离散分布

这一章我们讲一些分布。关于方差和均值的证明我重新开一篇文章说明。

伯努利分布

伯努利分布的一个例子就是硬币了，因为只能取到正反面（竖着的可以忽略）。伯努利分布指出，它只取两个值，我们可以设这两个值为0和1。
那么伯努利分布可以是

> **如果随机变量X满足$Prob(X=1)=p$，$Prob(X=0)=1-p$。那么X就服从参数为$p\in[0,1]$，我们把结果1看作成功，把结果0看成失败，记作$X\sim Bern(p)$，X称为二元标示随机变量。**

既然提到了随机变量，我们是要来计算均值和方差的
那么有

$$\mu x=1*p+0*(1-p)=p$$

怎么理解呢？因为只有成功和不成功，所以是1和0。还记得我们的彩票例子吗，那个均值的计算是根据这个式子来的，1代表的是中奖的金额，后面的因为是0所以一般不写进去。不过这个公式在博弈论有很多的应用。那么我们讲讲投掷一枚硬币多次的情况

二项分布

我们在上一章讲了二项分布其实，我们在这里再整合一下。

我们考虑二项分布的主要原因是，我们有n枚硬币，每一枚硬币出现成功的概率都是p，同时抛掷他们记录正面出现的次数，这些观点都很有用。但其实都是独立的事件，所以掷硬币n次和掷n枚硬币是一样的。

那么我们为什么要加一个二项分布呢，这是因为达到结果的目的不止一种，有很多种，我们可以通过组合学求出来n种里面，k种能得到答案的结果，再去乘概率，就能得出k种结果的概率是多少。

二项分布：设n是一个正整数，并设$p \in [0,1]$如果随机变量X满足：

$$Prob(X=k)=\begin{cases}\binom{n}{k}p^k(1-p)^{n-k} & 若k\in\{1,2,...,n\}\\ 0 & 其他\end{cases}$$

那么X就是服从参数为n,p的二项分布，记为$X\sim$ $Bin(n,p)$均值是np,**方差是**$np(1-p)$

我们的二项式定理是

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^ky^{n-k}$$

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^ky^{n-k}=\sum_{k=0}^{n}P(X=k)=(p+(1-p))^n=1$$

所以确实是一个概率分布
它的均值是np，方差则是$np(1-p)$

多项分布

我本来想在3.1讲的，但是放到现在。

我们现在推出的多项式定理，是由二项式定理推得来的。
假设我们进行了n次试验，每次试验有k个不相容的结果，那么它们的概率分别为$p_1,p_2,...p_k$，我们设一个包含多变量的函数$f(x_1,x_2,...,x_n)$表示在这n次的试验中，第i种可能的结果。注意到，这是从x_1到x_n的函数。那么根据二项式定理我们知道两个变量，即$f(x,y)$的二项式定理是

$$f(x)=\binom{n}{x},\ f(y)=\binom{n-x}{y}$$

显而易见，那么$f(x,y)=f(x)f(y)$
分布为

$$f(x,y)=\binom{n}{x}\binom{n-x}{y}$$

如果我们推广到第k个变量，那么就是
设$x_1,x_2,...,x_k$为随机变量，他们的取值是

$$f(x_1,x_2....,x_k)=\binom{n}{x_1}\binom{n-x_1}{x_2}...\binom{n-x_1-x_2-,,,x_{k-1}}{x_{k-1}}\binom{n-x_1-x_2-...-x_k}{x_k}$$

我们稍微的化简一下

$$f(x_1,...,x_k)=\frac{n!}{(n-x_1)!x_1!}\cdot\frac{(n-x_1)!}{(n-x_1-x_2)!x_2!}\cdot....\cdot\frac{(n-x_1-x_2...x_{k-1})!}{(n-x_1-...-x_k)!x_k!}$$

化简得到

$$\frac{n!}{x_1!x_2!...x_k!}$$

这个东西我们叫做多项式系数，记作

$$\binom{n}{x_1,x_2,...,x_k}$$

根据二项式系数，我们知道，p_k的k是二项式分布的幂次，也就是玩了n次游戏正确的次数。那么一个系数的就是p^{x_1}，两个就是$p^{x_1}p^{x_2}$，所以我们的二项式分布其实就是求总结果里面抽取某个特定样本然后求其概率。多项式分布是一样的，只不过样本的变量多了。所以求多项式概率的式子如下

$$f(x_1,x_2,...x_k)=\frac{n!}{x_1!x_2!...x_k!}p^{x_1}...p^{x_n}$$

几何分布

这个分布是伯努利分布推广来的。你想一下，我们在搞一个东西，但是我们一直失败，没有成功，就像你抽奖那样。想一发入魂？那怎么可能。老老实实氪金，干脆直接走人。那么我们的几何分布就是描述这么个东西，描述到达第一次成功的概率。

设$p \in [0,1]$如果随机变量X满足

$$Prob(X=n)=\begin{cases}p(1-p)^{n-1} & 若n \in \{1,2,...,n\}\\ 0 & 其他\end{cases}$$

那么我们说X是服从参数为p的几何分布，均值为$\frac{1}{p}$，方差为$\frac{1-p}{p^2}$

负二项分布

我们从伯努利推广一下，我们从多次失败到第一次成功上，推广出了伯努利分布，那么第二次，第三次成功到第r次成功呢?

我们给出一个随机变量X，那么他r次成功的概率的多少，我们抛掷了n次硬币，如果n不是一个整数，或者$n \leq r-1$，那么一定有$Prob(X=n)=0$，为什么呢？因为我们的成功是r次，如果出现了抛掷的次数比成功的次数还少1次或者等于，那么明显这是不可能的，你凭空多出了1次，所以我们只考虑$n \neq r$，这是极端的情况，因为我们丢了n次硬币，刚刚好出现了r次正面或者反面。**注意的是，在最后一次抛掷中，一定是成功的**。否则在前方的$n-1$次或更少的次数中，就已经取得了r次成功。我们的问题主要求的是抛掷n次，有r次成功的概率，在前面的n次试验恰好得到了r次成功，那么这说明在前面的$n-1$次就成功了$r-1$次，那么$n-1$选$r-1$次成功就是我们的二项式系数$\binom{n-1}{r-1}$，且概率为$p^r(1-p)^{n-r}$，那么我们就得到了一个东西，它的概率密度函数

$$Prob(X=n)=\begin{cases}\binom{n-1}{r-1}p^r(1-p)^{n-r} & 若k \in \{1,2,...\}\\ 0 & 其他\end{cases}$$

它的均值是$\frac{pr}{1-p}$方差为$\frac{pr}{(1-p)^2}$，记作$X \sim NegBin(r,p)$
证明均值我们还是继续的使用我们的微分恒等式。

泊松分布

我们前面研究的离散分布跟伯努利有关，虽然接下来的跟伯努利有点关系，但没那么大。泊松分布其实可以被定义成n,p的二项分布的极限，其中$n \to \infty$且$np_n \to \lambda$，p_n不是一个常数。给出泊松分布的定义:

$$Prob(X=n)=\begin{cases}\lambda^n e^{-\lambda}/n! & 若n \in \{0,1,2,...\}\\ 0 & 其他\end{cases}$$

那么X就服从参数为λ的泊松分布，记为X~$Pois(\lambda)$,均值和方差都是λ

这是一个明显的概率分布，我们可以利用泰勒级数求。

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

那么有

$$\sum P(X) = \sum \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1$$

所以是一个概率分布。

那么这个λ怎么算呢？，我们知道二项式分布的期望是$np = \mu$，那么反过来$p = \frac{\mu}{n}$
我们只需要把p替换成$\frac{\mu}{n}$就行了那么我们用如下极限

$$\lim_{n\to\infty} \binom{n}{k} (\frac{\mu}{k})^k (1 - \frac{\mu}{n})^{n-k} = \lim_{x\to\infty}$$

那么我们根据重要极限分一下

$$\lim_{n\to\infty} \binom{n}{k} (\frac{\mu}{n})^k (1 - \frac{\mu}{n})^{-k} (1 - \frac{\mu}{n})^{n} = \lim_{x\to\infty} = e^{-u} \lim_{n\to\infty} \binom{n}{k} (\frac{\mu}{n})^k (1 - \frac{\mu}{n})^{-k}$$

注意右边的等式，在$e^{-\mu}$右边的等式中$\binom{n}{k} (1 - \frac{\mu}{n})^{-k}$的极限$= 1$，然后我们用λ替换$\mu$就得到我们的公式了，注意到二项式系数是

$$\binom{n}{k} = \frac{n(n-1)...(n-k+1)}{k!}$$

因为是乘法，我们把k和n替换变个位置为

$$\frac{n(n-1)...(n-k+1)}{n^k} \frac{u^k}{k!}$$

左边的极限是1，右边就变成一个常数了。那么

$$\lim_{n\to\infty}(1-\frac{\mu}{n})^k=1$$

最后的结果就是

$$e^{-\mu}\frac{\mu^k}{k!}$$

把μ替换成λ，k换成n就是我们的泊松分布了。

离散均匀分布

我们现在考察最后一种对伯努利分布的推广，我们考察的是对于有n个结果，发生的概率为$\frac{1}{n}$
那么：

$$Prob(X=a)=\begin{cases}\frac{1}{n} & 若a\in\{a_1,a_2,...,a_n\}\\ 0 & 其他\end{cases}$$

那么X就是一个服从离散均匀分的离散变量。X的均值是$a+\frac{n-1}{2}$方差为$\frac{n^2-1}{12}$

德国坦克问题

在二战内，西方盟军试图统计德国人制造的坦克数量，为此他们收集了被摧毁的坦克序列号，分析坦克轮子，估算当时使用多少种车轮模具，利用这些信息成功的预计了有大约270辆坦克被生产出来，事实上德军一共制作了276辆坦克。这种分析方法预测了**无放回抽样**的最大值，我们现在抽象成数学问题，设一序列的范围在1到N上。而我们想求出N，

首先我们用N表示坦克总数，那么N是最大的序列号，1是最小的序列号，假设我们记录了k个序列号，其中m是我们记录到的最大的序列号，那么现在我们知道k，m想求出N，显然，m至少应该和N一样大，因为几乎可以肯定甚至会更大。难点在于，最优解N为多少呢?

接下来我们用N表示M的期望值，从$\{1,2,...,N\}$中取出k个观测值，随机变量M表示这k个观测值中的最大值。这意味着我们必须选择m而不是比m更大的数。求$Pr(M=m)$，我们首先注意到从N选k种可能有$\binom{N}{k}$种方法，最大值为m的k元组有$\binom{m-1}{k-1}$，因为我们的最大值设为$m=M$，所以必须是1~$(m-1)$中选一个。因为是不放回抽样，我们抽走一个，剩下$m-1$种抽$k-1$的可能。记得我前面讲的圆桌问题吗，这是一样的，我们固定了一个数，这样子才能够进行抽样，否则会无限循环下去。所以是$(m-1)$

为了方便，我们得重新讲讲圆桌问题，我们知道有一张圆桌，有a,b,c,d,e五个人。那么我们让他们按顺序进坐，那么第一个人可以选一个进入，剩下的只能在b,c,d,e中选择一个，那么我们知道，我们可以通过旋转让某个人在某个特定的位置，所以我们只需要计算接下来4个人的排序方法就能的出来5个人排序的圆桌问题概率。对于德国坦克问题也是一样的，我们固定第一个人的选择方法，所以是其他四个人的排序跟上被固定的一个人的排列方法，所以我们只需要在接下来的$m-1$中寻找人就行了，其实不必考虑每个数字的随机位置（如果真的纠结，其实利用平移我们也能够得到各种组合是一样的。），因为不放回抽样，他们的概率是一样的。也就是$1\times(m-1)!$。否则我们得到循环计数。当我们想从m选k个数的时候，我们已经固定了一个人，那么剩下的人里面选择自然就少了一种选择，就是$k-1$所以我们有

$$P(M=m)=\frac{1\times\binom{m-1}{k-1}}{\binom{N}{k}}，那么\sum_{m=k}^{N}\frac{\binom{m-1}{k-1}}{\binom{N}{k}}=1$$

那么我们开始计算期望值

$$E[M]=\sum_{m=k}^{N}mPr(M=m)=m\sum_{m=k}^{n}\frac{\binom{m-1}{k-1}}{\binom{N}{k}}$$

$$=\sum_{m=k}^{N}\frac{m\frac{(m-1)!}{(k-1)!(m-k)!}}{\frac{N!}{k!(N-k)!}}$$

由于

$$(m-1)!=(m-1)(m-2)...(m-m+1)，而m!=m(m-1)(m-2)...(m-m+1)$$

所以$m\times(m-1)!=m(m-1)(m-2)...(m-m+1)=m!$

$$=\sum_{m=k}^{N}\frac{\frac{m!}{(k-1)!(m-k)!}}{\frac{N!}{k!(N-k)!}}$$

那么我们该如何简化这个式子呢？我们知道概率和等于1，我们令公式以一种奇妙的方式乘1。最后利用概率和为1来化简式子，我们从$N+1$个结果中取出$k+1$个样本，我们要令其变成$\binom{m}{k}$这种形式，所以乘上一个$\frac{k}{k}$，这样子就不会破坏期望值乘1了。下面也是，乘一个1变成$N+1$,因为这个期望的结果很像$k+1$的观测值中最大序列号为$m+1$的概率，所以我们要来计算这点。等式变为

$$\sum_{m=k}^{N}\frac{\frac{m!}{(k-1)!(m-k)!}\frac{k}{k}}{\frac{N!}{k!(N-k)!}\frac{k+1}{k+1}\frac{N+1}{N+1}}=\sum_{m=k}^{N}\frac{\binom{m}{k}k}{\binom{N+1}{k+1}\frac{k+1}{N+1}}=\frac{k}{k+1}(N+1)\sum_{m=k}^{N}\frac{\binom{m}{k}}{\binom{N+1}{k+1}}$$

因为我们知道

$$\sum_{m=k}^{N}\frac{\binom{m-1}{k-1}}{\binom{N}{k}}=1$$

等式变为

$$1\times\frac{k}{k+1}(N+1),\ N=\frac{k+1}{k}m-1$$

那么我们就求出了期望值，可以根据序列号反推大概生产了多少量坦克。