# Introdução à Distribuição Hipergeométrica

Instituto de Economia - UFRJ

29 de junho de 2023

# Distribuição Hipergeométrica

A distribuição hipergeométrica trata de experimentos sem reposição.

Ela é relevante quando temos uma amostra (ou população) de tamanho finito, e esse limite é relevante para o cálculo de probabilidades.

Vamos responder à seguinte pergunta: qual é a probabilidade de obter $X = k$ sucessos em $n$ tentativas, em uma população de tamanho $N$ com $K$ sucessos possíveis?

A distribuição hipergeométrica é semelhante à binomial, mas elas diferem em uma característica crucial: a independência dos eventos. **A probabilidade de sucesso, na hipergeométrica, não é dada, mas depende dos resultados anteriores.**

# Distribuições Hipergeométrica e Binomial

A distribuição binomial descreve experimentos em que cada tentativa é independente das demais. Exemplo clássico: lançamento de uma moeda.

A distribuição hipergeométrica lida com experimentos sem reposição. Ou seja, a probabilidade de cada evento muda com cada tentativa. Exemplo clássico: retirada de cartas de um baralho. A probabilidade de tirar uma determinada carta muda cada vez que uma carta é retirada, porque essa carta não é devolvida ao baralho.

Se $N$ e $K$ são grandes em relação a $n$, e $p$ não é próximo de 0 ou 1, as duas distribuições são similares.

# Função de Probabilidade da Distribuição Hipergeométrica

A função de probabilidade de uma variável aleatória X que segue uma distribuição hipergeométrica é dada por:

$$P(X = k) = \frac{C(K, k) \cdot C(N - K, n - k)}{C(N, n)} \tag{1}$$

onde:

- $N$ é o tamanho total da população.
- $K$ é o número de sucessos na população.
- $n$ é o número de observações (ou seja, o número de tentativas).
- $k$ é o número de sucessos que resultaram das observações.
- $C(a, b)$ é a combinação de $a$ itens tomados $b$ a $b$ (ou seja, o número de maneiras diferentes de escolher $b$ itens de um conjunto de $a$ itens).

# Obtendo a Função de Probabilidade da Distribuição Hipergeométrica (1)

Para obter a função de probabilidade da distribuição hipergeométrica, precisamos observar que estamos lidando com uma situação de amostragem sem reposição de uma população finita.

Vamos supor que temos uma população de tamanho $N$ com $K$ sucessos. Estamos interessados em calcular a probabilidade de obter exatamente $k$ sucessos em $n$ tentativas.

O número de maneiras de obter $k$ sucessos a partir de $K$ possíveis é dado por $C(K, k)$.

De forma similar, o número de maneiras de obter $n - k$ fracassos da população restante de $N - K$ é dado por $C(N - K, n - k)$.

# Obtendo a Função de Probabilidade da Distribuição Hipergeométrica (2)

Então, o produto desses dois termos nos dá todas as combinações possíveis de obter $k$ sucessos e $n - k$ fracassos.

Por fim, dividimos pelo número total de maneiras de selecionar $n$ itens da população de $N$, que é dado por $C(N, n)$, para obter a probabilidade desejada.

Isso resulta na função de probabilidade da distribuição hipergeométrica.

# Esperança

A **esperança** (ou valor esperado) de uma variável aleatória X que segue uma distribuição hipergeométrica é dada por:

$$E[X] = n \cdot \frac{K}{N} \tag{2}$$

Isto é, o número de tentativas vezes a probabilidade de sucesso em uma única tentativa.

# Variância

A **variância** de uma variável aleatória X que segue uma distribuição hipergeométrica é dada por:

$$Var[X] = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N-n}{N-1} \tag{3}$$

Esta fórmula captura a ideia de que a variância (uma medida da dispersão dos dados) aumenta com o número de tentativas e a probabilidade de sucesso, mas diminui se a amostra é uma fração grande da população total.

# Comparação com esperança da Binomial

É útil comparar a esperança e a variância das distribuições binomial e hipergeométrica para entender suas semelhanças e diferenças.

**Esperança**:

- Binomial: $E[X] = n \cdot p$
- Hipergeométrica: $E[X] = n \cdot \frac{K}{N}$

Ambas as esperanças são produtos do número de tentativas e a probabilidade de sucesso. A diferença reside na probabilidade de sucesso: na binomial, é uma constante $p$, enquanto na hipergeométrica, é $K/N$, que muda conforme os sucessos são retirados da população.

# Comparação com variância da Binomial

**Variância**:

- Binomial: $Var[X] = n \cdot p \cdot (1 - p)$
- Hipergeométrica: $Var[X] = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N-n}{N-1}$

Na variância, notamos a adição de um termo $\frac{N-n}{N-1}$ na hipergeométrica. Este termo é o fator de correção de finitez, que reduz a variância quando a amostra é uma grande fração da população total.

# Exemplo de Distribuição Hipergeométrica

Vamos considerar um exemplo para ilustrar a distribuição hipergeométrica.

Suponha que temos uma urna com 20 bolas, 8 das quais são vermelhas e as restantes 12 são azuis. Vamos retirar 5 bolas da urna sem reposição. Estamos interessados em calcular a probabilidade de obter exatamente 2 bolas vermelhas.

Neste cenário, temos:

- $N$ (tamanho total da população) $= 20$
- $K$ (número total de sucessos na população) $= 8$
- $n$ (número de tentativas ou retiradas) $= 5$
- $k$ (número de sucessos que estamos interessados) $= 2$

Usando a função de probabilidade da distribuição hipergeométrica:

$$P(X = 2) = \frac{C(8, 2) \cdot C(12, 3)}{C(20, 5)} \tag{4}$$

# Esperança e Variância

No exemplo anterior:

$$E[X] = n \cdot \frac{K}{N} = 5 \cdot \frac{8}{20} = 2 \tag{5}$$

$$Var[X] = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N-n}{N-1} = 5 \cdot \frac{8}{20} \cdot \left(1 - \frac{8}{20}\right) \cdot \frac{20-5}{20-1} \approx 0.84 \tag{6}$$

O desvio padrão, que neste caso seria aproximadamente 0.92, indicando que a maioria das retiradas teria um número de bolas vermelhas dentro de 0.92 da média.