2.4 方差与标准差

如果有一个数据集，比如说：

$$X=\{x_1,x_2,\cdots,x_n\}$$

那么可以如下来衡量这些数据的"集中"程度：

$$S=\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n}$$

这个思路也可以用来衡量随机变量X的"集中"程度，不过对于随机变量，我们不使用算术平均，而是用加权平均，或者说期望。

2.4.1 方差的定义

代数式：

$$\operatorname{Var}(X)=E\left[(X-E(X))^2\right]$$

或（在浙江大学的教材中用的是下列符号）：

$$D(X)=E\left[(X-E(X))^2\right]$$

称为随机变量X的方差（Variance），也可记作σ^2或者σ_X^2。

2.4.2 二阶矩

方差可以写作：

$$\sigma^2=E\left[(X-\mu)^2\right]=\sum_i(x_i-\mu)^2p(x_i)$$

其中$(x_i-\mu)^2$是二次项，对比一阶矩$E(X)$而言，方差σ^2也可以称作二阶矩（这在物理中对应惯性矩，因为不是高中知识了，这里就不再介绍了）。

2.4.3 方差的各种计算公式

$$E\left[(X-\mu)^2\right]=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n}$$

$$E\left[(X-\mu)^2\right]=\sum_i(x_i-\mu)^2p(x_i)$$

$$\begin{aligned}E\left[(X-\mu)^2\right]&=\mathrm{E}\left[X^2-2X\mathrm{E}[X]+(\mathrm{E}[X])^2\right]\\&=\mathrm{E}\left[X^2\right]-2\mathrm{E}[X]\mathrm{E}[X]+(\mathrm{E}[X])^2\\&=\mathrm{E}\left[X^2\right]-(\mathrm{E}[X])^2\end{aligned}$$

2.4.4 标准差

方差有一个问题，单位与随机变量不同，由于方差的运算中存在平方，因此方差的单位为随机变量单位的平方。为了与随机变量比较，我们需要将方差的单位换算成随机变量的单位，这里我们引入标准差的概念：

假如随机变量X的方差为$Var(X)$，则称：

$$\sigma(X)=\sqrt{Var(X)}$$

为标准差，也可以记作σ或者σ_X。

2.4.5 方差的性质

- **化简**：可以通过下式来化简运算：

$$Var(X)=E\left(X^2\right)-\mu^2$$

- **常数**：若c为常数，则：

$$Var(c)=0$$

- **运算**：若a、b为常数，则：

$$Var(aX+b)=a^2Var(X)$$

2.4.6 二项分布的方差

	伯努利分布	二项分布
PMF	$p(x)=\begin{cases}p, & x=1\\ 1-p, & x=0\end{cases}$	$p(x)=\binom{n}{x}p^x(1-p)^{n-x}$
μ	p	np
$Var(X)$	$p(1-p)$	$np(1-p)$

（1）伯努利分布的方差可以如下计算：

$$E(X^2)=1^2\times p+0^2\times(1-p)=p$$

然后：

$$Var(X)=E\left(X^2\right)-E(X)^2=p-p^2=p(1-p)$$

（2）二项分布的方差可以如下计算：

$$
\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} (k+1-1) k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} (k-1) k \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} (k-1) k \binom{n}{k} p^k (1-p)^{n-k} + np \\
&= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} (1-p)^{(n-2)-(k-2)} + np \\
&= n(n-1)p^2 + np
\end{aligned}
$$

然后:

$$
Var(X) = E\left(X^2\right) - E(X)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)
$$

> 例题
>
> 某人进行射击练习，已知他每一次命中靶心的概率为p,则是否仅当$p=0$时，射击n次后，命中靶心次数的方差最小
>
> 设$X=$“命中靶心次数”
>
> 射击每次独立，随机变量X满足二项分布
>
> $$X \sim b(n,p)$$
>
> 根据二项分布的方差，方差为:
>
> $$Var(X) = np(1-p)$$
>
> 由于$0 \le p \le 1$,$n > 0$,因此有:
>
> $$np(1-p) \ge 0$$
>
> 当$p=0$或$p=1$时，取得最小值。本题说法不正确。

2.4.7 马尔可夫不等式

2.4.7.1 平均身高的例子

“中国男人平均身高1.718米”，也就是说身高X的$E(X)=1.718$，既然是平均身高，那么自然会有高有矮，会不会存在身高171.8米的巨人?

这个问题用数学的语言来讲的话，就是要求：

$$P(X=171.8)=?$$

但这里只知道平均身高，也就是数学期望$E(X)$，怎么求出具体的概率值呢？回想下数学期望的定义：

$$E(X)=\sum_{i=0}^{\infty}x_iP(X=x_i)$$

由于身高一定大于0的，所以上面累加的每一项都是正数，从而可得：

$$171.8P(X=171.8)\leq E(X)$$

进而可以推出：

$$P(X=171.8)\leq\frac{E(X)}{171.8}=1\%$$

这说明171.8米的巨人存在的概率很低。

单独的171.8米的巨人存在的可能性很低，但有没有可能存在一群171.8米及以上的巨人呢？

这个问题也就是问：

$$P(X\geq 171.8)=?$$

可以这么来考虑，根据定义（下面这个式子并不严格，只是示意）：

$$P(X\geq 171.8)=P(X=171.8)+P(X=172.8)+\cdots$$

在身高为正的前提下，下面不等式是成立的：

$$171.8P(X=171.8)+172.8P(X=172.8)+\cdots\leq E(X)$$

缩放一下：

$$171.8P(X=171.8)+171.8P(X=172.8)+\cdots\leq E(X)$$

进而推出：

$$171.8(P(X=171.8)+P(X=172.8)+\cdots)=171.8P(X\geq 171.8)\leq E(X)$$

最后得到：

$$P(X\geq 171.8)\leq\frac{E(X)}{171.8}=1\%$$

2.4.7.2 马尔可夫不等式定义

设X为取非负值的随机变量，则对于任何$a>0$，有：

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

2.4.7.3 马尔可夫不等式的应用

- 不超过1/5的人口会有超过5倍于人均收入的收入。

- 证明一个非负的随机变数，其平均值μ和中位数m满足$m \leq 2\mu$的关系

> 例题
>
> 根据官方数据，中国人均收入为51350元（假设收入皆为正数），那么年入超过百万的人会超过10%吗?
>
> 根据马尔可夫不等式:
>
> $$P(X \geq 1000000) \leq \frac{51350}{1000000} \approx 5.14\%$$
>
> 所以不会超过10%。
>
> 当然据说官方统计出来的数据为，年入超过百万的人真正占比为万分之四，所以马尔可夫不等式给出的估计偏差还是比较大的。

2.4.8 切比雪夫不等式

2.4.8.1 切比雪夫不等式的定义

设X是一随机变量，均值μ和方差σ^2有限，则对任何$k > 0$有:

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

2.4.8.2 切比雪夫不等式的证明

由于$(X-\mu)^2$为非负的随机变量，利用马尔可夫不等式（其中$a = k^2$），得:

$$P\left[(X-\mu)^2 \geq k^2\right] \leq \frac{E\left[(X-\mu)^2\right]}{k^2}$$

由于$(X-\mu)^2 \geq k^2$与$|X-\mu| \geq k$是等价的，因此上述方程可以等价于:

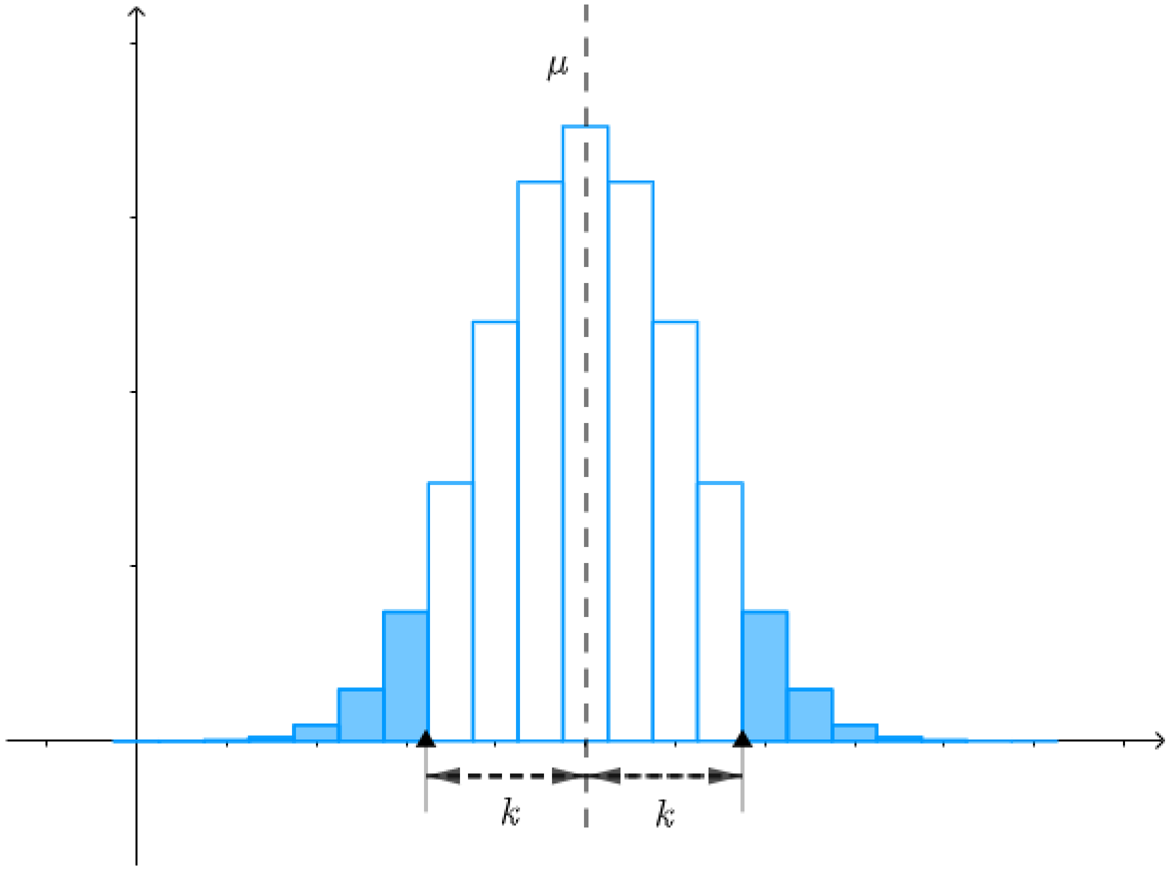
$$P(|X-\mu| \geq k) \leq \frac{E\left[(X-\mu)^2\right]}{k^2} = \frac{\sigma^2}{k^2}$$

> 例题
>
> 根据官方数据有:
>
> - 中国人均收入为51350元
> - 收入的标准差为44000元
>
> 那么有多少人收入超过百万?
>
> 根据题干，我们实际上知道了$\mu = 51350$，$\sigma = 44000$，那么人均收入超过百万的包含在下列不等式中:
>
> $$|X - 51350| \geq 948650$$
>
> 根据切比雪夫不等式有:

> $$P(|X-51350| \geq 948650) \leq \frac{44000^2}{948650^2} \approx 0.22\%$$
>
> 也就是千分之二左右，比之前的马尔可夫不等式估计出来的值要小多了。

2.4.8.3 切比雪夫不等式的意义

$P(|X-\mu| \geq k)$指的是与μ距离大于k的概率，也就是下图中蓝色阴影部分的面积：



当k越大，那么$\frac{\sigma^2}{k^2}$会越小，结合切比雪夫不等式：

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

也就是说X越远离μ，则概率越小；或者通俗地说，X大概率会围绕在μ附近。

2.5 泊松分布

$$X \sim b(5000, 0.001)$$

要求的概率为：

$$P(X \leq 5) = \sum_{k=0}^{5} \binom{5000}{k} 0.001^k 0.999^{5000-k}$$

这个概率计算量很大，由于n很大，p很小，所以可以用$\lambda = np = 5$的泊松分布近似：

$P(X \le 5) \approx \sum_{k=0}^{5} \frac{5^k}{k!} e^{-5} = 0.616$2.5 泊松分布

2.5.1 定义 期望与方差

对于随机变量X的概率质量函数：

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \cdots$$

称为随机变量X的泊松分布，也可以记为：

$$X \sim P(\lambda)$$

其数学期望和方差为：

$$E(X) = \lambda, \quad Var(X) = \lambda$$

期望可以这么计算：

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \end{aligned}$$

其中$\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$为$e^{\lambda}$的泰勒级数，即：

$$e^{\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

所以上式为：

$$E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

通过之前的推导，我们知道λ其实是对应的二项分布的μ，所以泊松分布的期望和对应的二项分布的期望相同。

又有：

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda}$$

$$= \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} e^{-\lambda}$$

$$= \sum_{k=1}^{\infty} [(k-1)+1] \frac{\lambda^k}{(k-1)!} e^{-\lambda}$$

$$= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$
$$= \lambda^2 + \lambda$$

所以：

$$Var(X) = E(X^2) - \mu^2 = \lambda$$

2.5.2 泊松分布的条件

一般地，在某一段时间T内发生特定事件的次数，如果满足以下假设，都可以看作泊松分布：

- **平稳性**：在此时间段T内，此事件发生的概率相同（在实际应用中大致相同就可以了）
- **独立性**：事件的发生彼此之间独立（或者说，关联性很弱）
- **普通性**：把T切分成足够小的区间ΔT，在ΔT内恰好发生两个、或多个事件的可能性为0（或者说，几乎为0）

> 例题：
>
> 记录1克放射性物质在1秒内放射出的α粒子数，如果从过去实验中得知，这个数目的平均值为3.2，则放出的α粒子数不超过2的概率约为：
>
> 放射出的α粒子数是服从泊松分布。
>
> 那么这里已知1秒内放射出的α粒子数平均数为3.2，也就是说：
>
> $$X \sim P(3.2)$$
>
> 这里的所求概率为：
>
> $$P(X \le 2) = \frac{3.2^0}{0!} e^{-3.2} + \frac{3.2^1}{1!} e^{-3.2} + \frac{3.2^2}{2!} e^{-3.2}$$
>
> $$\approx 0.3799$$

2.5.3 欧松定理

泊松分布实际上是二项分布的极限：

$$\lim_{n\to\infty}\binom{n}{k}\left(\frac{\mu}{n}\right)^k\left(1-\frac{\mu}{n}\right)^{n-k}=\frac{\mu^k}{k!}e^{-\mu}$$

所以在泊松分布的λ固定的情况，二项分布的n越大（对应的$p=\dfrac{\lambda}{n}$越小），此时两者会非常接近。

这也启发了我们，在二项分布n较大，p较小的时候可以用泊松分布来近似，即：

在n重伯努利实验中，记事件A在一次实验中发生的概率为p_n(与实验次数n有关)，如果当$n\to\infty$时，有$np_n\to\lambda$,则

$$\lim_{n\to\infty}\binom{n}{x}p_n^x(1-p_n)^{n-x}=\frac{\lambda^x}{x!}e^{-\lambda}$$

> 例题
>
> **1**
>
> 已知某种疾病的发病率为0.001，某单位共有5000人，问该单位患有这种疾病的人数不超过5人的概率为多少?
>
> 设该单位患有这种疾病的人数为X，则有：
>
> $$X\sim b(5000,0.001)$$
>
> 要求的概率为：
>
> $$P(X\leq 5)=\sum_{k=0}^{5}\binom{5000}{k}0.001^k0.999^{5000-k}$$
>
> 这个概率计算量很大，由于n很大，p很小，所以可以用$\lambda=np=5$的泊松分布近似：
>
> $$P(X\leq 5)\approx\sum_{k=0}^{5}\frac{5^k}{k!}e^{-5}=0.616$$
>
> **2**
>
> 有n对夫妇组成的$2n$个人,随机地分成n组，每组两人。当n足够大时，求没有妇女与她丈夫分在一组的概率的近似值。
>
> 当n很大时，根据泊松定理可知，妇女和她的丈夫结成对的数目近似服从泊松分布。
>
> 设W_i表示事件"第i个"妇女正好与她丈夫分在一组。则
>
> $$P(W_i)=\frac{1}{2n-1}$$
>
> 因此其均值近似为
>
> $$\sum_{i=1}^{n}P(W_i)=\frac{n}{2n-1}\approx\frac{1}{2}$$
>
> 设X表示配对成功的夫妻数。因此

> $$X \sim P(\frac{1}{2})$$
>
> 继而
>
> $$P(X=0)=e^{-\frac{1}{2}}$$