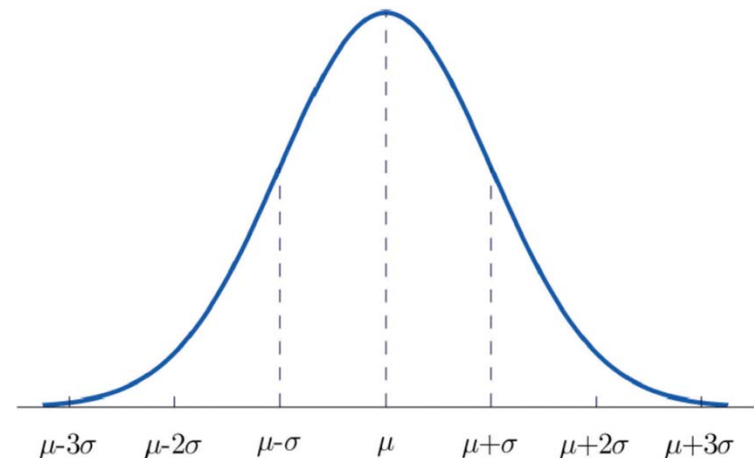
$$X \sim N(\mu, \sigma^2)$$

$$\text{PDF: } f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{Support: } (-\infty, +\infty)$$

$$\text{Expectation: } E[X] = \mu$$

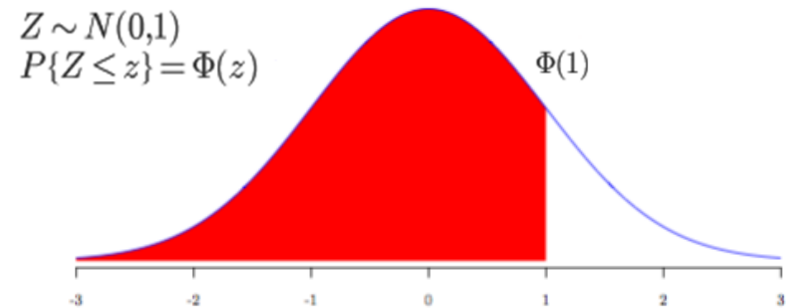
$$\text{Variance: } \text{Var}(X) = \sigma^2$$



# CDF of Normal Distribution

• تابع توزیع تجمعی توزیع نرمال استاندارد

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990



$$F_X(x) = P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

## توزیع نرمال: مثال

$$\begin{aligned}P(|X - \mu| \leq 3\sigma) &= P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \\&= \Phi\left(\frac{\mu + 3\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 3\sigma - \mu}{\sigma}\right) \\&= \Phi(3) - \Phi(-3) \\&= 2\Phi(3) - 1 \approx 1\end{aligned}$$



# توزیع نمایی

- مثال: مدت زمان تا اتفاق بعدی.

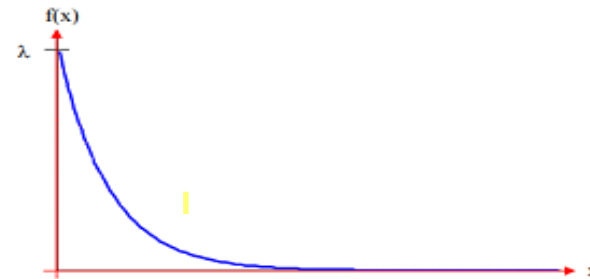
$$X \sim \text{Exp}(\lambda)$$

$$\text{PDF: } f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Support: } [0, \infty)$$

$$E[X] = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$



## توزیع نمایی: مثال

- بازدیدکنندگان سایت شما بعد به طور متوسط بعد از ۵ دقیقه سایت را ترک می کند. احتمال اینکه یک بازدید کننده بعد از ۱۰ دقیقه سایت را ترک کند چقدر است.
- $X$ : مدت زمانی که طول می کشد که یک بازدید کننده سایت را ترک کند:  $X \sim \text{Exp}(\lambda), \lambda = ?$
- بازدیدکنندگان به طور متوسط بعد از ۵ دقیقه سایت را ترک می کنند.  $E(X) = 5 \rightarrow \lambda = 0.2$
- تابع توزیع تجمعی:  
$$F_X(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda u} du = (1 - e^{-\lambda x})$$
- $$P(X \geq 10) = 1 - F_X(10) = e^{-0.2 \times 10} = 0.135$$

## خاصیت منحصر به فرد توزیع نمایی

$$P(X > t + s | X > s) = P(X > t)$$

• بی حافظه بودن

- مثال ۵ دقیقه از ورود آخرین مشتری گذشته احتمال اینکه یک دقیقه دیگر مشتری وارد شود:
- $X$ : مدت زمانی که طول می کشد تا مشتری جدید وارد شود:

$$P(X > 6 | X > 5) = P(X > 1)$$

# توزیع برنولی

- دو مقدار می گیرد: مقدار ۱ را با احتمال  $P$  و مقدار ۰ را با احتمال  $1-P$
- پرتاب یک سکه، تولید یک بیت اطلاعات

$$X \sim \text{Bern}(p)$$

$$P(X = 1) = p$$

$$P(X = 0) = q = 1 - p$$

$$E(X) = p$$

$$\text{Var}(X) = pq$$

# توزیع دوجمله ای

- تعداد دفعات موفقیت در  $n$  بار انجام یک آزمایش.
- مثال: تعداد بارهایی که در پرتاب  $n$  سکه خط حاصل می شود . تعداد بارهایی که یک تبلیغ توسط بازدیدکنندگان یک سایت کلیک می شود

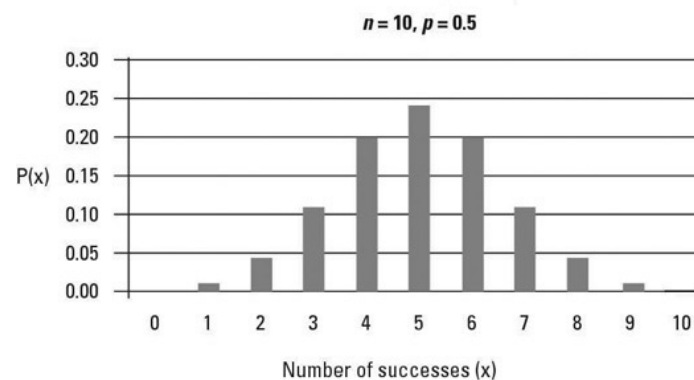
$$X \sim \text{Bin}(n, p)$$

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

$$k = 0, 1, \dots, n$$

$$E(X) = np$$

$$\text{Var}(X) = npq$$



# توزیع پواسن

- تعداد وقایعی که در یک بازه اتفاق می افتد مشروط به دانستن تعداد متوسط وقوع در این بازه .
- مثال : تعداد درخواست هایی برای تاکسی اینترنتی در یک منطقه در ۱۰ ثانیه آینده، تعداد زمین لرزه ها در یک سال

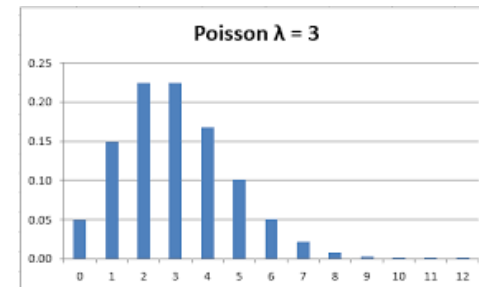
$$X \sim \text{Poi}(\lambda)$$

$$P(X = k) = \lambda^k \frac{e^{-\lambda}}{k!}$$

$$k = 0, 1, \dots, \infty$$

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda$$



## توزیع پواسن تقریبی از توزیع دو جمله ای

- توزیع پواسن می تواند توزیع نرمال را تقریب بزند. تحت چه شرایطی؟
- $n$  بزرگ باشد،  $p$  کوچک باشد و  $np$  مقداری متوسط

$$\binom{n}{k} p^k q^{n-k} \approx \lambda^k \frac{e^{-\lambda}}{\lambda!}$$

$$E(X) = np = \lambda$$
$$Var(X) = np(1-p) = \lambda$$

- مثال:  $n=1000$  لامپ داریم احتمال خراب بودن هر لامپ  $p=0.01$  است احتمال اینکه فقط یک لامپ خراب باشد چقدر است.

$$X \sim Bin(1000, .01)$$

$$P(X = 1) = \binom{1000}{1} 0.01 \times 0.99^{999}$$

$$= 0.00043630732$$

$$X \sim Poi(10)$$

$$P(X = 1) = 10 \frac{e^{-10}}{10!}$$

$$= 0.0004539992$$

# تولید اعداد تصادفی

- مولد اعداد شبه تصادفی (pseudo random number generator: PRNG) :
- " الگوریتمی برای تولید دنباله ای از اعداد است که خواص یک دنباله اعداد تصادفی را تقریب می زند."
- دنباله اعداد تولید شده توسط PRNG تصادفی به معنای واقعی نیستند به نحوی که با داشتن مقدار اولیه (seed) قابل تعیین هستند.
- کاربرد: شبیه سازی مونت کارلو، بازی های کامپیوتری و رمز نگاری

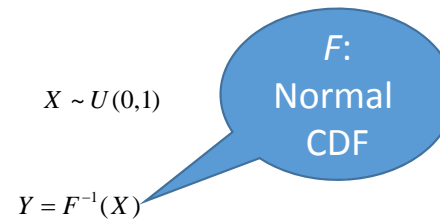


# تولید اعداد تصادفی نرمال

- زبان های برنامه نویس رایج مولد های PRN دارند.
- با اغماض می توانیم بگوییم اعدادی که تولید می شوند تقریبا مستقل و دارای توزیع یکنواخت هستند
- اگر دنباله تولیدی را اینگونه نمایش دهیم.  $U_1, U_2, U_3, \dots$   
 $U_i \sim \text{Uniform}[0,1]$
- برای تولید نمونه های تصادفی از توزیع های دیگر بر اساس این دنباله روش های مختلفی وجود دارد

# تولید نمونه های تصادفی نرمال

- متغیر تصادفی  $Y$  دارای توزیع نرمال خواهد بود :



$$P(Y \leq y) = P(F^{-1}(X) \leq y) = P(X \leq F(y)) = F(y)$$

## شبیه سازی

- تولید متغیر تصادفی در پایتون
  - رسم تابع چگالی احتمال، تابع توزیع احتمال ، رسم هیستوگرام
  - بررسی رابطه توزیع پواسن و توزیع دوجمله ای
  - حل یک مسئله احتمال با استفاده از شبیه سازی
- طول عمر المان A دارای توزیع نمایی  $X \sim \text{Exp}(\lambda_x)$  و طول عمر المان B دارای توزیع نمایی با  $Y \sim \text{Exp}(\lambda_y)$  احتمال اینکه المان B بیش از المان A عمر کند چقدر است.

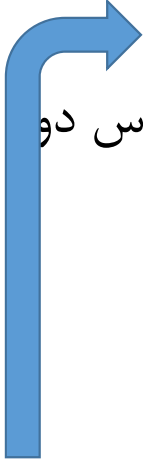
$$P(Y > X) = \frac{\lambda_x}{\lambda_x + \lambda_y}$$

## درس سوم

- متغیرهای تصادفی توام
- استقلال متغیرهای تصادفی
- کواریانس و همبستگی
- شبیه سازی

# توزیع توام

- روی یک فضای نمونه می توان بیش از یک متغیر تصادفی تعریف کرد
- مثال: پرتاپ دو تاس  $X$ : مقدار تاس اول  $Y$ : مقدار تاس دوم
- مثال: پرتاپ دو تاس  $X$ : مقدار تاس اول  $Y$ : جمع مقادیر تاس اول و تاس دوم



$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

MIT18\_05S14\_class7slides

## توزیع توام

- در اکثر شرایط نیاز به دو یا چند متغیر تصادفی برای توصیف مسئله داریم
- مثال: شاخص های فیزیولوژیکی مختلف از جامعه بیماران
- مثال: سن افراد عضو فیسبوک و تعداد دوستان آن ها
- مثال: نرخ بیماری های تنفسی و سطح آلودگی هوا در یک شهر
- تابع توزیع مشترک: برای محاسبه واقعه هایی که دو یا چند متغیر تصادفی در آن نقش دارند و بررسی رابطی بین متغیرهای تصافی
- ساده ترین حالت: متغیر های تصادفی مستقل. در غیر این صورت از کواریانس یا ضریب همبستگی به عنوان یکی از پارامترهای که برای تعیین توزیع مشترک لازم است. استفاده می شود

# توزیع توام

## پیوسته

- تابع چگالی احتمال مشترک
- تابع توزیع مشترک

$$F_{XY}(x, y) = P\{X \leq x, Y \leq y\}$$

$$= \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) du dv$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

## گسسته

- تابع چگالی جرم مشترک
- تابع توزیع مشترک

$$X: \{x_1, x_2, \dots, x_n\} \quad Y: \{y_1, y_2, \dots, y_m\}$$

$$p_{X,Y}(x_i, y_j) = P(\{X=x_i\} \cap \{Y=y_j\})$$

$$F_{XY}(x, y) = \sum_{x \leq x_i} \sum_{y \leq y_j} P(x_i, y_j)$$

$$P(X = x_i) = \sum_{j=1}^m P(x_i, y_j)$$

# استقلال دو متغیر تصادفی

- دو واقعه مستقل:

$$P(A \cap B) = P(A)P(B)$$

- دو متغیر تصادفی مستقل:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

- مثال:

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$



## کواریانس - همبستگی

- کواریانس: میزانی برای اندازه گیری تغییرات دو متغیر تصادفی نسبت به هم. تغییر یک متغیر تصادفی چگونه بر متغیر تصادفی دیگر تاثیر می گذارد.

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

- همبستگی: نرمالیزه کواریانس

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

## ماتریس کواریانس

$$C = \begin{bmatrix} \underbrace{E(X-E(X))^2}_{\sigma_x^2} & \underbrace{E(X-E(X))(Y-E(Y))}_{\rho\sigma_x\sigma_y} \\ \underbrace{E(X-E(X))(Y-E(Y))}_{\rho\sigma_x\sigma_y} & \underbrace{E(Y-E(Y))^2}_{\sigma_y^2} \end{bmatrix}$$

## تابع توزیع گوسی توام دو بعدی

$$f_{XY}(x, y) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{\begin{bmatrix} X-E(X) \\ Y-E(Y) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} X-E(X) \\ Y-E(Y) \end{bmatrix}}{2}}$$

$$\begin{bmatrix} X-E(X) \\ Y-E(Y) \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} X-E(X) \\ Y-E(Y) \end{bmatrix} = K$$

سطوح تراز بیضی هستند

## نکاتی از جبر خطی

- ماتریس کواریانس متقارن و positive definite است.
- در ماتریس متقارن، بردارهای ویژه بر هم عمود هستند.
- جهت بردارهای ویژه در ماتریس کواریانس جهت بیشترین میزان پراکندگی داده را نشان می دهد و مقدار ویژه متناظر میزان این پراکندگی را نشان می دهد.
- سطوح تراز تابع چگالی احتمال مشترک گوسی بیضی هستند و محورهای اصلی آن منطبق بر بردارهای ویژه هستند.

# تولید نمونه تصادفی از توزیع گوسی مشترک

- برای نمونه برداری از یک توزیع گوسی مشترک با میانگین و واریانس دلخواه می توان
- از توزیع گوسی استاندارد نمونه برداری کرد و با داشتن ماتریس کواریانس و میانگین آن ها را به نمونه هایی از توزیع مورد نظر تبدیل کرد

$$X \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$Y = LX + \vec{\mu} \quad Y \sim N(\vec{\mu}, \Sigma_y)$$

$$\Sigma_y = E((Y - E(Y))(Y - E(Y))^T)$$

$$= L \Sigma_x L^T = LL^T$$



Cholesky  
decomposition

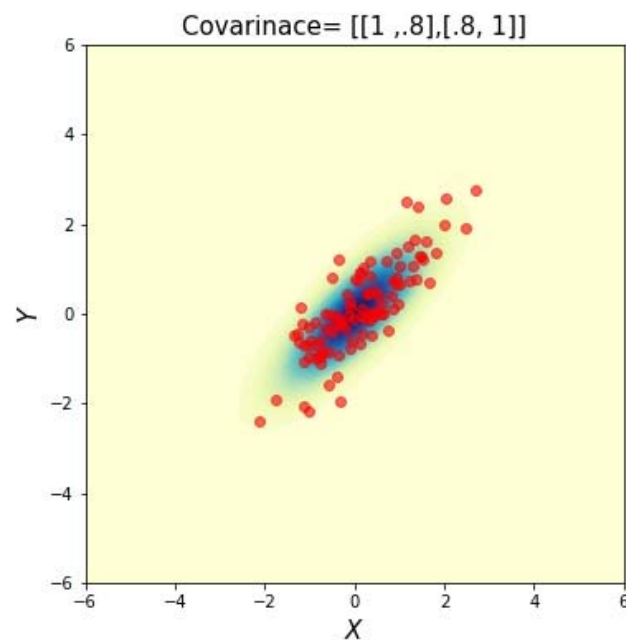
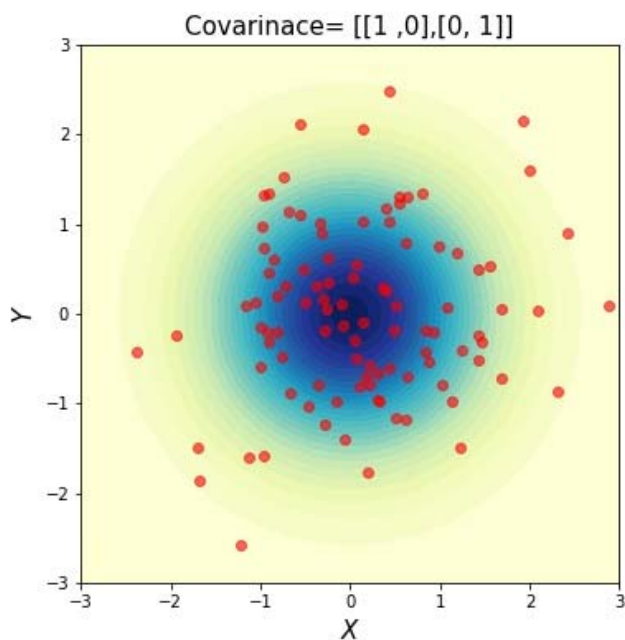
Eigen  
decomposition

$$\Sigma_y = \underbrace{Q}_{\text{eigen-vectors}} \Lambda Q^T = Q \sqrt{\Lambda} (Q \sqrt{\Lambda})^T$$

$$T = \underbrace{Q}_{\text{rotation}} \underbrace{\sqrt{\Lambda}}_{\text{scaling}}$$

$$Y = TX + \vec{\mu}$$

# تولید نمونه تصادفی از توزیع گوسی مشترک



## شبیه سازی

- تابع چگالی احتمال مشترک گوسی
- بررسی کواریانس

## درس چهارم

- قضیه لاپلاس-دموآور
- قضیه حد مرکزی
- نامساوی مارکوف
- نامساوی چپی شف
- شبیه سازی



## قضیه دموآور-لاپلاس

- توزیع دوجمله ای تحت شرایطی با توزیع نرمال قابل تقریب است.

$$\text{Bin}(n, p) \rightarrow \binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{npq}} e^{-\frac{(k-np)^2}{2npq}} \leftarrow N(np, npq)$$
$$k = np + c\sqrt{npq}$$
$$n \rightarrow \infty$$

## قضیه دموآور-لاپلاس: مثال

- مثال: جمعیتی خاص که نیمی از آن ها مرد و نیمی از آن ها زن هستند در نظر بگیرد. ۱۰۰۰۰ نفر به تصادف از این جمعیت انتخاب شده اند. احتمال اینکه ۴۹٪ تا ۵۱٪ جمعیت افراد شده مرد باشد چقدر است؟

$$\text{Bin}(10000, \frac{1}{2})$$

$$P(4900 \leq X \leq 5100) = \sum_{k=4900}^{k=5100} \binom{10000}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{10000-k} = ?$$

$$E(X) = np = 5000$$

$$\text{Var}(X) = npq = 2500 \quad \sqrt{npq} = 50$$

$$\begin{aligned} P(4900 \leq X \leq 5100) &\approx \Phi\left(\frac{5100 - 5000}{50}\right) - \Phi\left(\frac{4900 - 5000}{50}\right) = \Phi(2) - \Phi(-2) \\ &= 2\Phi(2) - 1 \\ &= 2 \times 0.9772 - 1 \\ &= .9544 \end{aligned}$$

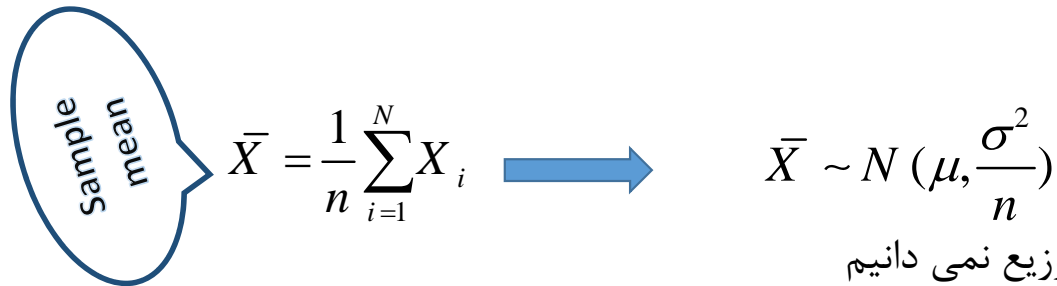
- X: تعداد مرد ها در میان ۱۰۰۰۰ نفر
- احتمال مرد بودن = احتمال زن بودن



# بررسی قضایای احتمال

- قضیه حد مرکزی:

$X_1, X_2, \dots, X_n$  متغیرهای تصادفی مستقل از هم و دارای توزیع یکسان (iid) و  $E(X_i) = \mu_i, \text{Var}(X_i) = \sigma^2$  در این صورت مجموع متغیرهای تصادفی زمانیکه  $n$  به اندازه کافی بزرگ باشد تقریبی از یک توزیع نرمال هست.


$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \longrightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

در عمل اکثر اوقات میانگین و واریانس توزیع نمی دانیم

میانگین و واریانس را با استفاده از مجموعه داده تخمین می زنیم و از

قضیه حد مرکزی استفاده می کنیم.

$$n = ?$$

## قضیه حد مرکزی

- مثال: ده تاس را پرتاب می کنیم continuity correction
- مجموع اعدادی که تاس ها نشان می دهد  $Y = X_1 + X_2 + \dots + X_{10}$
- $P(Y \leq 45)$  ؟

- با استفاده از قضیه حد مرکزی

$$E(X_i) = \frac{\sum_{i=1}^6 1 + 2 + \dots + 6}{6} = 3.5$$

$$\text{Var}(X_i) = E(X_i^2) - E(X_i)^2 = \frac{35}{12}$$

$$Y \sim N\left(10 \times 3.5, 10 \times \frac{35}{12}\right) \quad P(Y \leq 45) = \Phi\left(\frac{10}{\sqrt{10 \times \frac{35}{12}}}\right) = \Phi(1.8516) = 0.9678$$

- با استفاده از شبیه سازی: ۰/۹۷۵۲۶

## نامساوی مارکوف-چی شف

• نامساوی مارکوف:  $X$  متغیر تصادفی  
support :  $x \geq 0$

$$P(X > a) \leq \frac{E(X)}{a}$$



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

• نامساوی چیشف:  $X$  متغیر تصادفی

## نامساوی مارکوف: مثال

- $X$ : تعداد خط ها در  $n$  پرتاپ یک سکه سالم:  $P(X \geq n) = ?$   
 $X \sim \text{Bin}(n, \frac{1}{2}) \rightarrow E(X) = \frac{n}{2}$

- با استفاده از نامساوی مارکوف  
$$P(X \geq n) \leq \frac{1}{2}$$

- با استفاده از توزیع  
$$P(X \geq n) = P(X = n) = \left(\frac{1}{2}\right)^n$$

## نامساوی چیشف : مثال

- $X$  : تعداد خط ها در  $n$  پرتاپ یک سکه سالم:  $P(X \geq n) = ?$   
 $X \sim \text{Bin}(n, \frac{1}{2}) \rightarrow E(X) = \frac{n}{2}$

- با استفاده از نامساوی مارکوف  
$$P(X \geq n) \leq \frac{1}{2}$$

- با استفاده از توزیع  
$$P(X \geq n) = P(X = n) = \left(\frac{1}{2}\right)^n$$

## نامساوی چیشف: مثال

•  $X$ : تعداد خط ها در  $n$  پرتاپ یک سکه سالم:  $P(X \geq n) = ?$

$$X \sim \text{Bin}(n, \frac{1}{2}) \rightarrow \begin{aligned} E(X) &= \frac{n}{2} \\ \text{Var}(X) &= \frac{n}{4} \end{aligned}$$

$$P(X \geq \frac{3n}{4}) \leq \frac{2}{3}$$

• با استفاده از نامساوی مارکوف

$$\frac{\text{Var}(X)}{\left(\frac{n}{4}\right)^2}$$

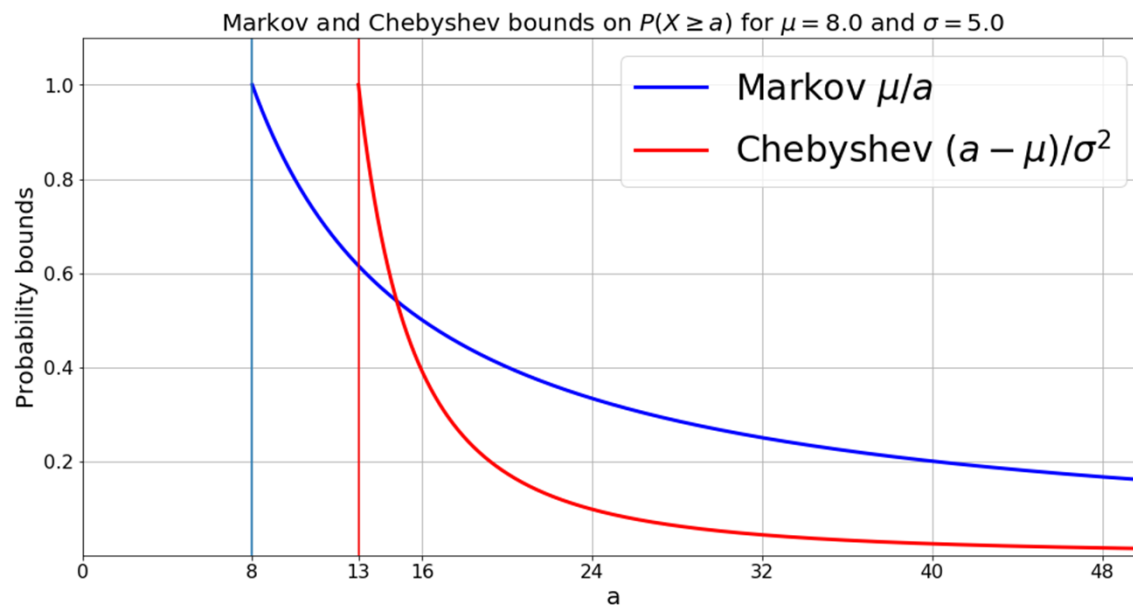
• با استفاده از نامساوی چپ شیف

$$P(X \geq \frac{3n}{4}) = P(X - \frac{n}{2} \geq \frac{n}{4}) \leq P\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) = P(|X - E(X)| \geq \frac{n}{4}) \leq \frac{4}{n}$$

$$P(X - \frac{n}{2} \geq \frac{n}{4}) + P(X - \frac{n}{2} \leq -\frac{n}{4})$$



# نامساوی چیشف-مارکوف



<https://github.com/kartikdube/Probability-and-Statistics-in-Data-Science-Using-Python/blob/master/09-Inequalities%20and%20Limit/Inequalities%20and%20Limit.ipynb>

## درس هفتم

- تخمین میانگین / تخمین واریانس
- شبیه سازی

## نمونه (sample)

- متغیرهای تصادفی  $X_1, X_2, \dots, X_n$  یک نمونه از توزیع (جامعه)  $F$  است اگر
- $X_i$  ها داری توزیع یکسان و از هم مستقل باشند (iid).
- $X_i$  دارای میانگین و واریانس یکسان باشند.  $E(X_i) = \mu$   $Var(X_i) = \sigma^2$
- $(X_1, X_2, X_3, X_4, X_5, X_6)$   $(0.7, 0.5, 0.3, 0.4, 0.33, -0.3)$
- می خواهیم میانگین و واریانس توزیع (جامعه)  $F$  را تقریب بزنیم.
- با توجه به اینکه تعداد اعضای یک جامعه زیاد است. در نظر گرفتن همه اعضا در محاسبه میانگین و واریانس کار بهینه ای نیست.
- از جامعه، نمونه انتخاب می کنیم و با استفاده از آن میانگین و واریانس را محاسبه می کنیم.

## میانگین نمونه

$$\bar{X} = \frac{1}{n} \sum_{i=0}^n X_i$$

- $\bar{X}$  یک تخمین برای  $\mu$  است.
- $\bar{X}$  یک متغیر تصادفی است و  $\mu$  یک مقدار مشخص است.
- $\bar{X}$  یک تخمین ناریب از  $\mu$  است:  $E(\bar{X}) = \mu$
- قضیه حد مرکزی: برای  $n$  های به اندازه کافی بزرگ  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

## خطای تخمین

- پارامتر  $\theta$  را با متغیر تصادفی  $\hat{\theta}$  تخمین می زنیم. خطای تخمین:  $E(\hat{\theta} - \theta)^2 = ?$

$$E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta} - \theta) + (E(\hat{\theta} - \theta))^2$$

$$= \underbrace{\text{var}(\hat{\theta})}_{\theta \text{ is not random}} + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{bias}}$$

$$E(\bar{X} - \mu)^2 = \frac{\sigma^2}{n} + \underbrace{0}_{\text{bias}}$$

- خطای (تخمین گر) میانگین نمونه:
- برای  $n$  های بزرگ به سمت صفر میل می کند
- خطای استاندارد  $\sqrt{\frac{\sigma^2}{n}}$

## فاصله اطمینان برای میانگین نمونه

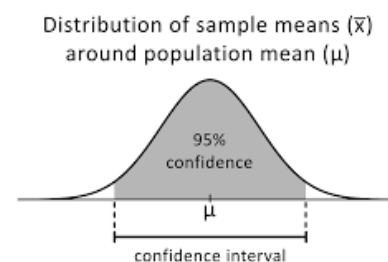
$$\left( \bar{X} - z^* \frac{\sigma}{\sqrt{n}} \quad \bar{X} + z^* \frac{\sigma}{\sqrt{n}} \right)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$P\left(\left| \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right| < z^*\right) = CI$$

$$P\left(\mu - z^* \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z^* \frac{\sigma}{\sqrt{n}}\right) = 2\Phi(z^*) - 1$$

• فاصله اطمینان



C	$z^*$
99%	2.576
98%	2.326
95%	1.96
90%	1.645

<https://www.omnicalculator.com/statistics/confidence-interval>

## واریانس نمونه

- در صورتی که مقدار میانگین را داشته باشیم

$$\sigma^2 = E((X_i - E(X_i))^2) \Rightarrow$$

$$\overline{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

- در صورتی که مقدار میانگین را نداشته باشیم ابتدا آن را با میانگین نمونه تخمین می زنیم و در این صورت واریانس نمونه (تخمین واریانس) به صورت زیر خواهد بود. ضریب (n-1) ناریب بودن تخمین را تضمین می کند.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$