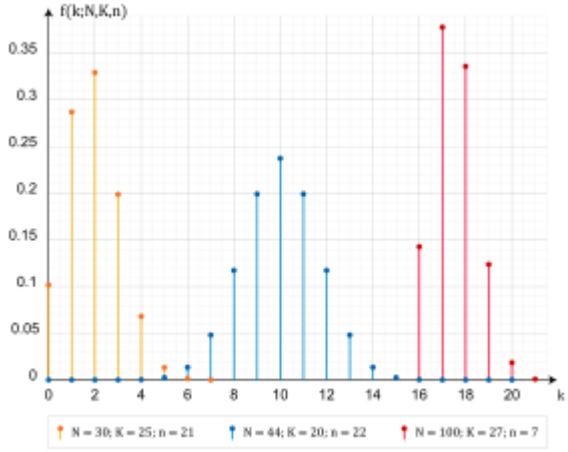
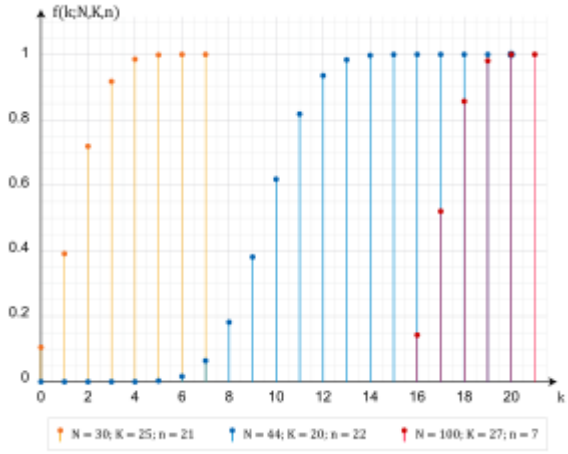
Distribuição hipergeométrica

Distribuição hipergeométrica



Função distribuição de probabilidade para alguns valores de N,K e n



Função distribuição acumulada para alguns valores de N,K e n

Parâmetros	$N \in \{0,1,2,\ldots\}$ $K \in \{0,1,2,\ldots,N\}$ $n \in \{0,1,2,\ldots,N\}$
Suporte	$k \in \{\max(0, n+K-N), \ldots, \min(n, K)\}$
f.d.p.	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
f.d.a.	$1 - \frac{\binom{n}{k+1}\binom{N-n}{K-k-1}}{\binom{N}{K}} \, {}_3F_2\left[\begin{matrix} 1,\ k+1-K,\ k+1-n \\ k+2,\ N+k+2-K-n \end{matrix}; 1\right]$, em que ${}_pF_q$ é a função hipergeométrica generalizada
Média	$n\frac{K}{N}$

Moda	$\left\lfloor \frac{(n+1)(K+1)}{N+2} \right\rfloor$
Variância	$n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$
Obliquidade	$\frac{(N-2K)(N-1)^{\frac{1}{2}}(N-2n)}{[nK(N-K)(N-n)]^{\frac{1}{2}}(N-2)}$
Curtose	$\frac{1}{nK(N-K)(N-n)(N-2)(N-3)}\cdot \left[(N-1)N^2\left(N(N+1)-6K(N-K)-6n(N-n)\right)+ 6nK(N-K)(N-n)(5N-6)\right]$
Função Geradora de Momentos	$\frac{\binom{N-K}{n}\,{}_2F_1(-n,-K;N-K-n+1;e^t)}{\binom{N}{n}}$
Função Característica	$\frac{\binom{N-K}{n}\,{}_2F_1(-n,-K;N-K-n+1;e^{it})}{\binom{N}{n}}$

Em teoria das probabilidades e estatística, a **distribuição hipergeométrica** é uma distribuição de probabilidade discreta que descreve a probabilidade de k sucessos em n retiradas, sem reposição, de uma população de tamanho N que contém exatamente K sucessos, sendo cada retirada um sucesso ou um fracasso. Em contraste, a distribuição binomial descreve a probabilidade de k sucessos em n retiradas com reposição.

Em estatística, o **teste hipergeométrico** usa a distribuição hipergeométrica para calcular a significância estatística de obtenção de um número específico k de sucessos (a partir de um total de n retiradas) a partir da população acima mencionada. O teste é frequentemente usado para identificar quais subpopulações estão super-representadas ou sub-representadas em um amostra. Por exemplo, um grupo de *marketing* poderia usar o teste para compreender sua base de consumidores ao testar um conjunto de consumidores desconhecidos para avaliar a super-representação de vários subgrupos demográficos (como mulheres ou pessoas abaixo de 30).

Índice

Definição

Identidades combinatórias

Aplicação e exemplo

Aplicação no *Texas hold 'em*

Simetrias

Teste hipergeométrico

Relação com o teste exato de Fisher

Ordem das retiradas

Distribuições relacionadas

Limites de cauda

Distribuição hipergeométrica multivariada

Exemplo

Ver também

Referências

Ligações externas

Definição

As seguintes condições caracterizam a distribuição hipergeométrica:

- O resultado de cada retirada (os elementos da população que compõem a amostra) pode ser classificado em uma de duas categorias mutuamente excludentes (por exemplo, aprovação ou reprovação, empregado ou desempregado);
- A probabilidade de um sucesso muda a cada retirada, conforme cada retirada diminui a população (amostragem sem reposição a partir de uma população finita).

Uma variável aleatória X segue a distribuição hipergeométrica se a função massa de probabilidade for dada por[1]

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

em que

- N é o tamanho da população,
- K é o número de estados de sucessos na população,
- n é o número de retiradas,
- k é o número de sucessos observados,
- $\binom{a}{b}$ é um coeficiente binomial.

A função massa de probabilidade é positiva quando $\max(0, n+K-N) \leq k \leq \min(K, n)$.

A função massa de probabilidade satisfaz a relação de recorrência

$$(k+1)(N-K-(n-k-1))P(X=k+1) = (K-k)(n-k)P(X=k)$$

com

$$P(X=0) = \frac{\binom{N-K}{n}}{\binom{N}{n}}.$$

Identidades combinatórias

Como é de se esperar, a soma das probabilidades resulta em 1:

$$\sum_{0\le k\le n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = 1$$

Esta é essencialmente a identidade de Vandermonde da combinatória.

A seguinte identidade também se aplica:

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}.$$

Isto segue da simetria do problema, mas isto também pode ser mostrado expressando os coeficientes binomiais em termos de fatoriais e rearranjando os últimos.[2]

Aplicação e exemplo

A aplicação clássica da distribuição hipergeométrica é a amostragem sem reposição. Suponha uma urna com dois tipos de bolas, vermelhas e verdes. Defina a retirada de uma bola verde como um sucesso e a retirada de uma bola vermelha como um fracasso (o que é análogo à distribuição binomial). Se a variável N descrever o número de todas as bolas na urna e K descrever o número de bolas verdes, então $N - K$ corresponde ao número de bolas vermelhas. Neste exemplo, X é a variável aleatória cujo valor observado é k, o número de bolas verdes retiradas no experimento. Esta situação é ilustrada pela seguinte tabela de contingência:

	Retiradas	Não retiradas	Total
Bolas verdes	k	$K-k$	K
Bolas vermelhas	$n-k$	$N+k-n-K$	$N-K$
Total	n	$N-n$	N

Agora, assuma, por exemplo, que há 5 bolas verdes e 45 bolas vermelhas na urna. De pé ao lado da urna, você fecha seus olhos e retira 10 bolas sem reposição. Qual é a probabilidade de que exatamente 4 das 10 sejam verdes? Note que, apesar de estarmos observando sucessos e fracassos, os dados não são precisamente modelados pela distribuição binomial, porque a probabilidade de sucesso em cada triagem não é a mesma, já que o tamanho da população remanescente muda conforme removemos cada bola.

O problema está resumido pela seguinte tabela de contingência:

	Retiradas	Não retiradas	Total
Bolas verdes	$k=4$	$K-k=1$	$K=5$
Bolas vermelhas	$n-k=6$	$N+k-n-K=39$	$N-K=45$
Total	$n=10$	$N-n=40$	$N=50$

A probabilidade de retirar exatamente k bolas verdes pode ser calculada pela fórmula

$$P(X=k) = f(k;N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Assim, neste exemplo, calcula-se

$$P(X = 4) = f(4; 50, 5, 10) = \frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}} = \frac{5 \cdot 8145060}{10272278170} = 0.003964583\ldots.$$

Intuitivamente, é ainda mais improvável que todas as cinco bolas sejam verdes.

$$P(X = 5) = f(5; 50, 5, 10) = \frac{\binom{5}{5}\binom{45}{5}}{\binom{50}{10}} = \frac{1 \cdot 1221759}{10272278170} = 0.0001189375\ldots.$$

Conforme esperado, a probabilidade de retirar cinco bolas verdes é aproximadamente 35 vezes menor do que a probabilidade de retirar 4 bolas verdes.

Outro exemplo se refere a um jogo de loteria que consiste em selecionar seis números de um conjunto de cem, que vão de de 00 a 99, com uma bola para cada número e sem reposição. Em um cartão de aposta, o jogador pode escolher de 6 a 12 números. Qual é a probabilidade de que o jogador acerte a quina, ou seja, cinco números, ao marcar 10 números no volante? Temos

- N: total de números, $N = 100$;
- n: total de números sorteados, $n = 6$;
- K: total de números escolhidos, $K = 10$;
- X: total de sucessos, deseja-se $X = 5$.

$$P(X = 5|100, 10, 6) = \frac{\binom{10}{5}\binom{100-10}{6-5}}{\binom{100}{6}} = \frac{252 * 90}{1.192.052.400} = 0,000019.$$

A probabilidade de que o jogador acerte a quina é de aproximadamente 0,000019%.

O mesmo problema pode ser resolvido de outra forma. Pode-se pensar que a escolha aleatória é feita pelo jogador, mas que os números "premiados" já estão definidos *a priori*, sem que o jogador saiba. Logo, existem dois tipos de números, os "premiados" e os "não premiados". O jogador escolhe aleatoriamente (ou não, desde que seu critério de escolha seja independente dos números "premiados") os 10 números do seu jogo. Assim:

- N: total de números, $N = 100$;
- n: total de números sorteados/escolhidos pelo jogador, $n = 10$;
- K: total de números premiados, $K = 6$;
- X: total de sucessos, deseja-se $X = 5$.

$$P(X = 5|100, 6, 10) = \frac{\binom{6}{5}\binom{100-6}{10-5}}{\binom{100}{10}} = \frac{6 * 54.891.018}{17.310.309.456.440} = 0,000019.$$

O resultado é o mesmo.

Aplicação no *Texas hold 'em*

No pôquer *Texas hold 'em*, jogadores fazer a melhor mão que podem combinando duas cartas em suas mãos com as cinco cartas (cartas comunitárias) eventualmente distribuídas sobre a mesa. O baralho tem 52 cartas, 13 de cada naipe. Para este exemplo, assuma que um jogador tem duas cartas de paus na mão e há três cartas na mesa, duas das quais também são de paus. O jogador gostaria de saber a probabilidade de que uma das duas próximas cartas a serem mostradas seja uma carta de paus para completar o *flush*.

Note que as chances calculadas neste exemplo assumem que nenhuma informação é conhecida sobre as cartas nas mãos dos outros jogadores. Entretanto, jogadores de pôquer experientes podem levar em conta como outros jogadores fazem suas apostas ao considerar as probabilidades para cada cenário. Estritamente falando, a abordagem ao calcular probabilidades de sucesso aqui descrita é precisa em um cenário em que há apenas um jogador na mesa. Em uma partida com vários jogadores, estas probabilidades podem ser ajustadas de alguma forma com base nas apostas dos oponentes.

Há quatro cartas de paus à mostra, então há nove cartas de paus ocultas. Há cinco cartas à mostra (duas na mão e três na mesa, então há $52 - 5 = 47$ ainda ocultas.

A probabilidade de que uma das duas próximas cartas a serem mostradas seja uma carta de paus pode ser calculada usando a hipergeométrica $k = 1$, $n = 2$, $K = 9$ e $N = 47$, sendo cerca de 31,6%.

A probabilidade de que as duas próximas cartas a serem mostradas sejam duas cartas de paus pode ser calculada usando a hipergeométrica $k = 2$, $n = 2$, $K = 9$ e $N = 47$, sendo cerca de 3,3%.

A probabilidade de que nenhuma das duas próximas cartas a serem mostradas seja uma carta de paus pode ser calculada usando a hipergeométrica $k = 0$, $n = 2$, $K = 9$ e $N = 47$, sendo cerca de 65,0%.

Simetrias

Invertendo os atributos das bolas verdes e vermelhas, temos:

$$f(k; N, K, n) = f(n - k; N, N - K, n).$$

Invertendo os atributos das bolas retiradas e não retiradas, temos:

$$f(k; N, K, n) = f(K - k; N, K, N - n).$$

Invertendo os atributos das bolas verdes e retiradas, temos:

$$f(k; N, K, n) = f(k; N, n, K).$$

Teste hipergeométrico

O teste hipergeométrico usa a distribuição hipergeométrica para medir a significância estatística da obtenção de uma amostra que consiste de um número específico de k sucessos (dentre um total n de retiradas) a partir de uma população de tamanho N contendo K sucessos. Em um teste para a super-representação de sucessos na amostra, o valor-p hipergeométrico é calculado como a probabilidade de obter aleatoriamente k ou mais sucessos a partir da população em um total n de retiradas. Em um teste para sub-representação, o valor-p é a probabilidade de obter aleatoriamente k ou menos sucessos.

Relação com o teste exato de Fisher

O teste baseado na distribuição hipergeométrica, o teste hipergeométrico, é idêntico à versão unicaudal correspondente do teste exato de Fisher.[3] Reciprocamente, o valor-p de um teste exato de Fisher bicaudal pode ser calculada como a soma de dois testes hipergeométricos apropriados.[4]

O biólogo e estatístico britânico Ronald Fisher

Ordem das retiradas

A probabilidade de retirar qualquer sequência de bolas brancas e pretas, a distribuição hipergeométrica, depende apenas do número de bolas brancas e pretas, não da ordem em que elas aparecem, isto é, é uma distribuição intercambiável. Como resultado, a probabilidade de retirar uma bola branca na i-ésima retirada[5]

$$P(W_i) = \frac{K}{N}.$$

Distribuições relacionadas

Considere $X \sim$ Hipergeométrica (K, N, n) e $p = K/N$.

- Se $n = 1$, então, X tem uma distribuição de Bernoulli com parâmetro p.
- Considere que Y tem uma distribuição binomial com parâmetros n e p. Isto modela o número de sucessos no problema análogo de amostragem com reposição. Se N e K forem grandes comparados a n e p não for próximo de 0 ou 1, então X e Y tem distribuições semelhantes, isto é, $P(X \leq k) \approx P(Y \leq k)$.
- Se n for grande, N e K forem grandes comparados a n e p não for próximo de 0 ou 1, então,

  $$P(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right),$$

  em que Φ é função distribuição normal padrão.

- Se as probabilidades de retirar uma bola branca ou preta não forem iguais (por exemplo, porque bolas brancas são maiores ou mais fáceis de pegar do que as bolas pretas), então, X tem uma distribuição hipergeométrica não central.
- A distribuição beta-binomial é *a priori* conjugada para a distribuição hipergeométrica.

A tabela abaixo descreve quatro distribuição relacionadas com o número de sucessos em uma sequência de retiradas:

	Com reposições	**Sem reposições**
Dado número de retiradas	Distribuição binomial	Distribuição hipergeométrica
Dado número de fracassos	Distribuição binomial negativa	Distribuição hipergeométrica negativa

Limites de cauda

Considere $X \sim$ Hipergeométrica (K, N, n) e $p = K/N$. Então, podemos derivar os seguintes limites:[6]

$$\begin{aligned}
\Pr[X \leq (p-t)n] &\leq e^{-n\mathrm{D}(p-t||p)} \leq e^{(-2t^2n)} \\
\Pr[X \geq (p+t)n] &\leq e^{-n\mathrm{D}(p+t||p)} \leq e^{(-2t^2n)}
\end{aligned}$$

em que

$$D(a||b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$$

é a divergência de Kullback-Leibler e $D(a,b) \geq 2(a-b)^2$ é usado.[7]

Se n for maior que $N/2$, pode ser útil aplicar simetria para "inverter" os limites, o que resulta no seguinte:[7][8]

$$\Pr[X \leq (p-t)n] \leq e^{-(N-n)\mathrm{D}(p+\frac{tn}{N-n}||p)} \leq e^{-2t^2n\frac{n}{N-n}},$$

$$\Pr[X \geq (p+t)n] \leq e^{-(N-n)\mathrm{D}(p-\frac{tn}{N-n}||p)} \leq e^{-2t^2n\frac{n}{N-n}}.$$

Distribuição hipergeométrica multivariada

O modelo de uma urna com bolas pretas e brancas pode ser estendida ao caso em que há mais de duas cores de bolas. Se houver K_i bolas de cor i na urna e forem retiradas n bolas aleatoriamente, sem reposição, então, o número de bolas de cada cor na amostra $(k_1, k_2, \ldots, k_c)$ tem distribuição hipergeométrica multivariada. Esta tem uma relação com a distribuição multinomial igual à que a distribuição hipergeométrica tem com a distribuição binomial — a distribuição multinomial é a distribuição "com reposição" e a a distribuição hipergeométrica multivariada é a distribuição "sem reposição".

As propriedades desta distribuição são dadas na tabela adjacente, em que c é o número de cores diferentes e $N = \sum_{i=1}^{c} K_i$ é o número total de bolas.

Distribuição hipergeométrica multivariada

Parâmetros	$c \in \mathbb{N} = \{0, 1, \ldots\}$ $(K_1, \ldots, K_c) \in \mathbb{N}^c$ $N = \sum_{i=1}^{c} K_i$ $n \in \{0, \ldots, N\}$
Suporte	$\left\{ \mathbf{k} \in \mathbb{Z}_{0+}^c \,:\, \forall i\, k_i \leq K_i, \sum_{i=1}^{c} k_i = n \right\}$
f.d.p.	$\frac{\prod_{i=1}^{c} \binom{K_i}{k_i}}{\binom{N}{n}}$
Média	$E(X_i) = \frac{nK_i}{N}$
Variância	$\mathrm{Var}(X_i) = \frac{K_i}{N}\left(1 - \frac{K_i}{N}\right) n\frac{N-n}{N-1}$

Exemplo

Suponha que uma urna contém cinco bolas pretas, dez bolas brancas e quinze bolas vermelhas. São selecionadas seis bolas sem reposição. A probabilidade de que sejam retiradas duas bolas de cada cor é

$$P(\text{2 pretas, 2 brancas, 2 vermelhas}) = \frac{\binom{5}{2}\binom{10}{2}\binom{15}{2}}{\binom{30}{6}} = 0.079575596816976.$$

Quando são retiradas seis bolas sem reposição, o número esperado de bolas pretas é $6 \times (5/30) = 1$, o número esperado de bolas brancas é $6 \times (10/30) = 2$ e o número esperado de bolas vermelhas é $6 \times (15/30) = 3$. Isto vem do valor esperado de uma distribuição binomial $E(X) = np$.

Ver também

- Amostragem (estatística)
- Distribuição geométrica
- Distribuição multinomial
- Keno

Referências

1. Rice, John A. (2007). *Mathematical Statistics and Data Analysis* (https://books.google.com.br/books?id=b6XHAAAACAAJ&dq=Mathematical+Statistics+and+Data+Analysis&hl=pt-BR&sa=X&redir_esc=y) (em inglês). [S.l.]: Thompson/Brooks/Cole. ISBN 9780495118688
2. Berkopec, Aleš (1 de junho de 2007). «HyperQuick algorithm for discrete hypergeometric distribution» (http://www.sciencedirect.com/science/article/pii/S1570866706000499). *Journal of Discrete Algorithms*. **5** (2): 341–347. doi:10.1016/j.jda.2006.01.001 (https://dx.doi.org/10.1016%2Fj.jda.2006.01.001)
3. Rivals, Isabelle; Personnaz, Léon; Taing, Lieng; Potier, Marie-Claude (15 de fevereiro de 2007). «Enrichment or depletion of a GO category within a class of genes: which test?» (https://academic.oup.com/bioinformatics/article/23/4/401/181853/Enrichment-or-depletion-of-a-GO-category-within-a). *Bioinformatics*. **23** (4): 401–407. ISSN 1367-4803 (https://www.worldcat.org/issn/1367-4803). doi:10.1093/bioinformatics/btl633 (https://dx.doi.org/10.1093%2Fbioinformatics%2Fbtl633)
4. «Interactive Fisher's Exact Test» (http://quantpsy.org/fisher/fisher.htm). *quantpsy.org*. Consultado em 18 de julho de 2017
5. Pollard, David (19 de setembro de 2005). «Chapter 4-Symmetry» (http://www.stat.yale.edu/~pollard/Courses/600.spring2010/Handouts/Symmetry%5BPolyaUrn%5D.pdf) (PDF). Universidade Yale. Consultado em 18 de julho de 2017
6. Hoeffding, Wassily (1 de março de 1963). «Probability Inequalities for Sums of Bounded Random Variables» (http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830). *Journal of the American Statistical Association*. **58** (301): 13–30. ISSN 0162-1459 (https://www.worldcat.org/issn/0162-1459). doi:10.2307/2282952 (https://dx.doi.org/10.2307%2F2282952)
7. Ahle, Thomas Dybdahl (8 de dezembro de 2015). «Another Tail of the Hypergeometric Distribution» (https://ahlenotes.wordpress.com/2015/12/08/hypergeometric_tail/). *ahlenotes*. Consultado em 19 de julho de 2017
8. Serfling, R. J. (janeiro de 1974). «Probability Inequalities for the Sum in Sampling without Replacement» (http://projecteuclid.org/euclid.aos/1176342611). *The Annals of Statistics* (em inglês). **2** (1): 39–48. ISSN 0090-5364 (https://www.worldcat.org/issn/0090-5364). doi:10.1214/aos/1176342611 (https://dx.doi.org/10.1214%2Faos%2F1176342611)

Ligações externas

- Distribuição hipergeométrica em *Wolfram MathWorld* (http://mathworld.wolfram.com/HypergeometricDistribution.html) (em inglês)
- Calculadora *on-line* de distribuição hipergeométrica (http://www.elektro-energetika.cz/calculations/distrhypgeo.php?language=portugues) (em português)
