Variáveis Aleatórias Discretas Parte II

Distribuição Hipergeométrica e Distribuição Poisson

Caso Multivariado Discreto

A variável aleatória X com distribuição Binomial(n,p) motiva duas outras variáveis aleatórias discretas importantes: Hipergeométrica e Poisson

Distribuição Hipergeométrica: *"sorteio de uma população finita"*

Suponha que 20 % dos indivíduos de uma população têm certa característica. Se sorteio, ao acaso, 5 deles. Qual é a probabilidade de k (k entre 0 e 5) terem essa característica?

Temos distribuição binomial, com n = 5 e p = 0,2?

Os sorteios são independentes?

Se a população for finita, não! Pois os sorteios não tem reposição.

(Por exemplo, suponha que "a população" tenha apenas 5 indivíduos...)

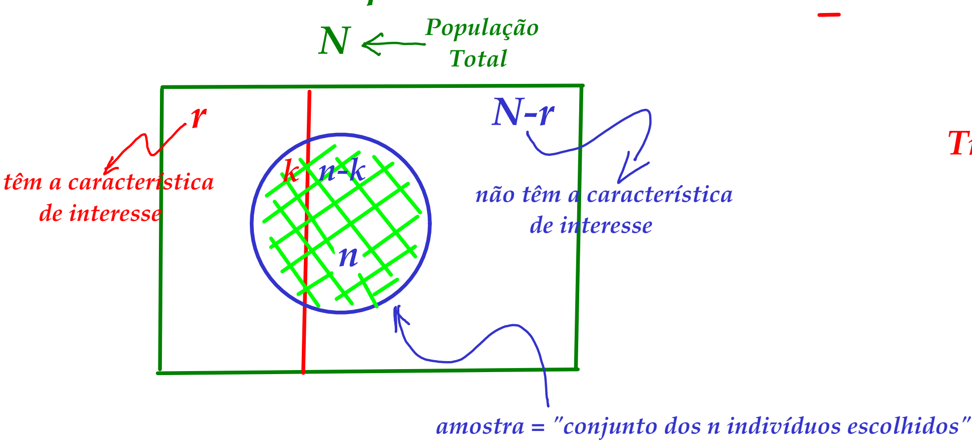
Suponha, por outro lado, que o grupo considerado seja "a População Brasileira" (mais de 200 milhões de indivíduos, e portanto "quase infinita").

Agora, para todos os efeitos práticos, os sorteios SÃO independentes! Temos (na prática) uma distribuição binomial com n=5 e p=0,2

A distribuição binomial corresponde, portanto, ao caso de um sorteio de uma "população infinita" (ou "muito grande"), uma situação comum na prática.

Sorteio de População finita:

Suponha uma população com N indivíduos, dos quais r tem certa característica que me interessa. Sorteio n deles (n ≤N) ao acaso.



Três parâmetros: N r n

Qual é a probabilidade de encontrar k sucessos (k indivíduos com a característica de interesse) nesta amostra de tamanho n?

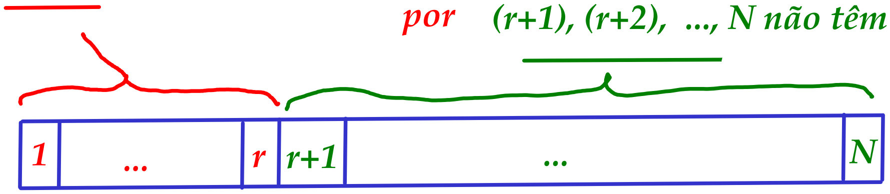
Ou, o que é a mesma coisa, queremos a distribuição de probabilidade da variável aleatória X, definida por

X = "número de indivíduos com a 'característica de interesse' encontrados na amostra de tamanho n escolhida ao acaso dessa população de tamanho N, das quais r têm essa característica".

Claro que a distribuição de probabilidade de X só pode ser encontrada após uma definição precisa do experimento aleatório em questão.

O experimento aleatório é : "escolho ao acaso n elementos de certo conjunto de N objetos dos quais r têm certa característica; estou interessado no evento 'k deles têm essa característica', k com valores de zero a n"

Temos N elementos identificados por 1, 2, ... N; os elementos identificados por 1,2,... r têm certa característica de interesse; os elementos identificados por (r+1), (r+2), ..., N não têm esta característica.



Definição do experimento aleatório:

Espaço amostral = "conjunto de todas as $\binom{N}{n}$ *escolhas de n desses N elementos"*

Probabilidades: "todas as possíveis escolhas têm a mesma probabilidade"

Ou seja, todas as $\binom{N}{n}$ *possíveis escolhas de n elementos de N, onde*

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$$

têm a mesma probabilidade (distribuição uniforme).

Portanto se A é um evento qualquer

$$P(A) = \frac{|A|}{|\Omega|} \qquad \text{onde} \qquad |\Omega| = \binom{N}{n}$$

Com isso podemos agora encontrar a distribuição de probabilidades da variável aleatória X definida acima:

X = "número de indivíduos com certa 'característica de interesse' encontrados numa amostra de tamanho n escolhida ao acaso dessa população de tamanho N, das quais r têm essa característica".

$|\{X=k\}|$ = número de maneiras de selecionar n objetos dentre os N tal que encontramos k com a caracteristica de interesse; dos N, r tem essa característica

Temos um "problema de contagem".

Para definir uma particular escolha em $\{X=k\}|$ temos que

1) definir quais são os k elementos escolhidos dentre os r possíveis; em seguida

2) definir quais são os n-k elementos restantes, escolhidos dentre os N-r possíveis.

Essas duas escolhas "não interferem em nada uma com a outra": são independentes.

Para a escolha 1) temos $\binom{r}{k}$ *possibilidades e*

Para a escolha 2) temos $\binom{N-r}{n-k}$ *possibilidades.*

Portanto

$$P(\{X=k\}) = \frac{|\{X=k\}|}{\binom{N}{n}} = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

Desde que k, $k \geqslant 0$, satisfaça

$$k \leqslant \min\{ n , r \} \quad e$$
$$n-k \leqslant N - r \quad ou\ seja\ k \geqslant n - N + r$$

Ou seja: $\underbrace{\max\{ 0, n - N + r \}}_{0} \leqslant k \leqslant \underbrace{\min\{ n , r \}}_{n}$

Situação usual: $0 \leqslant k \leqslant n$

Distribuição de Poisson:

Suponha que Y é uma v. a. com distribuição Binomial com parâmetros n e p

Y ~ Binomial (n,p) ou seja

$$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \le k \le n.$$

Uma situação interessante na prática: suponha que bits de informação (sinais de 0 ou de 1) são transmitidos sucessivamente a cada s segundos numa conexão wireless. Em situações concretas, é claro, s é um número MUITO pequeno. Cada sinal tem uma probabilidade p de ser transmitido errado. Suponha que erros são independentes, ou seja, cada sinal é transmitido corretamente, ou não, independentemente uns dos outros.

Então, se

Y = número de sinais transmitidos erradamente após n segundos, temos

Y ~Binomial(n, p)

Numa situação concreta, teriamos n muito grande e p muito pequeno, MAS com E(Y) = np (o número esperado de erros) assumindo certo valor, nem "muito grande" nem "muito pequeno".

Distribuição Posson:

n MUITO grande e
p MUITO pequeno
mas com np = $\lambda \in \mathbb{R}$

(letra grega lambda minúscula)

Matematicamente: o que acontece com a distribuição de probabilidade Binomial(n,p) no limite onde n cresce para infinito, de tal forma que np está fixo?

Definição: dizemos que X ~Poisson(λ) se

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad ; \quad k \geq 0$$

Exercício: verifique que

$$\sum_{k\geq 0} P(X=k) = \sum_{k\geq 0} \frac{e^{-\lambda}\lambda^k}{k!} = 1$$

Sugestão: use um resultado que você já deve ter aprendido num curso de análise:

$$\sum_{k\geq 0} \frac{x^k}{k!} = e^x \quad ; \quad \text{para todo } x \in \mathbb{R}$$

Fórmula da Poisson como limite da binomial.

Seja Y ~ Binomial (n,p)

$$\lim_{\substack{n\to\infty \\ np=\lambda}} P(Y=k) = \lim_{\substack{n\to\infty \\ np=\lambda}} \frac{n!}{(n-k)!\,k!} p^k (1-p)^{n-k}$$

$$= \lim_{n\to\infty} \left(\frac{1}{k!} \underbrace{\overbrace{n.(n-1)\dots(n-k+1)}^{k \text{ termos}} \frac{\lambda^k}{n^k}}_{1.(1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{k-1}{n})\lambda^k} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} \right)$$

$$\left(p = \frac{\lambda}{n}\right)$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1.(1-\tfrac{1}{n})(1-\tfrac{2}{n})\dots(1-\tfrac{k-1}{n}) \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} \right)$$

mas $\displaystyle \lim_{n\to\infty} \underbrace{1.(1-\tfrac{1}{n})(1-\tfrac{2}{n})\dots(1-\tfrac{k-1}{n})}_{k \text{ termos}} = 1$ (Porquê?)

$$\lim_{n\to\infty} \underbrace{\left(1-\frac{\lambda}{n}\right)^k}_{k \text{ termos}} = 1 \quad (\text{Porquê?})$$

$$e \quad \lim_{n\to\infty} \underbrace{\left(1-\frac{\lambda}{n}\right)^n} = e^{-\lambda} \quad (\text{Porquê?})$$

Uma demonstração usa o binômio de Newton e a identidade ✡ acima.

note que há n termos menores que 1 que tendem para 1 quando n cresce.

Então

$$\lim_{\substack{n\to\infty \\ np=\lambda}} P(y=k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \geq 0$$

Que é a distribuição Poisson com parâmetro $\lambda > 0$

Valor esperado da Poisson:

Se X ~Poisson(λ ***)***

então E(X) = λ

Demonstração: $\boxed{EX} = \sum_{k\geq 0} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k\geq 1} \frac{\lambda^{k-1}}{(k-1)!}$

$$= \boxed{\lambda}$$

pois $\sum_{k\geq 1} \frac{\lambda^{k-1}}{(k-1)!} = \sum_{\ell=0}^{\infty} \frac{\lambda^\ell}{\ell!} = e^{\lambda}$ (por ✡ acima)

Se X é uma variável aleatória, como medir o "típico erro entre X e seu valor esperado"?

Variância: *Se X é uma v.a. sua variância é definida por*

$$\mathrm{var}(X) = E\left[(X - EX)^2\right]$$

ou seja, var(X) é o valor esperado do "erro quadrático de X em relação ao seu valor esperado"

Exercício: verifique que

$$\mathrm{var}(X) = E(X^2) - (EX)^2$$

e, se a e b são constantes

$$\mathrm{var}(ax+b) = a^2\,\mathrm{var}(x)$$

Definição:

Desvio padrão de X é a raiz quadrada de sua variância, var(X).

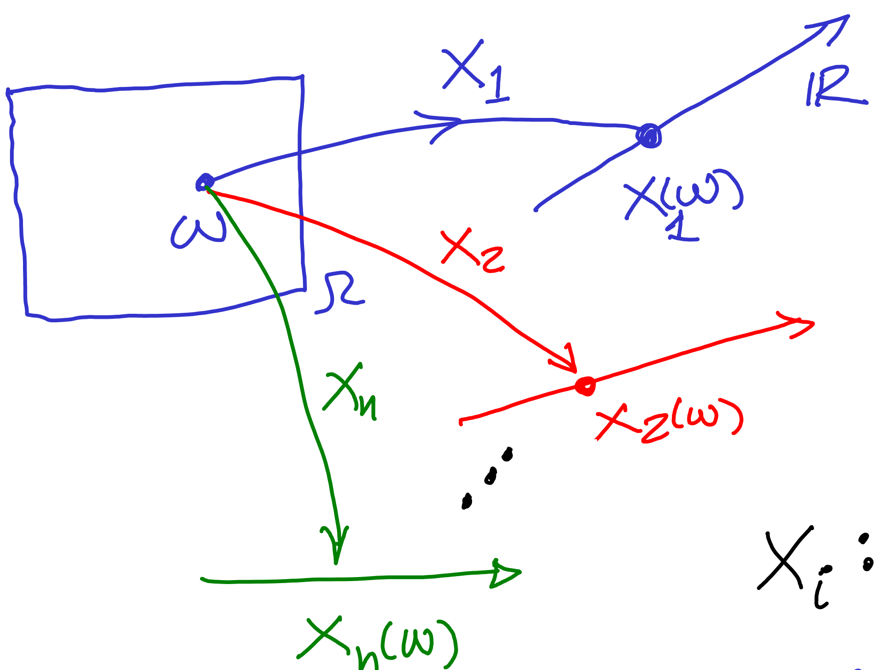
Coleção de Variaveis aleatórias.

(caso multivariado)

Uma situação usual é que estejamos interessados em mais que uma variável aleatória definida a partir do mesmo experimento aleatório.

Vamos começar supondo que todas essas variáveis aleatórias são discretas.

Sejam X_1 , X_2 , ... X_n definidas no mesmo experimento aleatório.



$X_i : \Omega \to \mathbb{R}$

$1 \leq i \leq n$

n variáveis aleatórias

Exemplo: Lanço, independentemente, dois dados honestos, um azul, outro verde.
Posso definir diversas variáveis aleatórias. Por exemplo:

X_1 = *"resultado do dado azul"*
X_2 = *"resultado do dado verde"*
X_3 = *"soma dos valores dos dois dados"*
$X_4 = \begin{cases} 1 \text{ se os dois resultados forem pares} \\ 0\text{, caso contrário} \end{cases}$
$X_5 = \begin{cases} 1 \text{ se os dois resultados forem iguais} \\ 0\text{, caso contrário.} \end{cases}$

Estamos considerando o caso discreto, ou seja todas as v.a. definidas no mesmo experimento aleatório são discretas.

Para fixar a notação, suponha que a variável aleatória X_i , para $i = 1 \ldots n$, só pode assumir os valores

$$x_1^i, x_2^i, x_3^i, \ldots$$

Definição: ***Distribuição de Probabilidades Conjunta das v.a.***

$X_1, X_2, \ldots X_n$ ***é a coleção de probabilidades***

$$\left\{ P(X_1 = x^1_{k_1}, \ldots, X_n = x^n_{k_n}) \right\}$$

para todos os possíveis valores $x^i_{k_i}$ ***, i = 1, ... , n.***

Exercício: Mostre que, se $X_1, X_2, \ldots, X_n$ é uma coleção qualquer de variáveis aleatórias e $a, a_1, a_2, \ldots, a_n$ são números reais então

$$E\left(a+\sum_{i=1}^{n} a_i X_i\right) = a + \sum_{i=1}^{n} a_i E X_i$$

Esta propriedade ("linearidade do valor esperado") vale para qualquer combinação linear de variáveis aleatórias.

Que podemos dizer sobre $\operatorname{var}\left(a+\sum_{i=1}^{n} a_i X_i\right)$?

Em particular: se X e Y são duas varíaveis aleatórias que posso dizer sobre var(X+Y)?

Definição: As variáveis aleatórias $X_1, X_2, \ldots X_n$ *definidas no mesmo experimento aleatório, são ditas Independentes se*

$$P(X_1 = x^1_{k_1}, \ldots, X_n = x^n_{k_n}) = \prod_{i=1}^{n} P(X_i = x^i_{k_i})$$

para toda coleção de possíveis valores $x^i_{k_i}$ *de cada v.a.* X_i, *para i =1, ..., n.*

Obs: Note que esta definição exige que os eventos

$$\{X_i = x^i_{k_i}\}, \quad 1 \le i \le n$$

sejam independentes, para todos os possíveis valores assumidos por estas variáveis aleatórias.

Exercício: considere o exemplo acima, do lançamento independente de dois dados honestos, um azul, outro verde. Verifique que $\{X_1, X_2\}$ é uma coleção (bivariada) de v.a. independentes enquanto que $\{X_1, X_2\ X_3\}$ não é uma coleção de variáveis independentes. E o par $\{X_1\ X_5\}$?

Resultado importante: Se duas variáveis aleatórias X e Y são independentes então:

$$E(X\,Y) = E(X)\,E(Y)$$

Demonstração: *(Mostramos para o caso discreto; mas vale em geral)*

Suponha que X só pode assumir valores

$$x_1, x_2, \ldots$$

e Y só pode assumir valores

$$y_1, y_2, \ldots$$

Então os possíveis valores da v.a. **$W=XY$** *são*

$$x_i y_j \,, \quad \textit{para } i \geq 1 \text{ e } j \geq j$$

Então, por definição

$$E(XY) = \sum_{i,j} x_i y_j P(X=x_i, Y=x_j)$$

soma sobre todos os valores de i e j

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j P(X=x_i) P(Y=y_j)$$

$$= \sum_{i=1}^{\infty} x_i P(X=x_i) \underbrace{\left(\sum_{j=1}^{\infty} x_j P(Y=y_j) \right)}_{EY}$$

$$= \underbrace{\sum_{i=1}^{\infty} x_i P(X=x_i)}_{EX} \; EY$$

$$= EX \cdot EY$$

Aplicações deste resultado.

Cálculo simples do valor esperado da Binomial(n,p).

Sejam $X_1, X_2, \ldots, X_n$ ***, v.a. independentes***

com X_i ***~ Bernoulli(p), para i =1, ..., n.***

Seja $X = X_1 + \ldots + X_n$

Exercício: Mostre que X ~ Binomial(n,p)

Exercício: Use a linearidade do valor esperado, e o valor esperado da Bernoulli(p) ($E(X_i) = p$***, para todo i=1, ..., n) para mostrar que***

$$E(X) = n\,p.$$

Note que NÃO preciso usar a independência entre essas v.a. Bernoulli.

Se X ~ Binomial(n,p), qual é a variância de X?

$$var(X) = E(X - EX)^2$$

$$= E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i)\right)^2\right]$$

$$= E\left[\sum_{i=1}^{n}(X_i - \underbrace{EX_i}_{p})\right]^2$$

$$= \sum_{i=1}^{n} \underbrace{E(X_i - p)^2}_{p(1-p)} + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \underbrace{E\left[(X_i - p)(X_j - p)\right]}_{\substack{(\text{indep}) \\ \underbrace{E(X_i - p)}_{0}\underbrace{E(X_j - p)}_{0}}}$$

$$= np(1-p)$$

Então se X ~ Binomial(n , p)

E(X) = np   e   var(X) = np(1-p)

Caso Bivariado.

Considere duas variáveis aleatórias X e Y definidas no mesmo experimento aleatório.

Definição: *Covariância entre X e Y*

$$cov(X,Y) = E\left[(X - EX)(Y - EY)\right]$$

Como o nome sugere, a covariância entre X e Y procura mensurar uma relação simples entre as variações conjuntas de duas variáveis aleatórias.

A covariância, cov(X,Y), pode assumir valores reais quaisquer. Uma situação interessante é quando cov(X,Y)=0, quando dizemos que as duas variáveis aleatórias são "não correlacionadas".

Exercício: Mostre que, se X e Y são independentes, então

$$cov(X,Y) = 0$$

Vale a relação inversa?

Não é verdade que cov(X,Y)=0 implique independência.

Podemos ver isso por um exemplo:

Sejam X e Y v.a. que só podem assumir valores -1, 0 ,1, com a seguinte distribuição conjunta:

Y \ X	-1	0	1
-1	0	1/4	0
0	1/4	0	1/4
+1	0	1/4	0

P(X=1,Y=-1)

P(X=1,Y=0)

etc...

Exercício: determine as distribuições de X e de Y isoladamente (distribuições marginais)

Exercício: X e Y são independentes?

Exercício: determine cov(X,Y).

Exercício: Mostre que, se X e Y são variáveis aleatórias discretas quaisquer então

$$var(X+Y) = var(X) + var(Y) + 2\,cov(X,Y)$$

Exercício: Escreva X ~ Hipergeométrica(N,r,n) vista acima como soma de v.a. Bernoulli(p). Essas v.a. Bernoulli serão independentes? Use esta decomposição para determinar E(X).