

Vorlesung 1b

Wiederholte rein zufällige Wahl

(aus endlich vielen möglichen Ausgängen)

mit dem Beispiel

“Wahrscheinlichkeit von Kollisionen”

1. Die Frage

$n = 25$ Individuen,

$r = 365$ Plätze.

Jedes Individuum wird auf einen
rein zufällig ausgewählten Platz gesetzt.

(Mehrfachbelegungen sind erlaubt!)

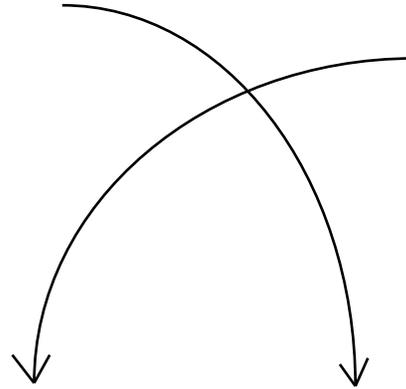
Wie wahrscheinlich ist es, dass dabei
keine Mehrfachbelegung auftritt?

Individuen

1

...

n



Plätze

1

...

r

Ein anderer Blick:

Jedes von n Individuen

ist mit je einem von r möglichen Kennzeichen versehen,
das vom Zufall bestimmt ist.

Wie stehen die Chancen,

dass alle Individuen verschieden gekennzeichnet sind?

Oder anders gesagt:

dass keine zwei der n Individuen gleich gekennzeichnet sind?

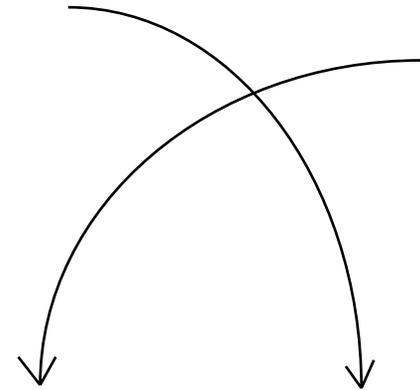
(ein Zitat aus “dem Buch”, Seite 1)

Individuen

1

...

n



Kennzeichen

1

...

r

Eine in der Informatik populäre Sicht:

Man denkt man bei den Individuen an Daten (“keys”)
und spricht bei den Kennzeichen
von Hash-Werten oder Fingerabdrücken.

Populäre Version:

$n = 25$ Leute auf einer Party

Kennzeichen ... Geburtstag ($\in \{1, 2, \dots, 365\}$)

Wie wahrscheinlich ist es,
dass keine zwei Leute am selben Tag Geburtstag haben?

2. Beschreibung durch
eine Zufallsvariable und ein Ereignis.

Die Individuen denken wir uns mit 1 bis n
und die Kennzeichen mit 1 bis r nummeriert.

Ein **Ausgang** der Kennzeichnung lässt sich beschreiben
durch das n -tupel

$$a = (a_1, \dots, a_n),$$

wobei a_i das Kennzeichen des i -ten Individuums bezeichnet

$$(1 \leq a_i \leq r).$$

Die Menge der möglichen Ausgänge der Kennzeichnung ist

$$S := \{1, \dots, r\}^n,$$

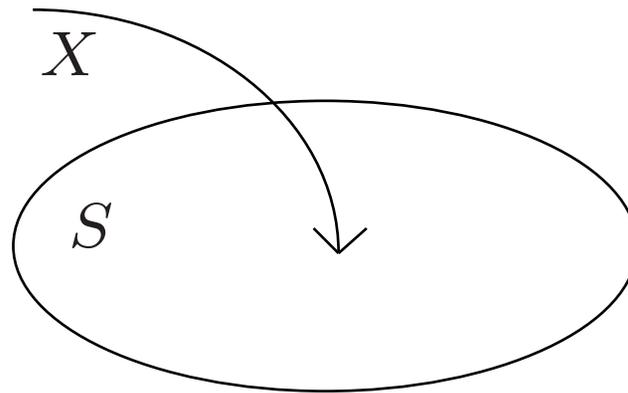
die Menge aller n -tupel (a_1, \dots, a_n) mit $a_i \in \{1, \dots, r\}$.

Manchmal schreiben wir dafür auch

$$S := \{1, \dots, r\}^{\{1, \dots, n\}},$$

die Menge aller Abbildungen von $\{1, \dots, n\}$ nach $\{1, \dots, r\}$.

Den zufälligen Ausgang der Kennzeichnung beschreiben wir durch eine *Zufallsvariable* X .



X kommt durch zufällige Wahl eines Elementes aus S zustande.

Die Menge S heißt *Zielbereich* (oder *Wertebereich*) der Zufallsvariable X .

Wie jedes Element (a_1, \dots, a_n) unserer Menge S besteht auch die Zufallsvariable X aus n Komponenten:

$$X = (X_1, \dots, X_n) .$$

Wir interessieren uns für das *Ereignis*,
dass **keine zwei Komponenten von X gleich** sind.

Dieses Ereignis schreiben wir als

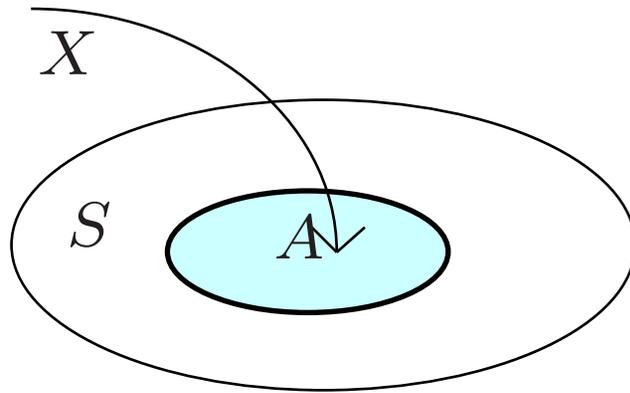
$$\{X_i \neq X_j \text{ für alle } i \neq j\}$$

oder auch als

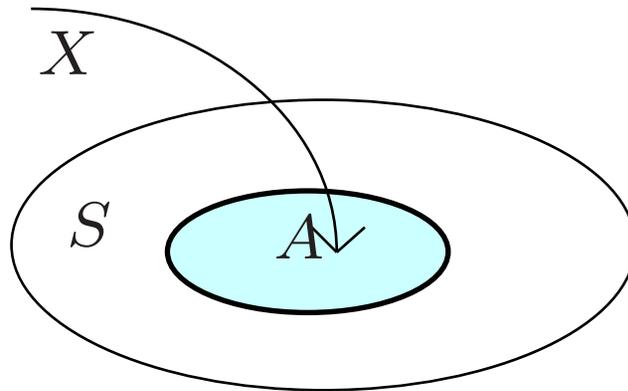
$$\{X \in A\}$$

mit der Teilmenge

$$A := \{a \in S : a_i \neq a_j \text{ für alle } i \neq j\} .$$



3. Die Wahrscheinlichkeit für “Kollisionsfreiheit”



Wie kommt man zur **Wahrscheinlichkeit**
des Ereignisses $\{X \in A\}$?

Zur Erinnerung:

Wahrscheinlichkeiten gehören zu Ereignissen
und messen deren Chance einzutreten
mit einer Zahl zwischen 0 und 1:

$$P(X \in A)$$

Zwei einleuchtende Regeln
für das Rechnen mit Wahrscheinlichkeiten:

$$\mathbf{P}(X \in S) = 1 .$$

d.h. das *sichere Ereignis* $\{X \in S\}$ hat Wahrscheinlichkeit 1

und (bei endlich vielen möglichen Ausgängen, d.h. $\#S < \infty$)

$$\mathbf{P}(X \in A) = \sum_{a \in A} \mathbf{P}(X = a)$$

(Additivität.)

Um die Wahrscheinlichkeit $\mathbf{P}(X \in A)$ berechnen zu können,

muss man eine **Modellannahme** treffen.

Eine prominente Modellannahme ist die einer

rein zufälligen Wahl.

Damit ist gemeint, dass für je zwei $a, a' \in S$

$$\mathbf{P}(X = a) = \mathbf{P}(X = a').$$

Das heißt: kein Ausgang ist bevorzugt.

Also:

$$P(X = a) = \frac{1}{\#S}, \quad a \in S.$$

Und

$$P(X \in A) = \frac{\#A}{\#S}.$$

Wir haben nun die Aufgabe des Abzählens der Mengen

$$S := \{1, \dots, r\}^n = \{(a_1, \dots, a_n) : 1 \leq a_i \leq r\}$$

und

$$A := \{a \in S : a_i \neq a_j \text{ für alle } i \neq j\}$$

$$\#S = r^n$$

$$S = \{(a_1, \dots, a_n) : 1 \leq a_i \leq r\}$$

$$A := \{a \in S : a_i \neq a_j \text{ für alle } i \neq j\}$$

$$\#A = ?$$

Für a_1 gibt es r mögliche Werte, für a_2 dann noch $r - 1$, usw.

Also:

$$\#A = r(r - 1) \cdots (r - (n - 1))$$

$$\mathbf{P}(X \in A) = \frac{r(r-1)\cdots(r-(n-1))}{r^n}$$

$$= \frac{r-1}{r} \frac{r-2}{r} \cdots \frac{r-(n-1)}{r}$$

$$\mathbf{P}(X \in A) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{r}\right).$$

4. Die Approximation durch Linearisieren von \exp

Eine Formel aus der Mathe 1 (“AnaLinA”):

$$e^{-h} = 1 - h + o(h) \quad \text{für } h \rightarrow 0.$$

Dabei ist $o(h)$ von kleinerer Ordnung als h , d.h. $\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$.

Damit bekommen wir die “AnaLinA - Approximation:”

$$\begin{aligned} \prod_{i=1}^{n-1} \left(1 - \frac{i}{r}\right) &\approx \prod_{i=1}^{n-1} \exp\left(-\frac{i}{r}\right) \\ &= \exp\left(-\sum_{i=1}^{n-1} \frac{i}{r}\right) = \exp\left(-\frac{(n-1)n}{2r}\right). \end{aligned}$$

Also: $\mathbf{P}(X \in A) \approx \exp\left(-\frac{(n-1)n}{2r}\right).$

$$\mathbf{P}(X \in A) \approx \exp\left(-\frac{(n-1)n}{2r}\right).$$

Wegen der Verwendung der Approximation

$$1 - \frac{j}{r} \approx e^{-\frac{j}{r}}, \quad j < n,$$

sollte hier $\frac{n}{r}$ klein sein, sprich $n \ll r$.

Gibt es auch eine Näherungsformel,
die brauchbar ist für alle $n < r$?

Wohlan!

5. Die Stirling-Approximation

$$\mathbf{P}(X \in A) = \frac{r(r-1) \cdots (r-(n-1))}{r^n}$$

$$= \frac{r}{r} \frac{r-1}{r} \frac{r-2}{r} \cdots \frac{r-(n-1)}{r}$$

$$\mathbf{P}(X \in A) = \frac{r!}{r^n (r-n)!}$$

mit $k! := 1 \cdot 2 \cdots k$, lies: k -Fakultät

Die Stirling-Formel:

$$k! \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k$$

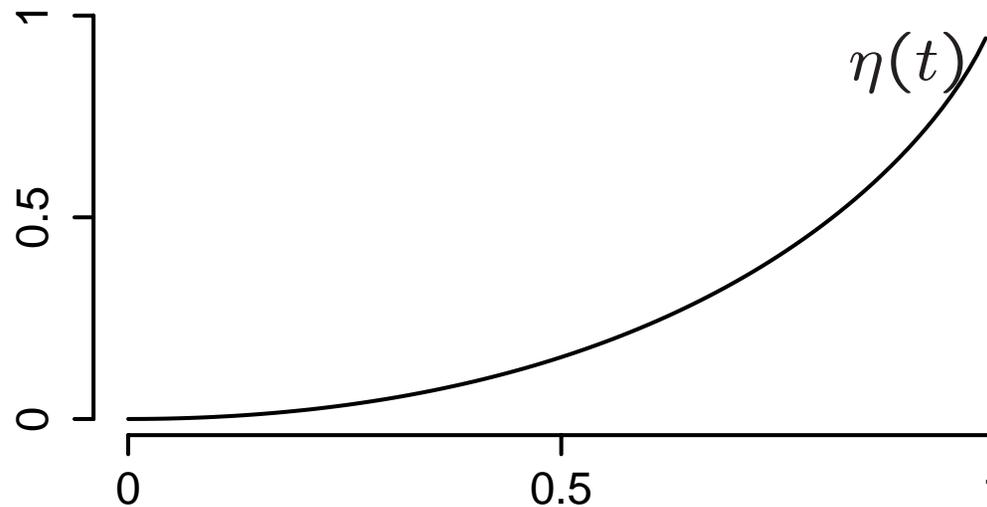


Abraham de Moivre (1667-1754)

$$\begin{aligned} e^{-n} \left(1 - \frac{n}{r}\right)^{n-r} &= \exp \left(-n + (n-r) \ln \left(1 - \frac{n}{r}\right) \right) \\ &= \exp \left(-r \left(\frac{n}{r} + \left(1 - \frac{n}{r}\right) \ln \left(1 - \frac{n}{r}\right) \right) \right) \\ &= \exp \left(-r \eta \left(\frac{n}{r} \right) \right) \end{aligned}$$

mit $\eta(t) := t + (1-t) \ln(1-t)$, $0 \leq t \leq 1$.

$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$

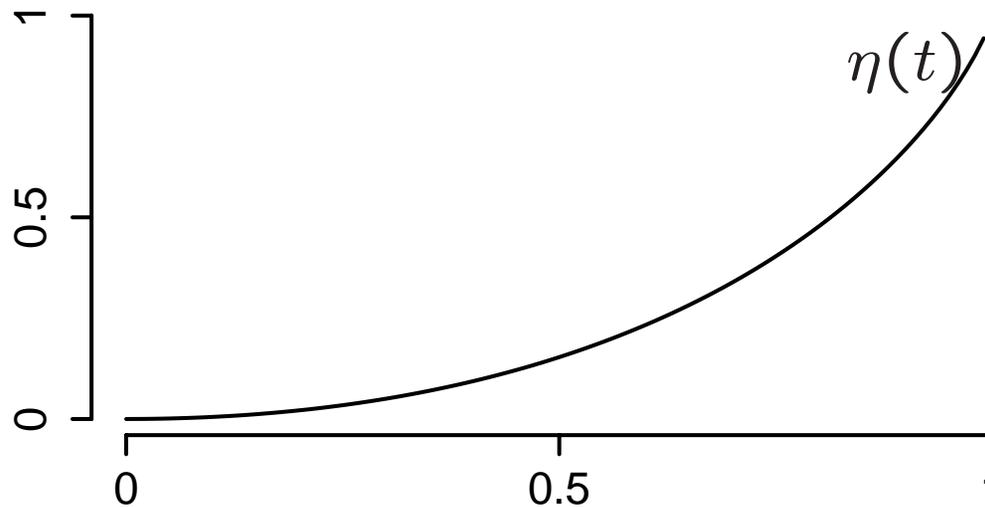


$$\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1 - \frac{n}{r}}} \exp\left(-r \eta\left(\frac{n}{r}\right)\right)$$

Stirling-Approximation

der Wahrscheinlichkeit für Kollisionsfreiheit

$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$

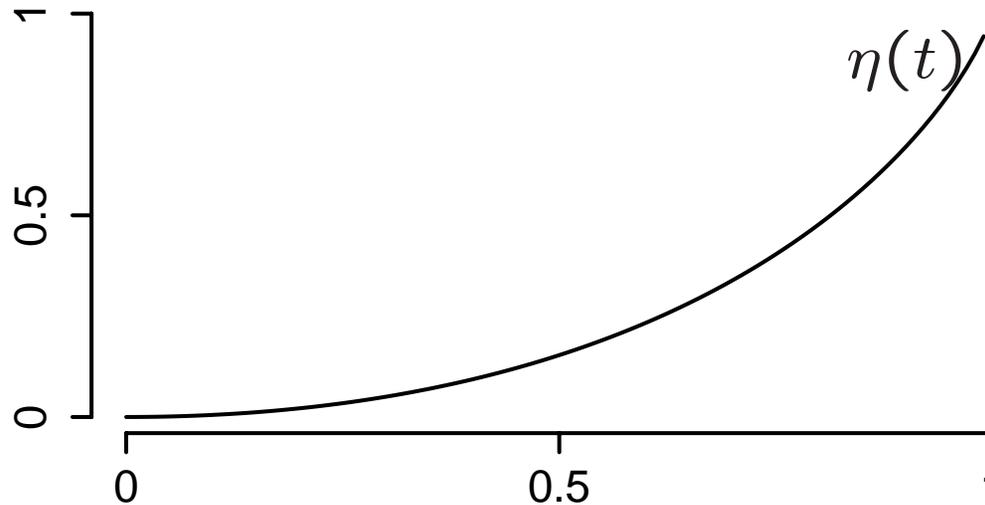


$$\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1 - \frac{n}{r}}} \exp\left(-r \eta\left(\frac{n}{r}\right)\right)$$

Für $n \ll r$, also $t := \frac{n}{r} \ll 1$,

können wir $\eta(t)$ quadratisch approximieren:

$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$

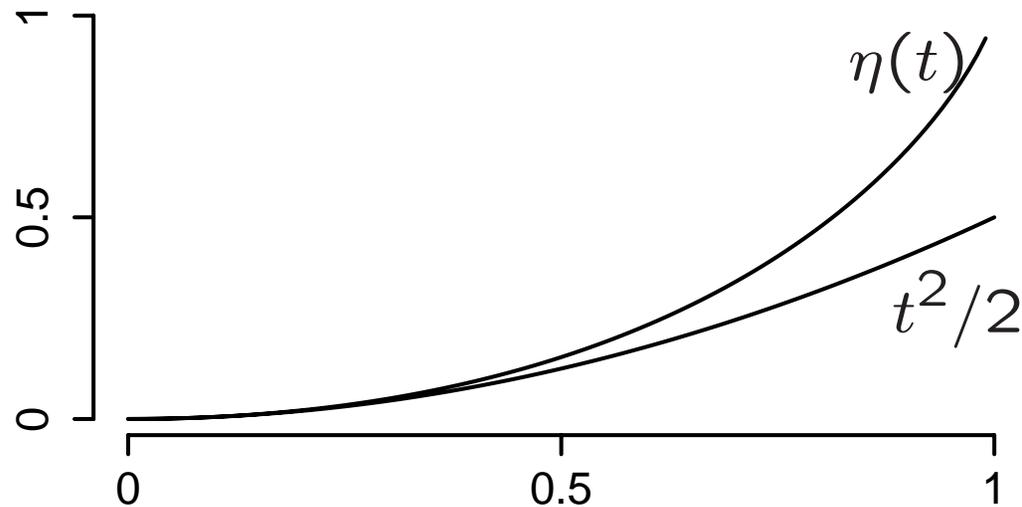


$$\ln(1 - t) = -\left(t + \frac{t^2}{2} + \frac{t^3}{3} + \dots\right)$$

$$(1 - t) \ln(1 - t) = -t + \frac{t^2}{2} + o(t^2) \text{ für } t \rightarrow 0$$

$$\eta(t) = \frac{t^2}{2} + o(t^2) \text{ für } t \rightarrow 0$$

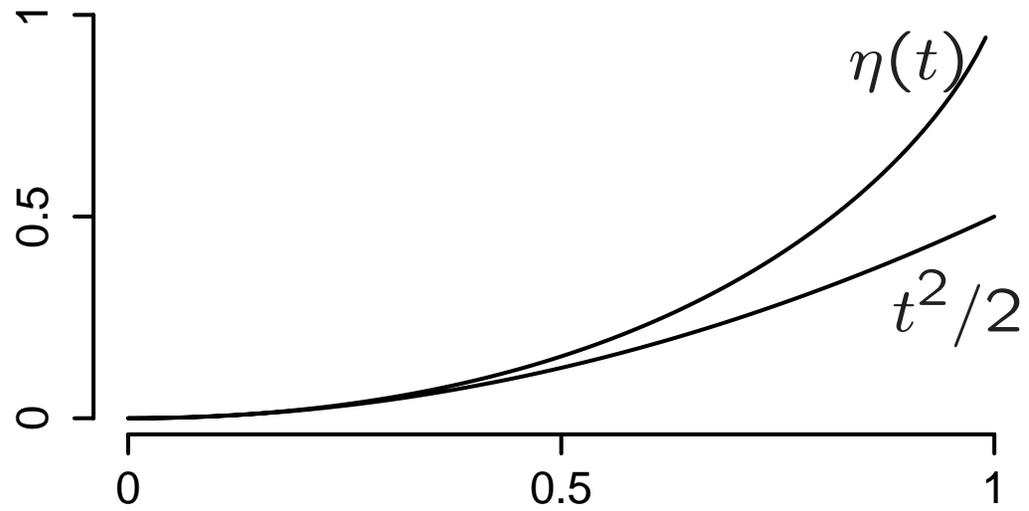
$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$



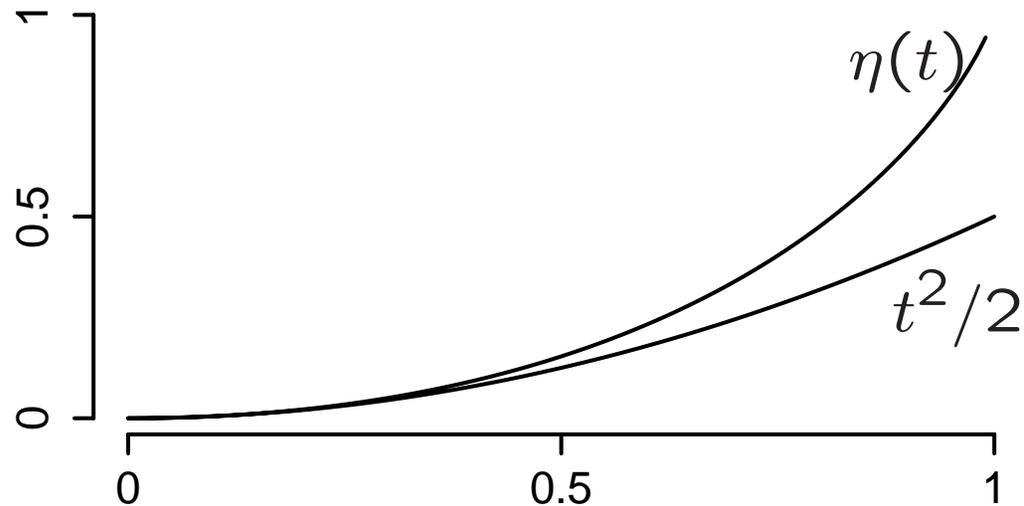
$$\ln(1 - t) = -\left(t + \frac{t^2}{2} + \frac{t^3}{3} + \dots\right)$$

$$(1 - t) \ln(1 - t) = -t + \frac{t^2}{2} + o(t^2) \text{ für } t \rightarrow 0$$

$$\eta(t) = \frac{t^2}{2} + o(t^2) \text{ für } t \rightarrow 0$$



$\frac{1}{2}t^2$ ist die quadratische Approximation von $\eta(t)$ um $t = 0$.



$\frac{1}{2}t^2$ ist die quadratische Approximation von $\eta(t)$ um $t = 0$.

Für $n \ll r$ (wie z. B. für $n = 25$, $r = 365$) ist also

$$\eta\left(\frac{n}{r}\right) \approx \frac{1}{2} \cdot \left(\frac{n}{r}\right)^2.$$

Fazit:

Für $n < r$ ist $\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1 - \frac{n}{r}}} \exp\left(-r \eta\left(\frac{n}{r}\right)\right)$

Fazit:

Für $n < r$ ist $\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1 - \frac{n}{r}}} \exp\left(-r \eta\left(\frac{n}{r}\right)\right)$

mit $\eta(t) = t + (1 - t) \ln(1 - t)$, $0 \leq t \leq 1$

(Stirling-Approximation).

Für $n \ll r$ ist $\mathbf{P}(X \in A) \approx \exp\left(\frac{n}{2r}\right) \exp\left(-\frac{n^2}{2r}\right)$
 $= \exp\left(-\frac{n(n-1)}{2r}\right)$

(Stirling+Taylor-Approximation)

Für $n^2 \ll r$ ist $\mathbf{P}(X \in A) \approx 1$.

Man beachte:

Die Stirling+Taylor Approximation

$$\mathbf{P}(X \in A) \approx \exp\left(-\frac{n(n-1)}{2r}\right)$$

ist identisch mit der “AnaLinA-Approximation”.

Beispiel: $n = 25, r = 365$:

$$\mathbf{P}(X \in A) = \frac{r(r-1) \cdots (r-n+1)}{r^n} = 0.431300$$

$$\frac{1}{\sqrt{1 - \frac{n}{r}}} \exp\left(-r \eta\left(\frac{n}{r}\right)\right) = 0.431308$$

$$\exp\left(-\frac{n(n-1)}{2r}\right) = 0.4396$$

6. Der Zeitpunkt der ersten Kollision.

Ein dynamisches Bild:

Wir denken uns r fest und lassen n laufen ($n = 1, 2, \dots$)

Vorstellung: Ein Individuum nach dem anderen wird auf einen (immer wieder neu) rein zufällig gewählten Platz gesetzt.

Die Folge (X_1, X_2, \dots) der gewählten Plätze ist dann eine rein zufällige $1 \dots r$ -Folge

$A_n :=$ “ keine Kollision bis (einschließlich) n ”

$T :=$ Zeitpunkt der ersten Kollision

(ist ablesbar aus (X_1, X_2, \dots)). Es gilt

$$A_n = \{T > n\}$$

also insbesondere auch

$$\mathbf{P}(A_n) = \mathbf{P}(T > n).$$

Wir haben somit AUCH die Verteilung der Zufallsvariablen T
(sowohl exakt als auch näherungsweise) berechnet.

Die Ferebee-Wandtner'schen Programme zur VL 1b (siehe Homepage)
geben dazu eine erhellende Illustration.