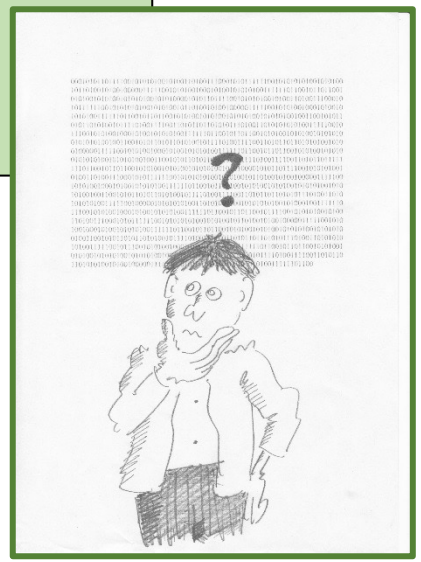
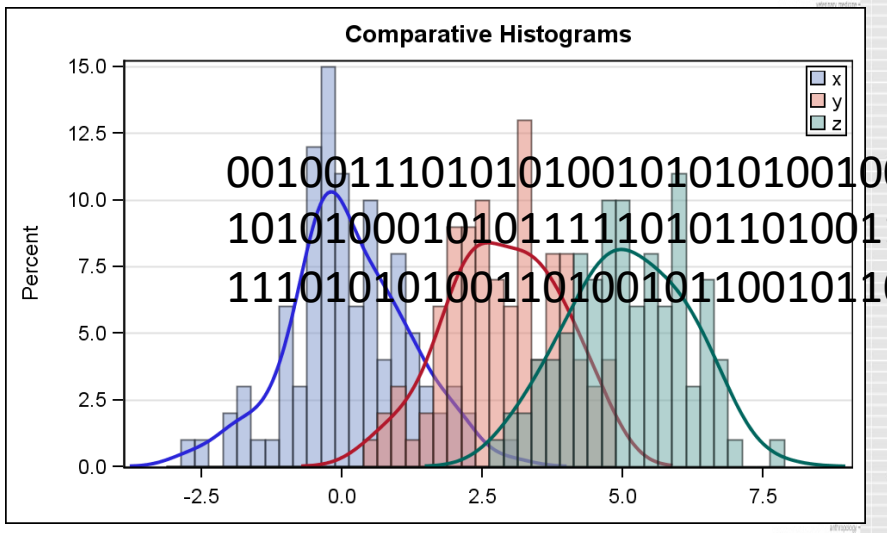
Hypothesis testing and the hypergeometric distribution

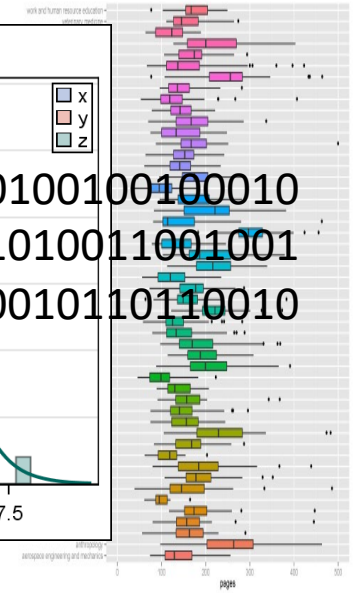
Statistics and data analysis

Zohar Yakhini, Leon Anavy

IDC, Herzeliya







Using the binomial distribution.
Example: Experimental treatment for Kidney Cancer

- Suppose we have $n = 40$ patients who will be receiving an experimental therapy (Tx) which is believed to be better than current treatments (standard of care = SoC).
  The latter has a historically derived 5-year survival rate of 20%. That is, under the SoC the probability of 5-year survival is $p = 0.2$
- We will now count 5-year survival under Tx and will then need to decide if we can confidently say that the new experimental treatment is better.

Results and “The Question”

- Suppose that using the new treatment we find that 16 out of the 40 patients survive at least 5 years past diagnosis.
- Q: Does this result suggest that the new therapy, Tx, has a better 5-year survival rate than that of the SoC?
  That is:
  is the probability that a patient survives at least 5 years greater than 0.2 when treated using the new therapy?

What do we consider in answering the question of interest?

We essentially ask ourselves the following:

- If we assume that new therapy is **no better** than the current then what is the probability of seeing the observed numbers? That is – how likely are they to occur, in such case, by chance alone?
- More specifically:
  What is the probability of seeing 16 or more successes out of 40 if the success rate of the new therapy is also 0.2?
- This is called estimating the **p-value** of the **OBSERVED RESULT** under the **NULL model**

Binomial null model

- This is a binomial experiment situation.
  - There are n = 40 patients and we are counting the number of patients that survive 5 or more years.
  - The individual patient outcomes are independent and under the NULL MODEL the probability of success is *p = 0.2* for all patients. (that is: we assume that Tx is NOT better than the SoC)

- So the random variable *X = # of "successes" in the clinical trial* is, under the NULL model, Binomial with *n = 40* and *p = 0.2:*

$$\underline{X \sim BIN(40,0.2)}$$

Binomial null model

So, the p-value will be calculated as

For $X \sim Binomial(n = 40, p = 0.2)$

$$p - value = P(X \geq 16) = 1 - CDF_X(15) = 0.0029$$

```
from scipy.stats import binom
rv = binom(40, 0.2)
x_16_and_up = 1 - rv.cdf(15)
print("{:.4f}".format(x_16_and_up))
```

```
0.0029
```

Conclusion (statistics helps decision making …)

Because it is highly unlikely (p = 0.0029) that we would see this many successes in a group of *40* patients if the new Tx had the same probability of success as the SoC we have to make a choice, either …

A) Tx's survival rate is less than 0.2 and we have obtained a very rare result by chance.

OR

B) our assumption about the success rate of the new Tx is wrong and in actuality it has a better than 20% 5-year survival rate making the observed result more plausible.

Caveat: other aspects of the null model can also be wrong …

Tx is better than the SoC treatment with p-value <0.003 under a binomial null model

Alternative notation

- **Null hypothesis H_0** - the conservative hypothesis on which we want to defend (Tx is no better than SoC → $p \leq 0.2$)
- **Alternative hypothesis H_1** - a new hypothesis that we want to check (Tx is better than SoC → $p > 0.2$)

- We assume H_0 is true until we decide otherwise!
- We can only reject H_0
- We can not verify an hypothesis, only fail to reject it

Alternative notation

- **Test statistic**
  - Can be calculated from the sample

    $\mathrm{X} = (\#"successes")$
  - We know its distribution under H_0 (if H_0 is true than the distribution is known)

    $\mathrm{X} \sim Binomial(40, 0.2)$ (under H_0, if H_0 is true and Tx is no better than SoC)

- **p-value**
  - Under $\boldsymbol{H_0}$, What is the probability to get a test statistic which is equal or "more extreme" than the observed.

    $P(S \geq 16)$ when $S \sim Bin(40, 0.2)$
  - If the **probability is low** than $\boldsymbol{H_0}$ **is rejected**
  - If the **probability is high** than $\boldsymbol{H_0}$ **cannot be rejected**

Tx is better than the SoC treatment with p-value <0.003 under a binomial null model

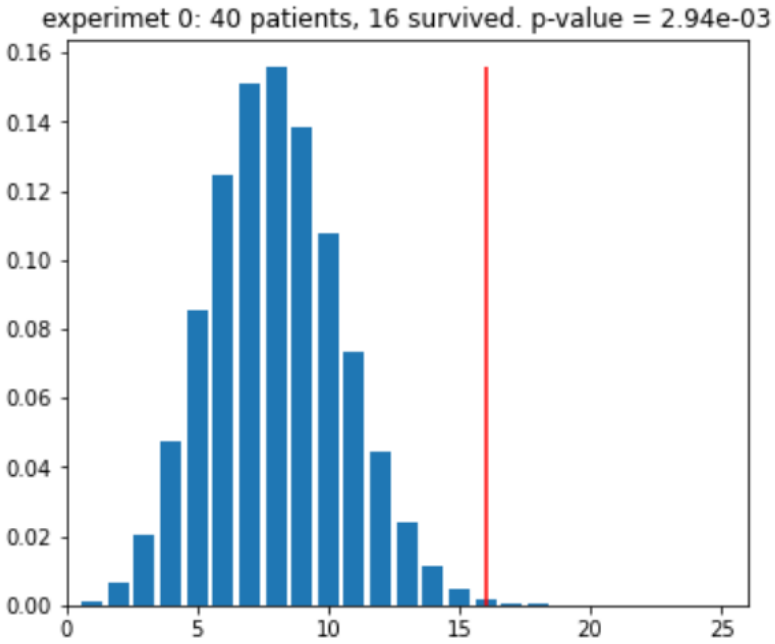
The observed 5 yrs survival in Tx has a Right side p-value <0.003 under a binomial null model that represents the SoC (Binom(40,0.2))

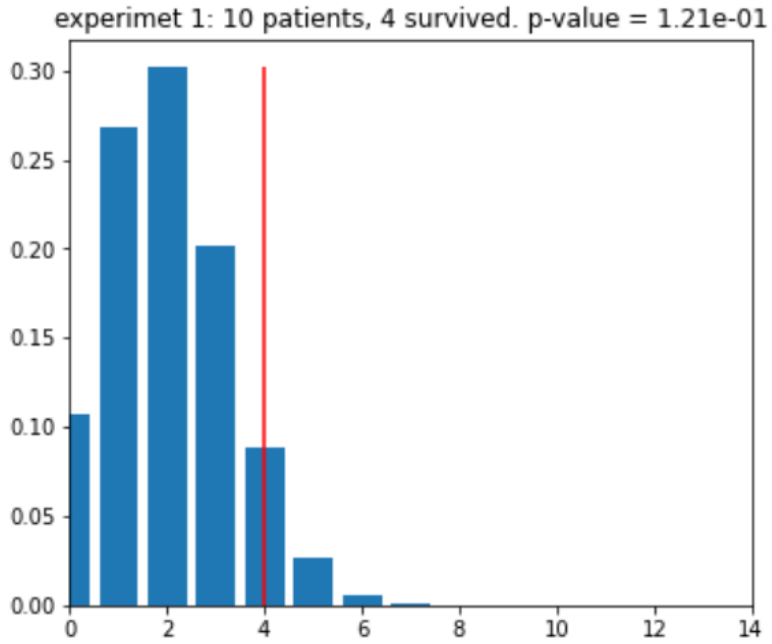
Sample size matters

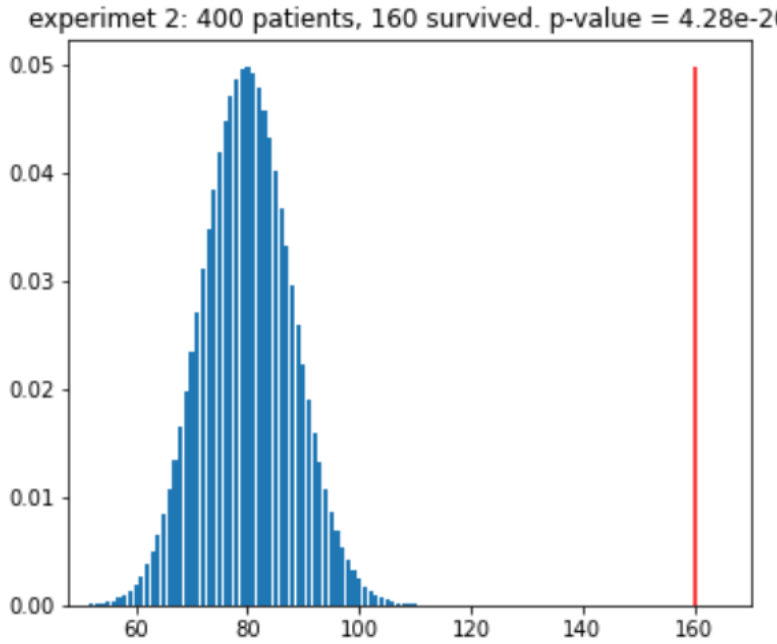
Experiment 1: 40 patients, 16 survived after 5-years

Experiment 2: 10 patients, 4 survived after 5-years

Experiment 2: 400 patients, 160 survived after 5-years







Approximating the binomial distribution

The central limit theorem:

Let $X_1, X_2, X_3, \ldots, X_n$ be random variables all sampled **independently** from **the same distribution** with mean μ and (finite non 0) variance σ^2.

Let $\overline{X_n}$ be the average of $X_1, X_2, X_3, \ldots, X_n$.

Then for any fixed number x we have

$$\lim_{n\to\infty} P\left(\frac{\sqrt{n}}{\sigma}\ (\overline{X_n} - \mu)\ \leq x\right) = \ \Phi(x)$$

where $\Phi(x)$ is the standard normal density function.

Approximating the binomial distribution

For (**almost**) any distribution, if we sample it (**sufficiently**) many times, and then average, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed with mean μ and variance σ^2/n.

$$\bar{X}_n \dot{\sim} N(\mu, \sigma^2/n)$$

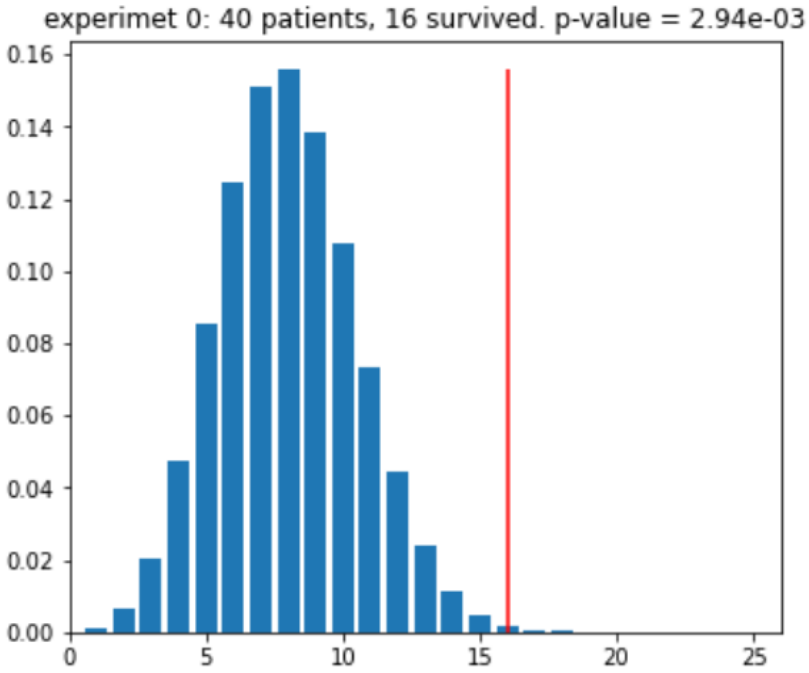
Or alternatively: the sum $S_n = \sum_{i=1}^{n} X_i$ is approximately normally distributed with mean $n\mu$ and variance $n\sigma^2$.

$$S_n \dot{\sim} N(n\mu, n\sigma^2)$$

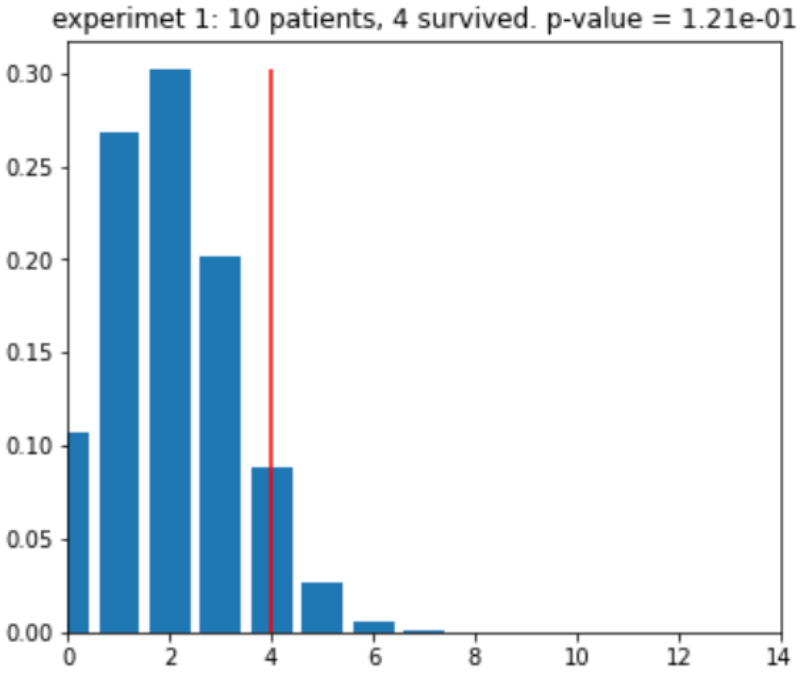
Approximating the binomial distribution

$$X \sim Binomial(n,p) \rightarrow X \dot{\sim} N\left(np, np(1-p)\right)$$

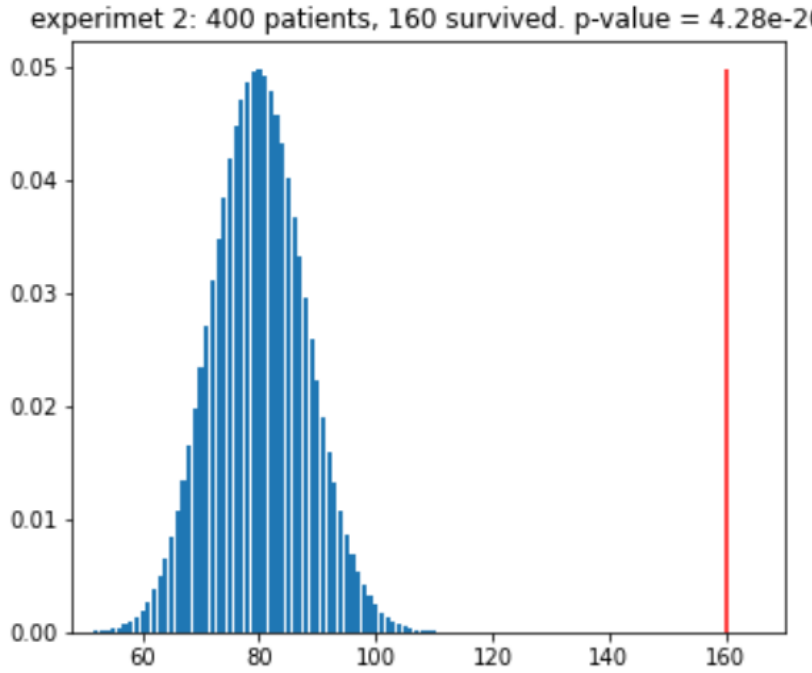
$X \dot{\sim} N(8,6.4)$
$P(X \geq 16) \cong$
$P(Z > 2.96) = 0.0015$



$X \dot{\sim} N(2,1.6)$
$P(X \geq 4) \cong$
$P(Z > 1.19) = 0.12$



$X \dot{\sim} N(80,64)$
$P(X \geq 160) \cong$
$P(Z > 9.94) < 10^{-22}$

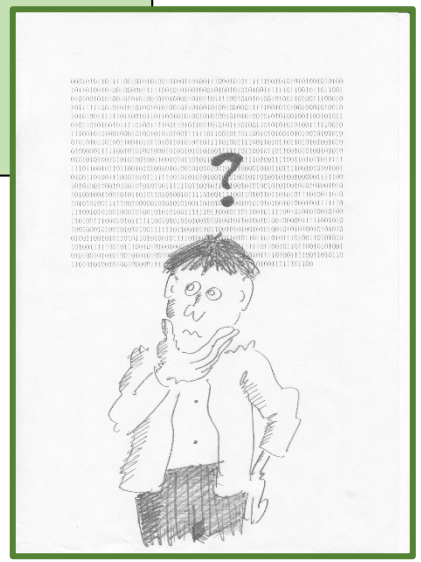


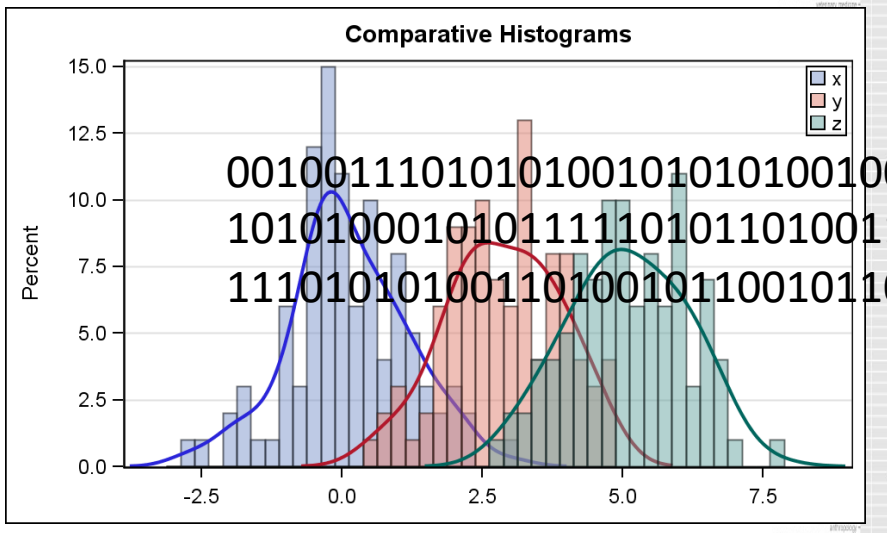
The hypergeometric distribution and COVID19

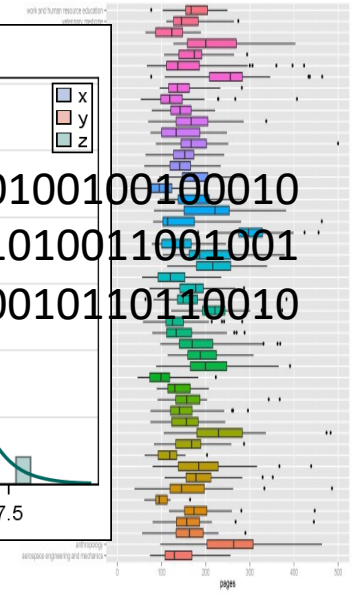
Statistics and data analysis

Zohar Yakhini, Leon Anavy

IDC, Herzeliya





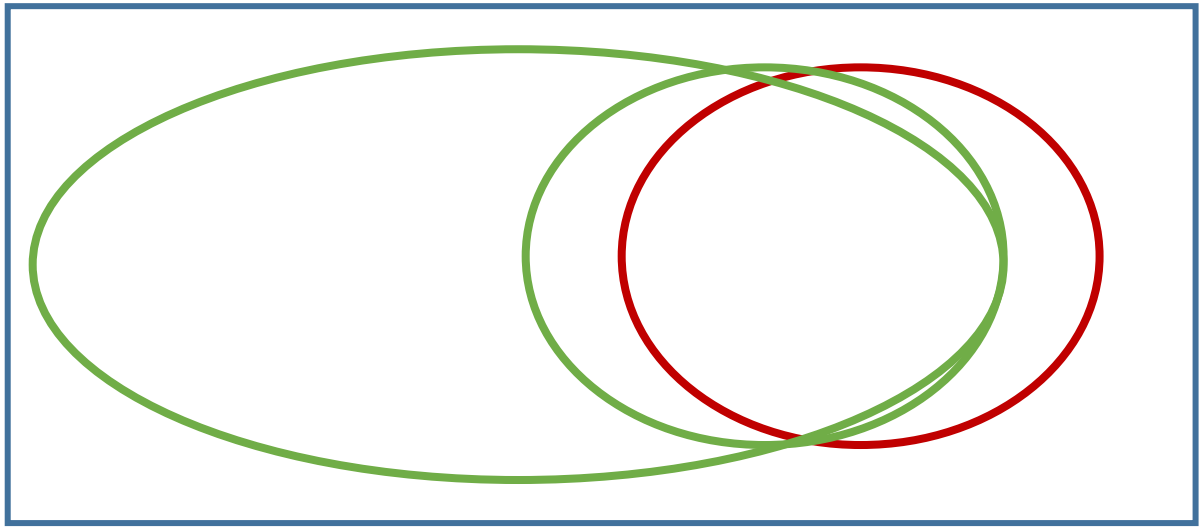


Pfizer/BioNTech announcement

A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo; BNT162b2 was 95% effective in preventing Covid-19 (95% credible interval, 90.3 to 97.6). Similar vaccine efficacy (generally 90 to 100%) was observed across subgroups defined by age, sex, race, ethnicity, baseline body-mass index, and the presence of coexisting conditions. Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient. The safety profile of BNT162b2 was characterized by short-term, mild-to-moderate pain at the injection site, fatigue, and headache. The incidence of serious adverse events was low and was similar in the vaccine and placebo groups.

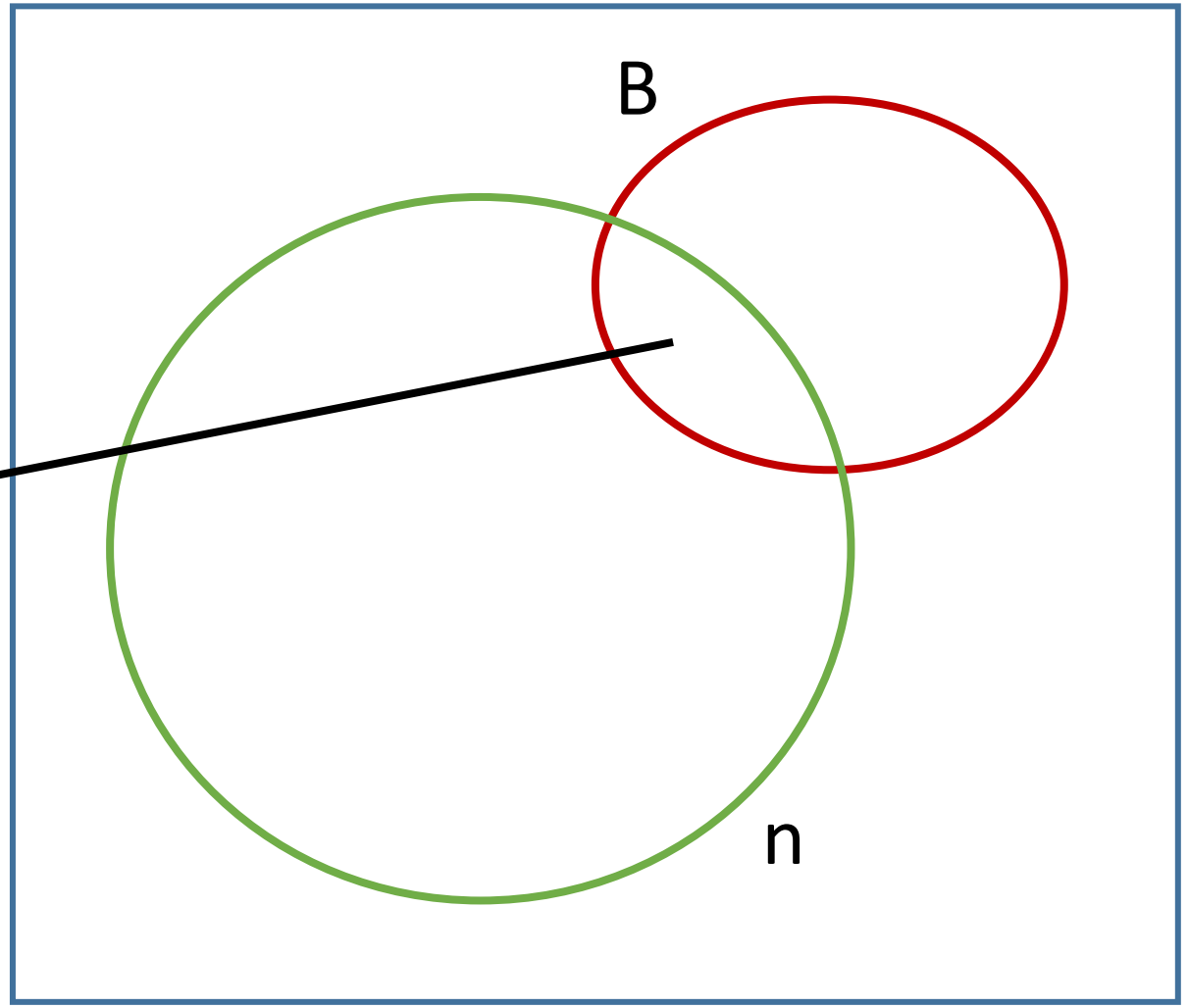
The Hyper-Geometric Distribution (HG)

Of 40 students in class, 10 received a grade of A in the exam. 8 of them have a last name starting with a letter in the range A-M.
Can we say something about the grade being related to the first letter of the last name?



The Hyper-Geometric Distribution (HG)

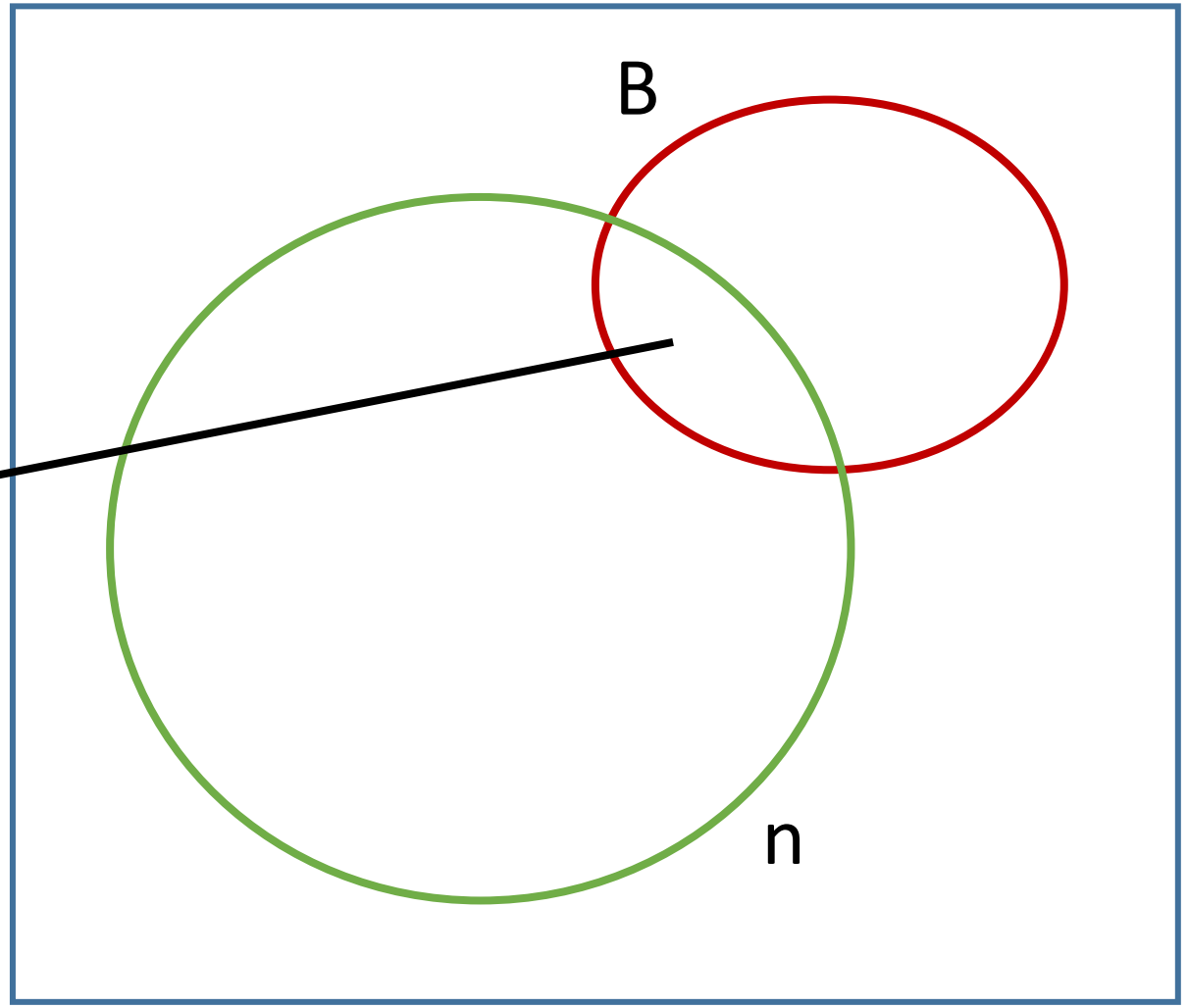
Size of the intersection –
what is the distribution??



The Hyper-Geometric Distribution (HG)

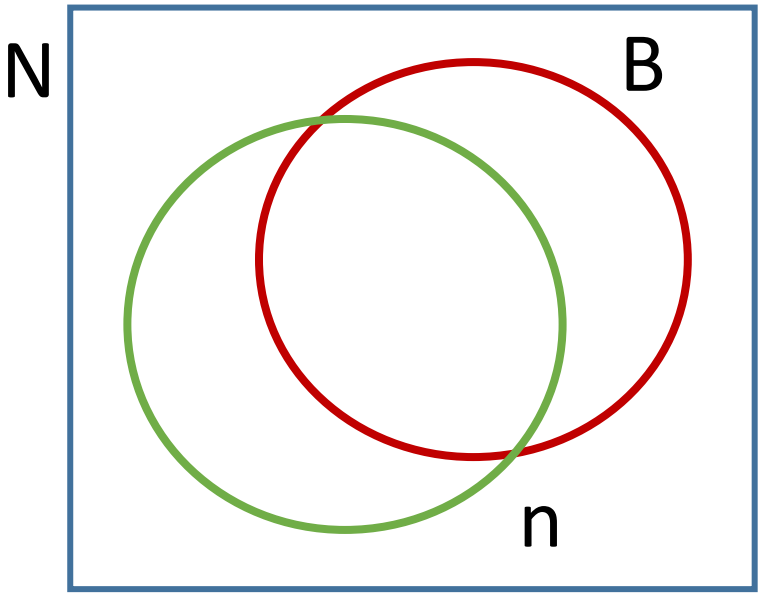
Assuming that everything is uniform then the probability of exactly b elements in the intersection is:

$$HG(N, B, n, b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$



Example – effectiveness of a cold treatment

Cold Length	Medicine Taken: yes	no	Total
1 -3 days	86	19	105
4 - 7 days	16	79	95
Total	102	98	200



N = 200

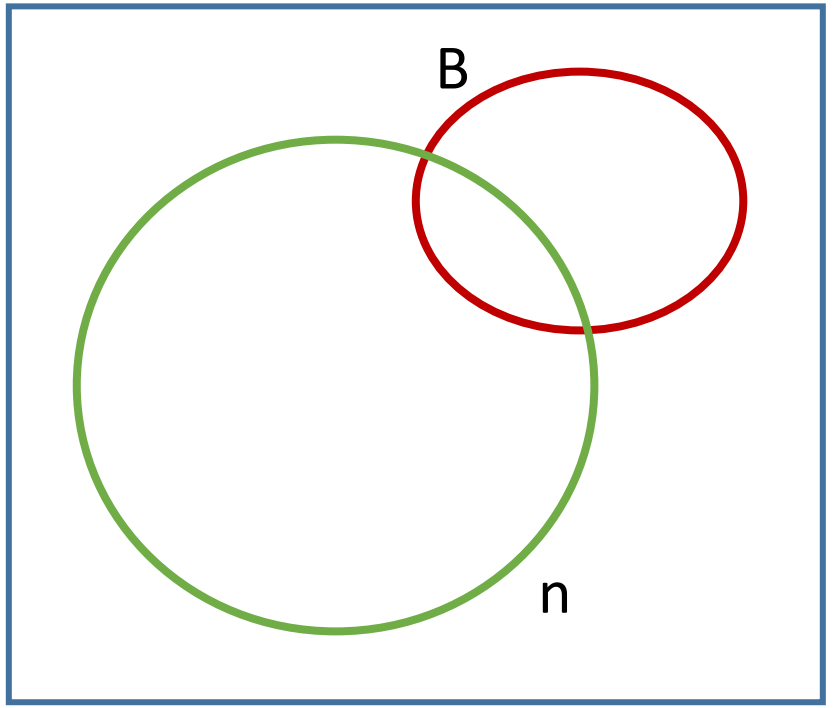
B = 105 people whose cold lasted <3 days

n = 102 people who took the Tx

b = 86, the ∩

Hyper-Geometric hypothesis test

- **Null hypothesis H_0 -** fix N and B and assume that all sets of size n are equiprobable



- **Test statistic $X \sim HG(N, B, n)$**

$$P(X = b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

- **p-value:**

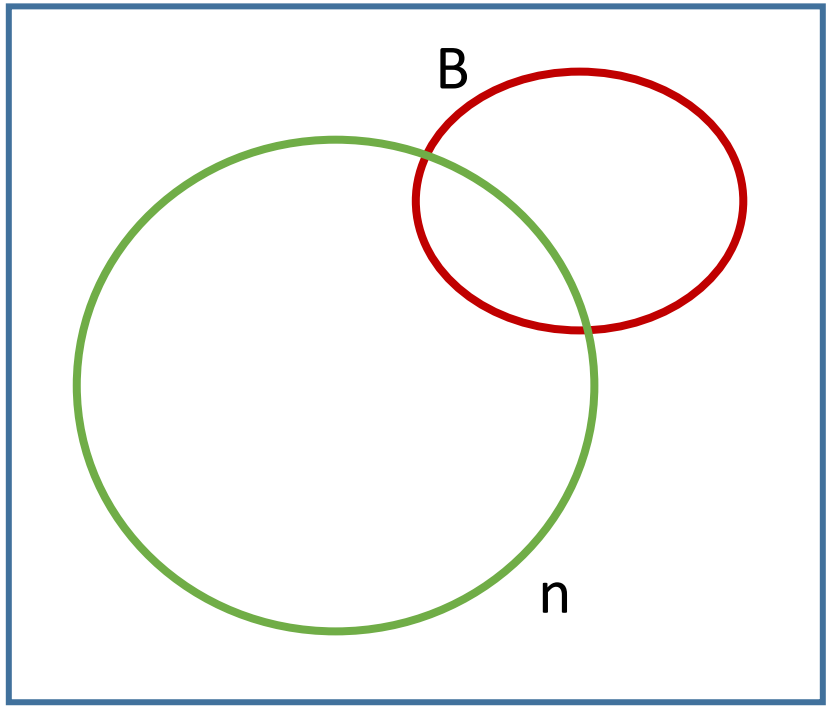
$$P(X \geq b) = HGT(N, B, n, b) = \sum_{t=b}^{min(n,B)} HG(N, B, n, t)$$

Hyper-Geometric hypothesis test

$$\boldsymbol{X \sim HG(200, 105, 102)}$$

$$\boldsymbol{P(X \geq 86)} = HGT(N, B, n, b) = \sum_{t=b}^{min(n,B)} HG(N, B, n, t)$$

$$= \mathbf{1} - \boldsymbol{CFD}(\mathrm{b} - 1)$$




```
from scipy.stats import hypergeom as hg
from math import comb
X = hg(M = 200, n = 102, N = 105) # Note: N->M, B->N
print(f'{1-X.cdf(85):.3e}')
```

```
8.327e-15
```

P-value < $\mathbf{10^{-14}}$ → Reject $\boldsymbol{H_0}$ → **We cannot say that there is no relationship between the medicine and the recovery**

Winning the Randomistan Lotto …

$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote 6 numbers on your lotto card ahead of time.
How many matches?

The probability of no matches:

Your numbers

$$\frac{\binom{6}{0}\binom{36}{6}}{\binom{42}{6}} \approx 0.37$$

Lotto numbers

$$HG(42,6,6,0) = \frac{\binom{6}{0}\binom{36}{6}}{\binom{42}{6}}$$

Winning the Lotto …

$$HG(N, B, n, b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote 6 numbers on your lotto card ahead of time.
How many matches?

The probability of 1 match:

Your numbers

$$\frac{\binom{6}{1}\binom{36}{5}}{\binom{42}{6}} \approx 0.43$$

Lotto numbers

$$HG(42,6,6,1) = \frac{\binom{6}{1}\binom{36}{5}}{\binom{42}{6}}$$

Winning the Lotto …

$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote 6 numbers on your lotto card ahead of time.
How many matches?

The probability of 5 matches:

Your numbers

$$\frac{\binom{6}{5}\binom{36}{1}}{\binom{42}{6}} \approx 0.00004$$

Lotto numbers

$$HG(42,6,6,1) = \frac{\binom{6}{5}\binom{36}{1}}{\binom{42}{6}}$$

Winning the Lotto …

$$HG(N, B, n, b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote 6 numbers on your lotto card ahead of time.
How many matches?

The probability of 6 matches:

Your numbers

$$\frac{\binom{6}{6}\binom{36}{0}}{\binom{42}{6}} \approx 0.0000002$$

Lotto numbers

$$HG(42,6,6,1) = \frac{\binom{6}{6}\binom{36}{0}}{\binom{42}{6}}$$

Winning the Lotto …

$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote 6 numbers on your lotto card ahead of time.
How many matches?
The probability of 6 matches:

```
from scipy.stats import hypergeom as hg
from math import comb
X = hg(M = 42, n = 6, N = 6) # Note: N->M, B->N
print(f'{1-X.cdf(5):.3e}')
print(f'{X.pmf(6):.3e}')
print(f'{1/comb(42,6):.3e}')
```

```
1.906e-07
1.906e-07
1.906e-07
```

Winning the Lotto …

$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote **7** numbers on your lotto card ahead of time.
How many matches?

The probability of 6 matches:

Your numbers

$$\frac{\binom{7}{6}\binom{35}{0}}{\binom{42}{6}} \approx 0.0000013$$

Lotto numbers

$$HG(42,6,6,1) = \frac{\binom{7}{6}\binom{35}{0}}{\binom{42}{6}}$$

Winning the Lotto …

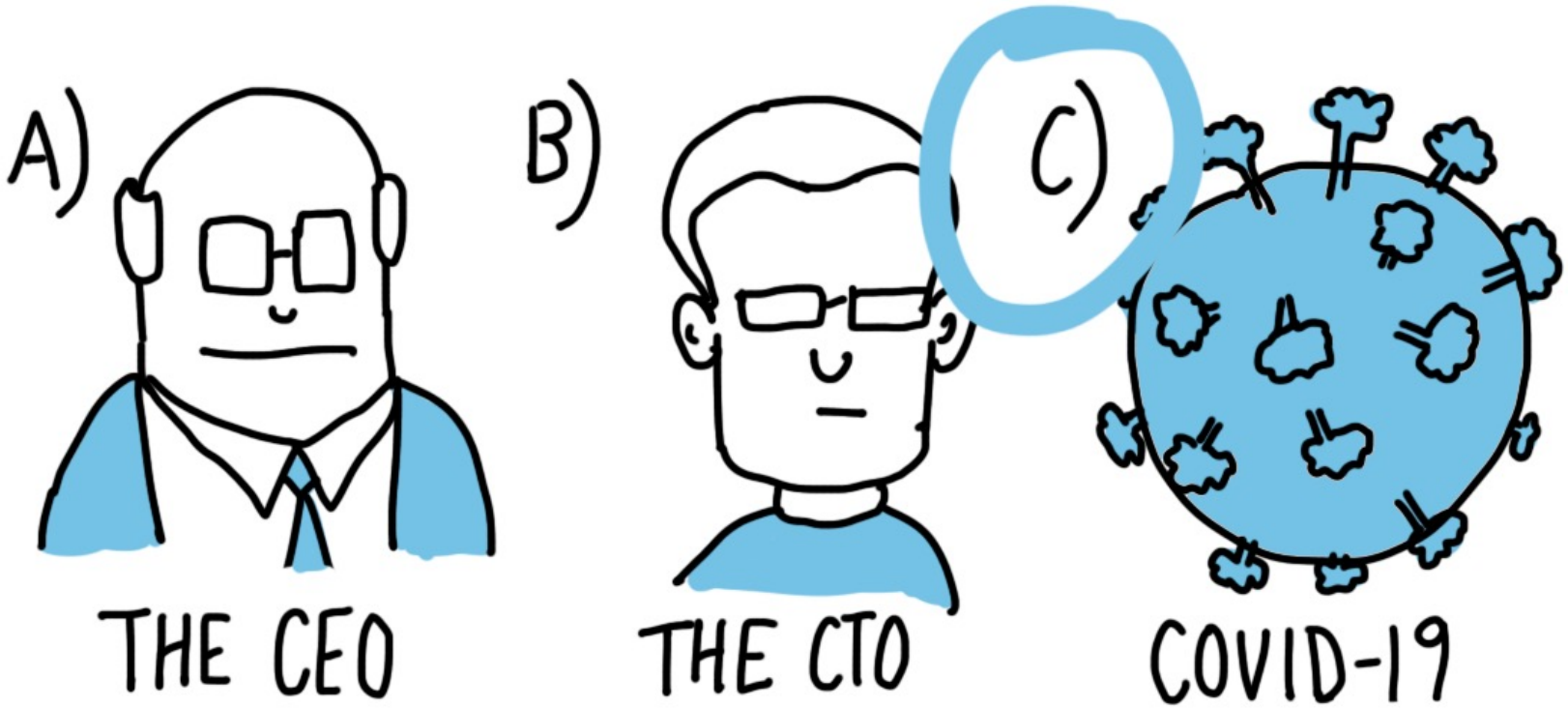
$$HG(N,B,n,b) = \frac{\binom{B}{b}\binom{N-B}{n-b}}{\binom{N}{n}}$$

6 of 42 numbered balls are drawn at random without replacement
You wrote **7** numbers on your lotto card ahead of time.
How many matches?
The probabilitv of 6 matches:

```
X = hg(M = 42, n = 6, N = 7) # Note: N->M, B->N
print(f'{1-X.cdf(5):.3e}')
print(f'{X.pmf(6):.3e}')
print(f'{comb(7,6)/comb(42,6):.3e}')
```

```
1.334e-06
1.334e-06
1.334e-06
```

WHO LED THE DIGITAL TRANSFORMATION OF YOUR COMPANY ?



BUSINESSILLUSTRATOR.COM

Pfizer/BioNTech announcement

A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo; BNT162b2 was 95% effective in preventing Covid-19 (95% credible interval, 90.3 to 97.6). Similar vaccine efficacy (generally 90 to 100%) was observed across subgroups defined by age, sex, race, ethnicity, baseline body-mass index, and the presence of coexisting conditions. Among 10 cases of severe Covid-19 with onset after the first dose, 9 occurred in placebo recipients and 1 in a BNT162b2 recipient. The safety profile of BNT162b2 was characterized by short-term, mild-to-moderate pain at the injection site, fatigue, and headache. The incidence of serious adverse events was low and was similar in the vaccine and placebo groups.

HG p-value for the BNT162b2 results

$$N = 43448, B = 21720, n = 162 + 8 = 170$$

$$X \sim HG(N, B, n) = HG(43448,21780,170)$$

$$X \cong Y \sim Binomial\left(n, p = {}^{B}/_{N}\right)$$

```
: from scipy.stats import binom, norm
  X = hg(M = 43448, n = 170, N = 21720) # Note: N->M, B->N
  print(f'{X.cdf(8):.3e}')
  Y = binom(n=170,p=21720/43448)
  print(f'{Y.cdf(8):.3e}')
```

```
8.056e-39
1.058e-38
```

Summary

- Hypothesis testing
- The hypergeometric distribution
- Normal and binomial approximations