Hypergeometrische verdeling (pag. 98-99)

- Beschouw een populatie van N elementen waarvan r aan een bepaalde eigenschap voldoen (dit noemen we succes of uitkomst 1) en de overige N – r niet aan de eigenschap voldoen (mislukking of uitkomst 0). Het aantal keer succes op een willekeurige steekproef van n verschillende elementen uit deze populatie is dan een hypergeometrisch verdeelde veranderlijke

- Parameters:
  - N = de grootte van de populatie
  - r = het aantal elementen uit de populatie die succes opleveren, die dus aan de eigenschap voldoen
  - n = de grootte van de steekproef die je neemt uit de populatie $n < N$

- Mogelijke waarden: $0, 1, 2, \cdots, n$

- Kansfunctie:

$$p_X(k) = P(X = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} = \frac{C_r^k C_{N-r}^{n-k}}{C_N^n}$$

voor elke k die voldoet aan: $\max(0, n-(N-r)) \le k \le \min(n, r)$

$p_X(k) = 0$ voor alle andere k.

Hoe komen we aan deze kansfunctie en aan die grenswaarden voor k?

De kansfunctie kunnen we afleiden via volgende schema:

$$\begin{array}{ccccc} N & = & r & + & N-r \\ \downarrow & & \downarrow & & \downarrow \\ n & = & k & + & n-k \end{array}$$

$$P(X = k) = \frac{\#\,gunstige\ gevallen}{\#\,mogelijke\ gevallen}$$

mogelijke gevallen: n elementen trekken uit een populatie van N, zonder herhaling en waarbij de volgorde niet belangrijk is: combinatie van n uit N.

$$\#\ mogelijke\ gevallen = C_N^n$$

#gunstige gevallen: een gunstig geval bekom je als er in de steekproef van n elementen, k komen uit het deel van de populatie die succes oplevert (bevat r elementen), en de overige n-k komen uit de resterende N-r elementen van de populatie.
Het aantal gunstige gevallen is dus het aantal mogelijkheden om k elementen te trekken uit r **en** n-k uit N-r, zonder terugleggen en waarbij de volgorde niet belangrijk is.
Dus: $\#\ mogelijke\ gevallen = C_r^k C_{N-r}^{n-k}$
Hieruit volgt onmiddellijk de kansfunctie.

Vanwaar komen de grenzen voor k?
De trekkingen die beschreven staan in bovenstaande afleiding hebben maar zin als de

volgende voorwaarden voldaan zijn:
$k \le r$

$n-k \le N-r \Leftrightarrow k \ge n-(N-r)$

En uiteraard voldoen de mogelijke waarden voor *k* (zie hoger) aan: $0 \le k \le n$

Samenvoegen geeft:

$k \le r$ en $k \le n$ dus $k \le \min(n,r)$

$k \ge n-(N-r)$ en $k \ge 0$ dus $k \ge \max(0, n-(N-r))$

In 1 keer geschreven: $\max(0, n-(N-r)) \le k \le \min(n,r)$

- Kenmerken (zonder bewijs):
  - Verwachtingswaarde: $E(X) = \frac{nr}{N}$
  - Variantie: $Var(X) = \frac{nr(N-r)(N-n)}{N^2(N-1)}$
- Voorbeelden:
  - Bij een tombola worden 200 loten verkocht, hierbij zijn 10 winnende loten. Een bepaalde persoon koopt 8 loten. Het aantal winnende loten uit de 8 gekochte loten is dan een hypergeometrische verdeling met parameters: $N = 200, r = 10, n = 8$
  - Extra uitleg bij voorbeeld 6.14 p 98.
    De bedoeling is van een schatting te maken van het aantal Belgen met een rekening in Nederland. Dit is de (onbekende) populatiegrootte N. Van deze populatie zijn er r = 3000 die dit aangeven aan de belastingen. Dit betekent dat er N-3000 niet overgeven dat ze een rekening hebben in Nederland.
    De Nederlandse overheid heeft een lijst van n = 5000 Belgen die een rekening hebben in Nederland. Deze lijst kan dus opgevat worden als een steekproef van 5000 uit de populatie van Belgen met een rekening in Nederland. Hierbij zijn er k = 2000 die het ook aangegeven hebben aan de belastingen (en dus 3000 die het niet aangegeven hebben).
    Het schema ziet er nu dus als volgt uit.

    $$\begin{array}{ccccc} N & = & 3000 & + & N-3000 \\ \downarrow & & \downarrow & & \downarrow \\ 5000 & = & 2000 & + & 3000 \end{array}$$

    De veranderlijke X die hypergeometrisch verdeeld is, is dus:
    X = het aantal Belgen uit de steekproef van 5000 die een rekening in Nederland hebben en dit aangegeven hebben aan de belastingen.
    De verwachtingswaarde van X is: $E(X) = \frac{nr}{N} = \frac{5000 \cdot 3000}{N}$
    Beschouwen het aantal uit de steekproef (2000) als benadering voor deze verwachtingswaarde dan kunnen we hieruit N schatten:

$$2000 = \frac{5000 \cdot 3000}{N} \Rightarrow N = \frac{5000 \cdot 3000}{2000} = 7500$$

- Extra uitleg bij vb 6.15 p 99
  Het schema dat je hier ter verduidelijking kan gebruiken is:
  Voor $P(X = 1)$:

100	=	5(defect)	+	95 (goed)
↓		↓		↓
10	=	1	+	9

  Voor $P(X = 0)$:

100	=	5(defect)	+	95 (goed)
↓		↓		↓
10	=	0	+	10