Hypergeometric Distribution

Egor Howell

Intuition

The hypergeometric distribution measures the probability of ***k*** success in ***n*** number of trials (a sample), *without replacement*, given some information about the population.

This is very similar to the ***binomial distribution*** bar the one key difference of sampling *without replacement*. Due to this, the probability of each success changes every trial, whereas in the binomial distribution the probability of a success (and failure) is fixed.

Intuition

An easy-to-understand example is determining the probability of drawing all ***4*** kings in a random sample of ***20*** cards from a standard deck of cards.

If we draw a king, the probability of drawing the subsequent king will be different from the first as the population composition has changed. Thus, the probability of a success is dynamic.

Theory

Probability Mass Function

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

Equation in LaTeX by author.

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

Equation in LaTeX by author.

- ***n*** is the number of trials
- ***k*** is the number of successes
- ***N*** is the population size
- ***K*** is the total number of successes in the population
- ***X*** is a random variable from the hypergeometric distribution

The interested reader can find a derivation of the PMF here.

Example

Let’s go back to our previous example of drawing 4 kings in a random 20 card sample from a regular deck of cards. The information we have is:

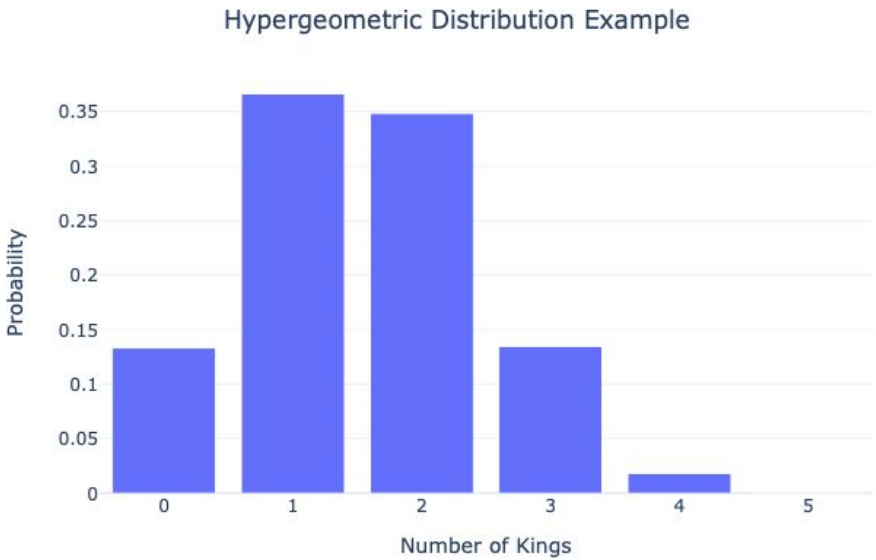
- ***N* = *52***: Number of cards in the deck
- ***n* = *20***: Number of cards we sample
- ***k* = *4***: Number of kings we want (successes)
- ***K* = *4***: Number of kings in the deck

$$P(X = 4) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{4}{4}\binom{52-4}{20-4}}{\binom{52}{20}} \approx 0.0179$$

Equation in LaTeX by author.

Distribution

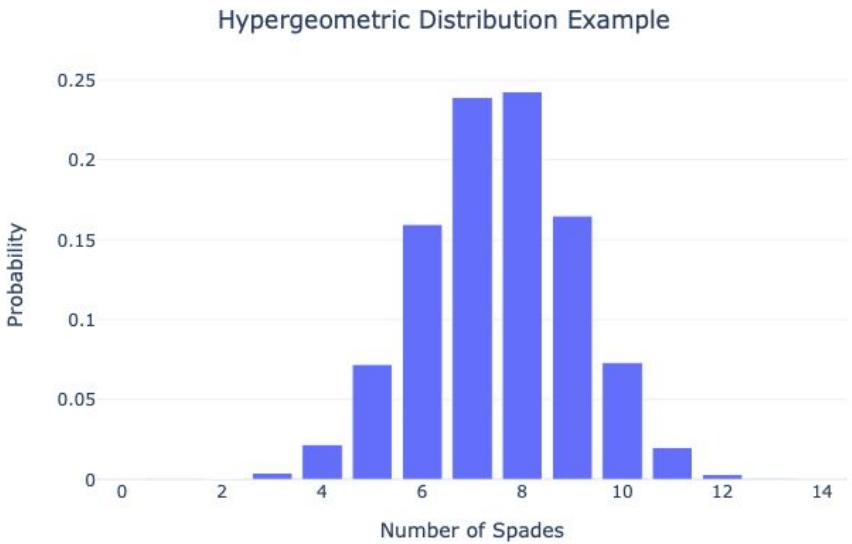
Drawing 4 kings in a random 20 card sample from a regular deck of cards




```python
# Import packages
import plotly.graph_objects as go
from scipy.stats import hypergeom

# Define the hypergeometric distribution parameters
N = 52
K = 4
n = 20

# Number of possible successes
k = list(range(0, K+1))

# The PMF
pmf = hypergeom.pmf(k, N, K, n)

# Plot the PMF
fig = go.Figure(data=[go.Bar(x=k, y=pmf, marker_color='rgba(176, 224, 230)')])

fig.update_layout(
    title="Hypergeometric Distribution Example",
    xaxis_title="Number of Kings",
    yaxis_title="Probability",
    width=650,
    height=400,
    template="plotly_white",
    margin=dict(l=80, r=40, t=60, b=40),
    title_x=0.5
)

fig.update_xaxes(range=[-0.5, K+1])
fig.show()
```

Another Example

Number of spades-suited cards from a random 30-card sample?




```python
# Define the hypergeometric distribution parameters
N = 52
K = 13
n = 30

# Number of possible successes
k = list(range(0, K+1))

# The PMF
pmf = hypergeom.pmf(k, N, K, n)

# Plot the PMF
fig = go.Figure(data=[go.Bar(x=k, y=pmf, marker_color='rgba(176, 224, 230)')])
fig.update_layout(
    title="Hypergeometric Distribution Example",
    xaxis_title="Number of Spades",
    yaxis_title="Probability",
    width=650,
    height=400,
    template="plotly_white",
    margin=dict(l=80, r=40, t=60, b=40),
    title_x=0.5
)
fig.update_xaxes(range=[-0.5, K + 1.5])
fig.show()
```

Applications

The hypergeometric distribution touches many fields including:

- *Probability of winning a hand in poker*
- *Voting populations analysis*
- *Quality control in manufacturing*
- *Genetic variations within a population*

Therefore, the hypergeometric is something you will most likely come across in your data science career and thus is worth knowing about.

Thanks

Understanding The Hypergeometric Distribution

Breaking down one of the less well-known distributions in data science

Egor Howell

Published in Towards Data Science · 5 min read · Jun 21

127

Photo by Roth Melinda on Unsplash

@egorhowell

Newsletter

Dishing The Data