

## Vorlesung 1b

# Wiederholte rein zufällige Wahl

(aus endlich vielen möglichen Ausgängen)

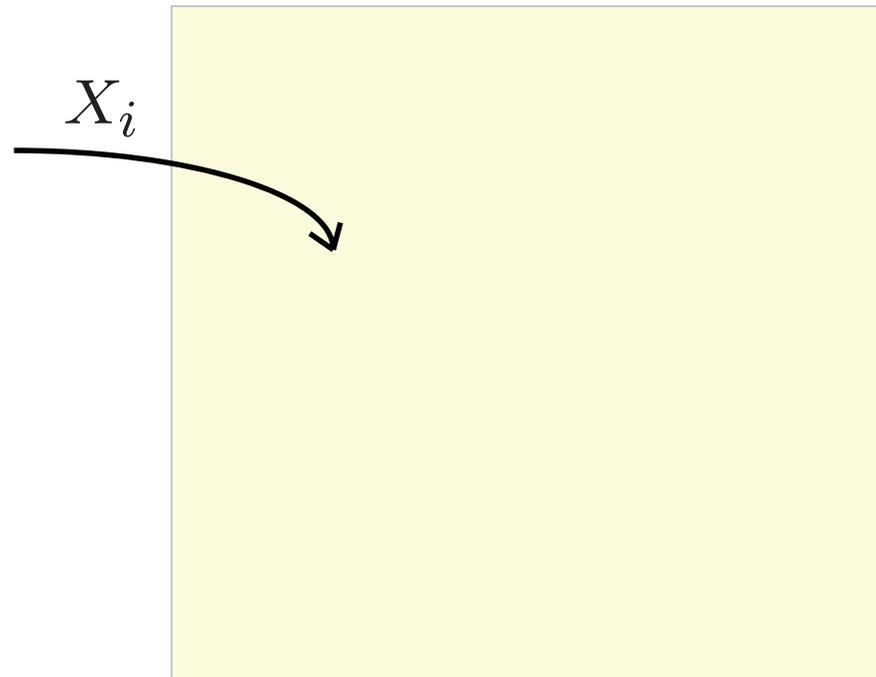
mit dem Beispiel

“Wahrscheinlichkeit von Kollisionen”

(Buch S. 1-5)

# 1. Erinnerung an die erste Vorlesung und Fragestellung

$n(= 100)$ -mal wiederholt wird rein zufällig  
ein Pixel aus  $g = 10^6$  Pixeln gewählt.



$i = 1, \dots, 100$

$n(= 100)$ -mal wiederholt wird rein zufällig  
ein Pixel aus  $g = 10^6$  Pixeln gewählt.

Letztesmal fragten wir  
nach (der Verteilung von) Trefferquoten.

Heute fragen wir:

Wie wahrscheinlich ist es, dass dabei lauter verschiedene  
Pixel gewählt werden (sprich: keine Kollision auftritt)?

Was schätzen Sie

für

$n = 100$  ?

$n = 1000$  ?

$n = 10000$  ?

Tema con variazioni:

$n$  Individuen,  
 $g$  Plätze.

Jedes Individuum wird auf einen  
rein zufällig ausgewählten Platz gesetzt.

(Mehrfachbelegungen sind erlaubt!)

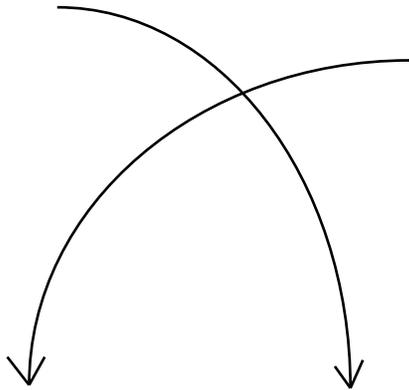
Wie wahrscheinlich ist es, dass dabei  
keine Mehrfachbelegung auftritt?

Individuen

1

...

$n$



Plätze

1

...

$g$

Ein anderer Blick:

Jedes von  $n$  Individuen

ist mit je einem von  $g$  möglichen Kennzeichen versehen, \*

das vom Zufall bestimmt ist.

\* Anders als im Buch Kersting/W. wird hier die Anzahl der Kennzeichen mit  $g$  (und nicht mit  $r$ ) bezeichnet - wegen des besseren typographischen Unterschieds von  $g$  zu  $n$  und der Assoziation von  $g$  mit **G**esamtzahl der Plätze, **G**eburtstage,...

Ein anderer Blick:

Jedes von  $n$  Individuen

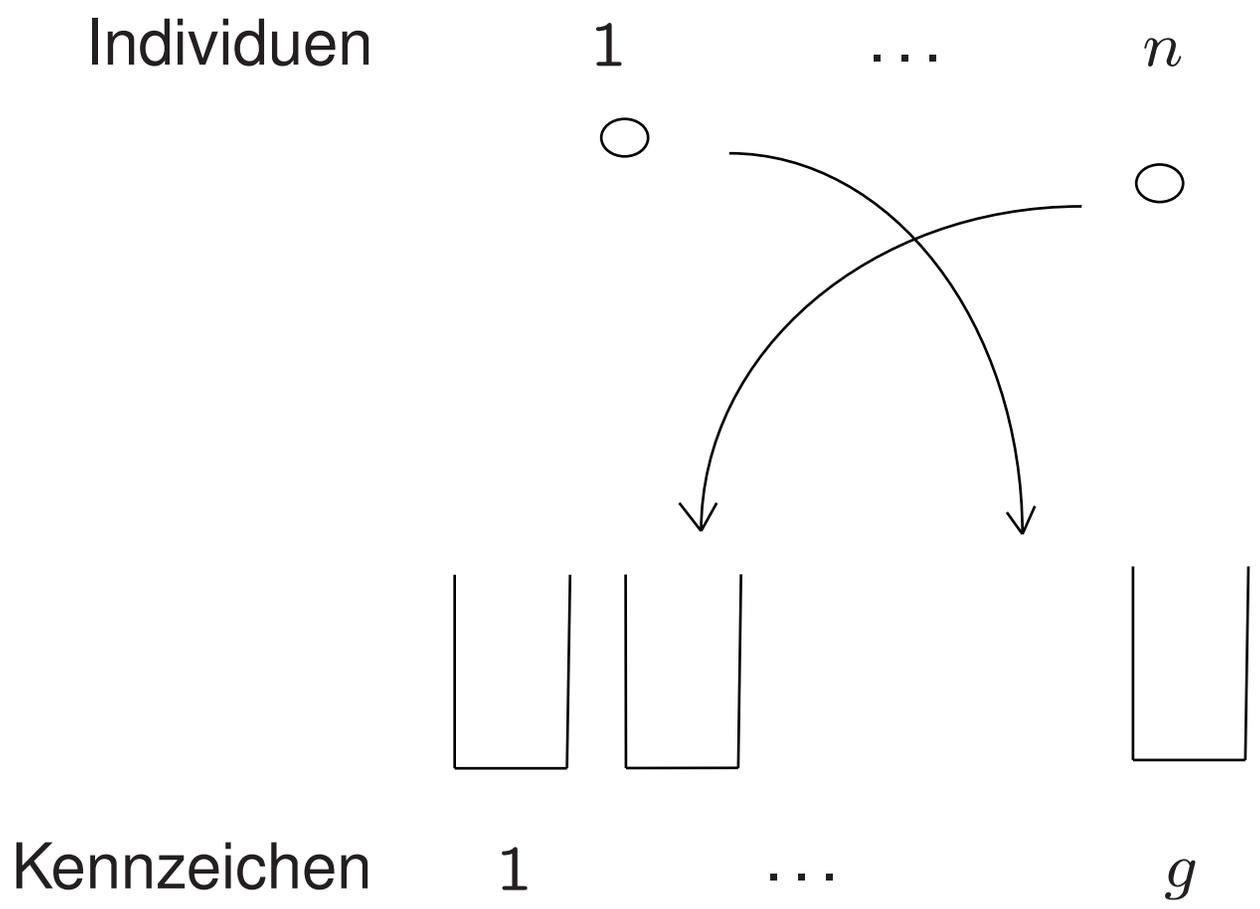
ist mit je einem von  $g$  möglichen Kennzeichen versehen.

Wie stehen die Chancen,

dass alle Individuen verschieden gekennzeichnet sind?

Oder anders gesagt:

dass keine zwei der  $n$  Individuen gleich gekennzeichnet sind?



Eine in der Informatik vertraute 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  $\leftrightarrow$  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.

(Buch S. 2-3)

Die Individuen denken wir uns mit 1 bis  $n$   
und die Kennzeichen mit 1 bis  $g$  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

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

Die Menge der möglichen Ausgänge der Kennzeichnung ist

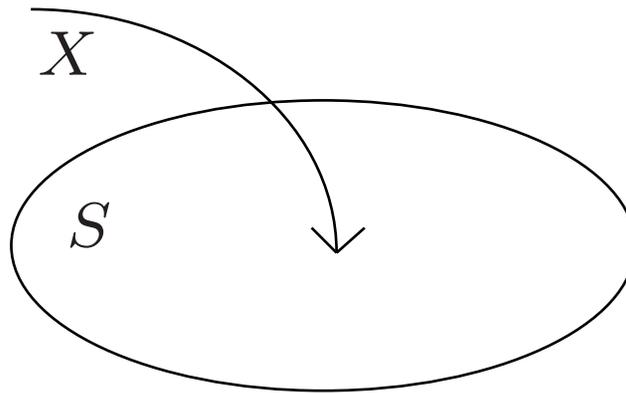
$$S := \{1, \dots, g\}^n,$$

die Menge aller  $n$ -tupel  $(a_1, \dots, a_n)$

mit Einträgen (Komponenten)

$$a_i \in \{1, \dots, g\}, \quad i = 1, \dots, n.$$

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

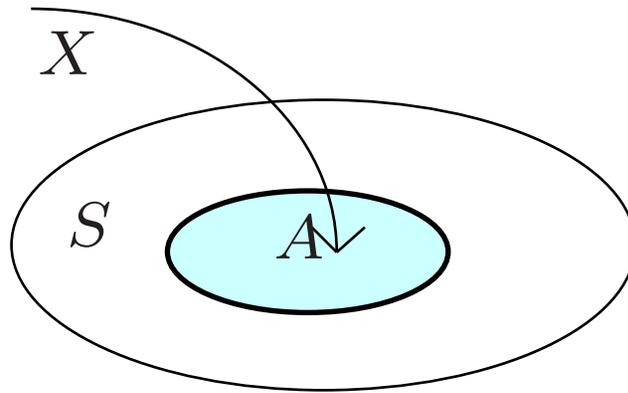
$$\{X_i \neq X_j \text{ für alle } i \neq j\}$$

oder auch als

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

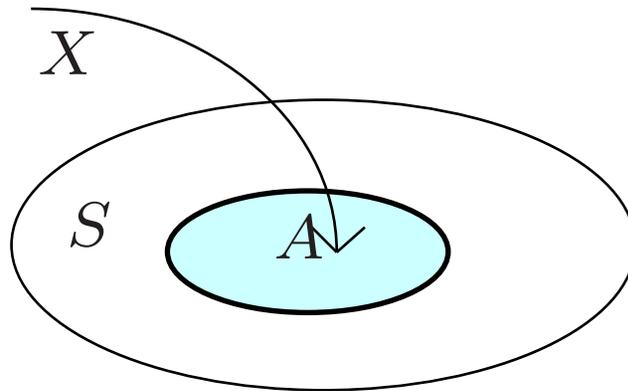
mit der Teilmenge

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



### 3. Die Wahrscheinlichkeit für Kollisionsfreiheit

(Buch S. 3)



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  
(Normierung auf Eins)

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 zwei Mengen

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

und

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

$$\#S = g^n$$

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

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

$$\#A = ?$$

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

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

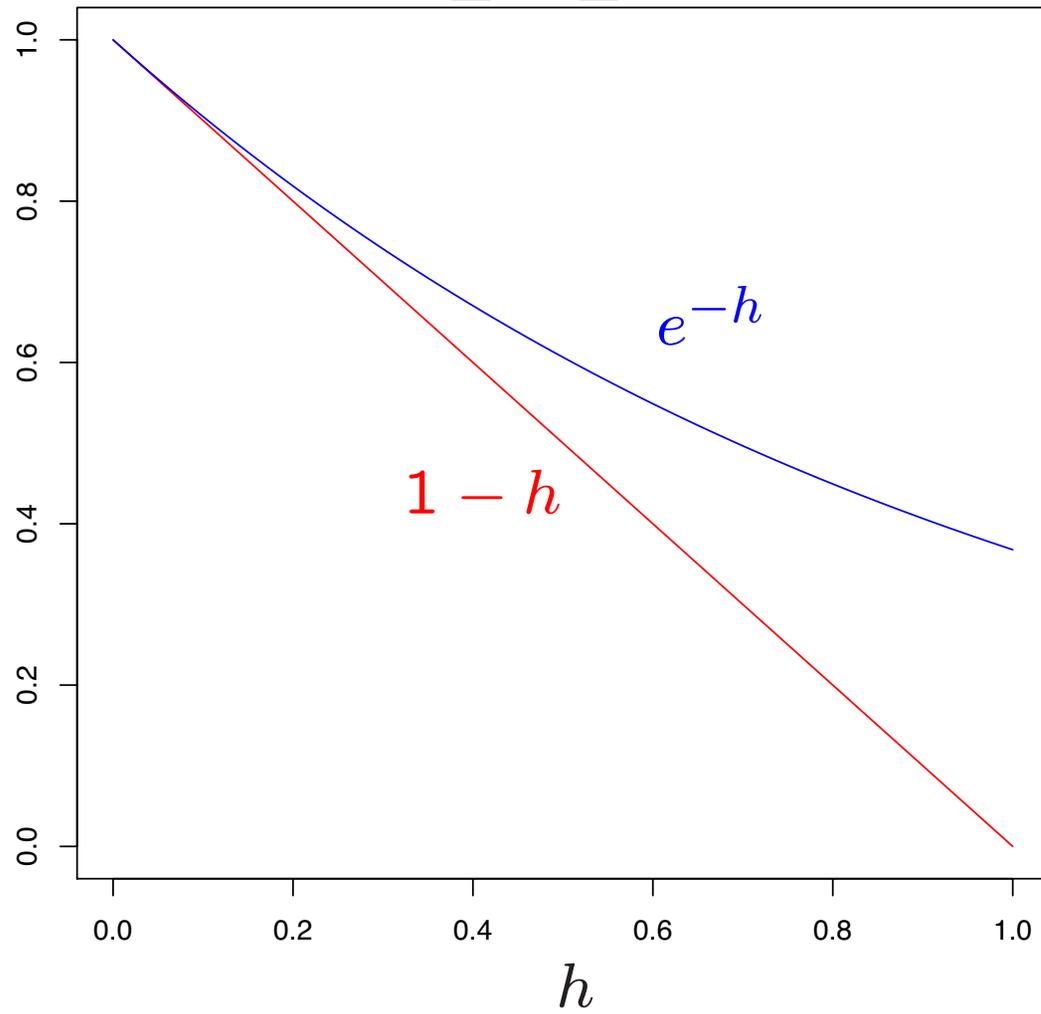
$$\mathbf{P}(X \in A) = \frac{g(g-1) \cdots (g-(n-1))}{g^n}$$

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

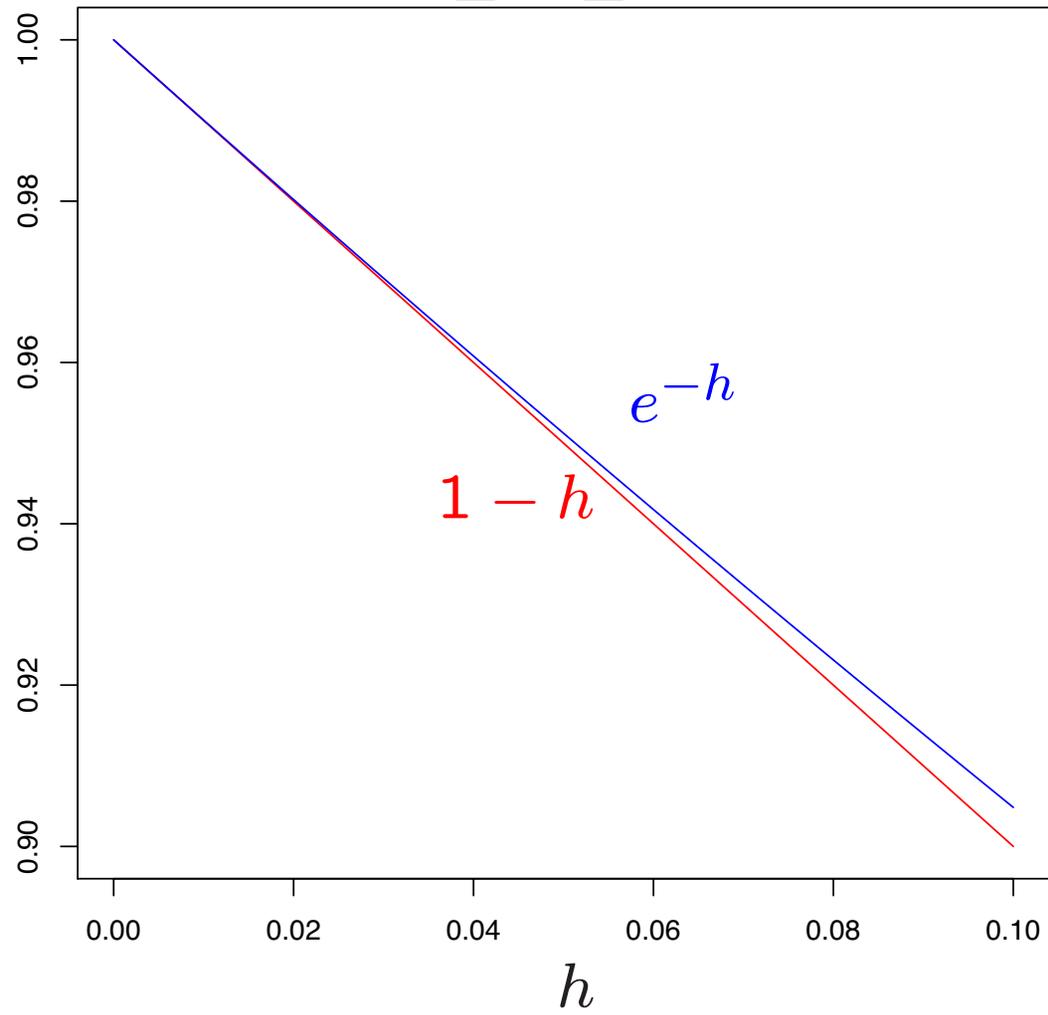
$$\mathbf{P}(X \in A) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{g}\right) .$$

# 4. Die Approximation über die Linearisierung von $\exp$

für  $0 \leq h \leq 1$ :



für  $0 \leq h \leq 0.1$ :



Eine Formel aus der AnaLinA:

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

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

Salopp geschrieben:

$$e^{-h} \approx 1 - h \quad \text{für kleine } h.$$

Damit bekommen wir für kleines  $\frac{n}{g}$  die Approximation

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

Also für kleines  $\frac{n}{g}$ , d.h. für  $n \ll g$ :

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

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

Wohlan!

# 5. Die Stirling-Approximation

(Buch S. 4-5)

$$\begin{aligned} \mathbf{P}(X \in A) &= \frac{g(g-1) \cdots (g-(n-1))}{g^n} \\ &= \frac{g(g-1) \cdots (g-(n-1)) (g-n)(g-n-1) \cdots 1}{g^n (g-n)(g-n-1) \cdots 1} \end{aligned}$$

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

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)

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

$$\begin{aligned} \frac{g!}{g^n(g-n)!} &\approx \frac{1}{g^n} \sqrt{\frac{g}{g-n}} \frac{g^g}{(g-n)^{g-n}} \frac{e^{g-n}}{e^g} \\ &= \sqrt{\frac{g}{g-n}} e^{-n} \left(\frac{g}{g-n}\right)^{g-n} \\ &= \frac{1}{\sqrt{1-\frac{n}{g}}} e^{-n} \left(1 - \frac{n}{g}\right)^{n-g} \end{aligned}$$

$$\frac{g!}{g^n (g-n)!} \approx \frac{1}{\sqrt{1 - \frac{n}{g}}} e^{-n} \left(1 - \frac{n}{g}\right)^{n-g}$$

Wir formen jetzt den Ausdruck

$$e^{-n} \left(1 - \frac{n}{g}\right)^{n-g}$$

weiter um. Ein Trick dabei ist es,

$$1 - \frac{n}{g} \quad \text{als} \quad \exp\left(\ln\left(1 - \frac{n}{g}\right)\right)$$

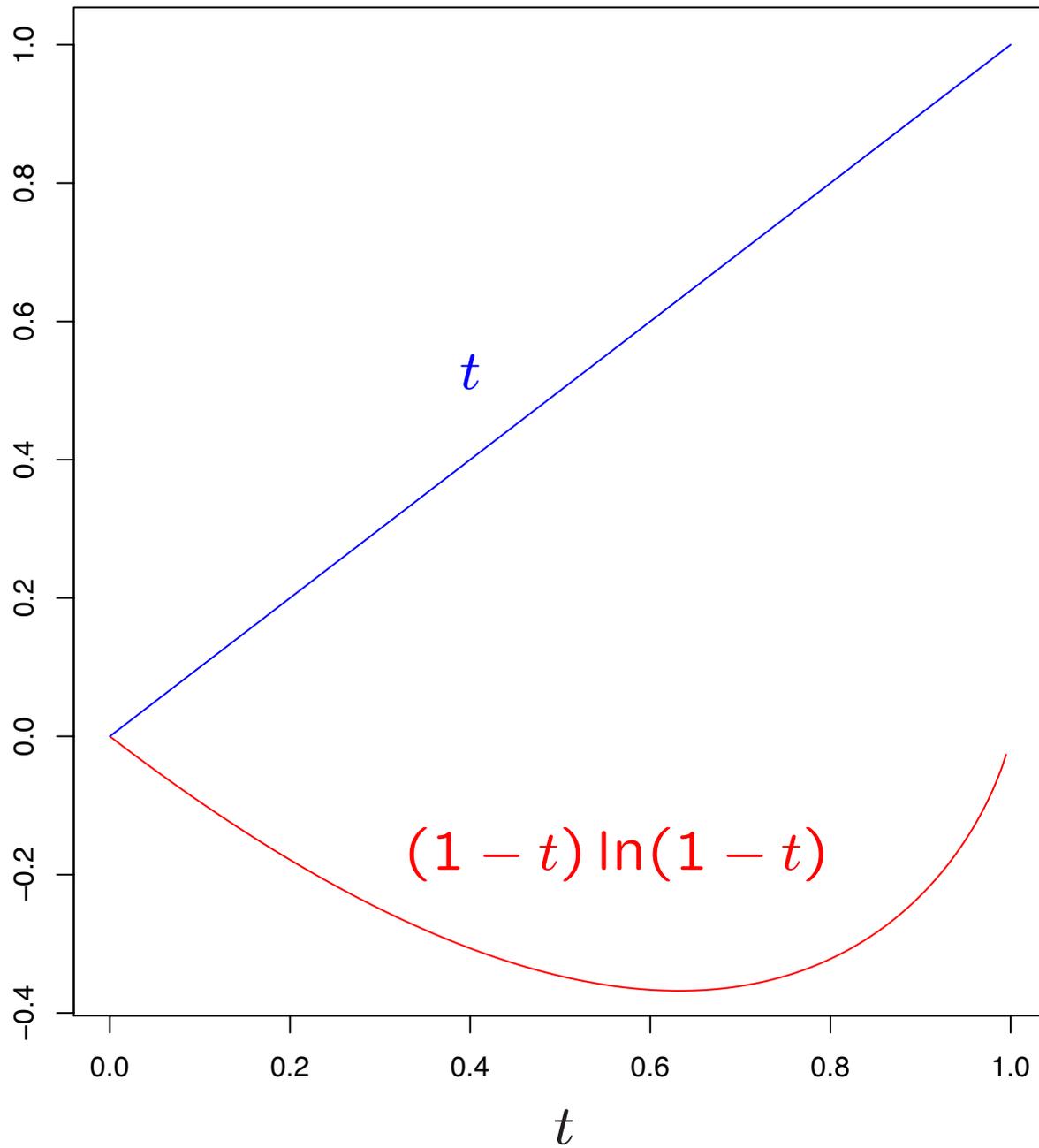
zu schreiben

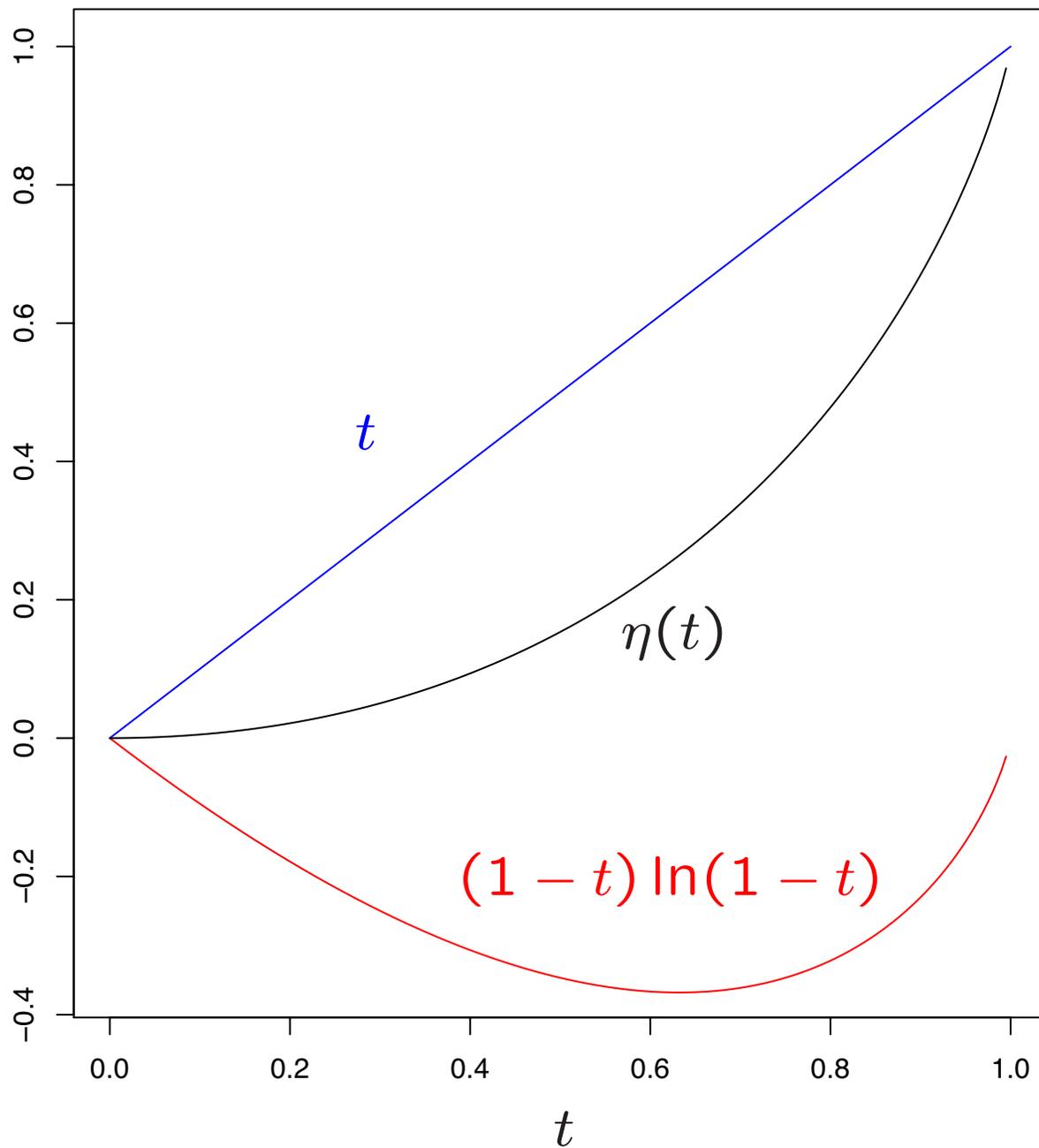
und dann die Beziehung  $e^a e^b = e^{a+b}$  zu verwenden.

$$e^{-n} \left(1 - \frac{n}{g}\right)^{n-g} = \exp \left( -n + (n-g) \ln \left(1 - \frac{n}{g}\right) \right)$$
$$= \exp \left( -g \left( \frac{n}{g} + \left(1 - \frac{n}{g}\right) \ln \left(1 - \frac{n}{g}\right) \right) \right)$$

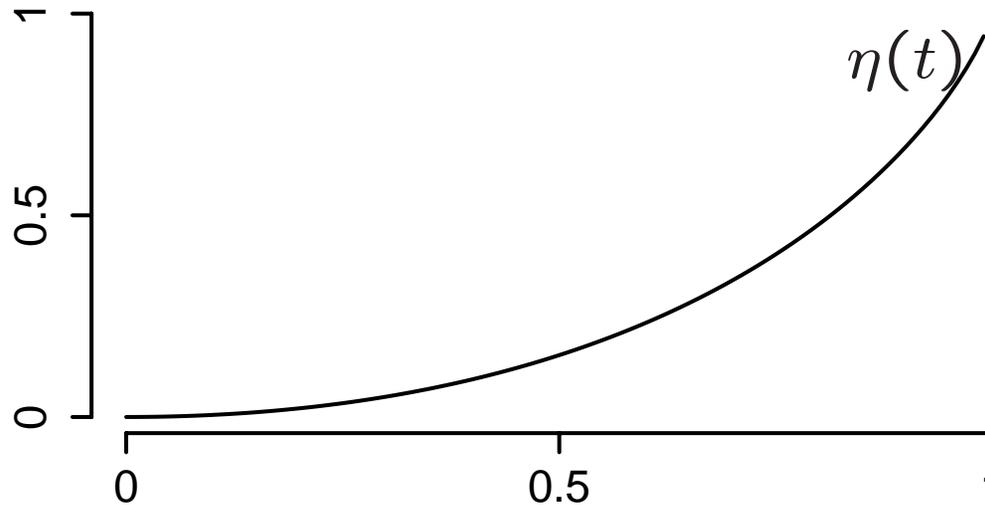
$$\begin{aligned}
e^{-n} \left(1 - \frac{n}{g}\right)^{n-g} &= \exp \left( -n + (n-g) \ln \left(1 - \frac{n}{g}\right) \right) \\
&= \exp \left( -g \left( \frac{n}{g} + \left(1 - \frac{n}{g}\right) \ln \left(1 - \frac{n}{g}\right) \right) \right) \\
&= \exp \left( -g \eta \left( \frac{n}{g} \right) \right)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

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





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

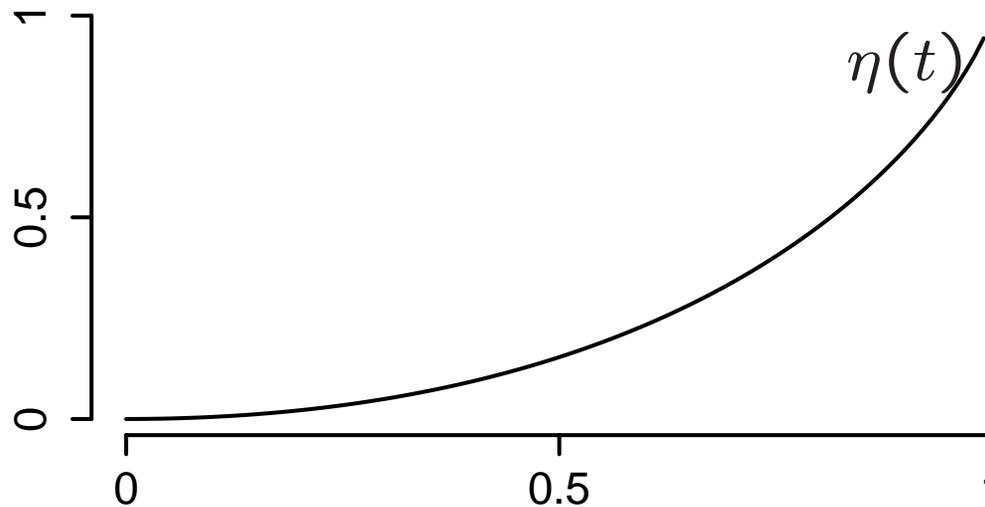


$$\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1 - \frac{n}{g}}} \exp\left(-g \eta\left(\frac{n}{g}\right)\right)$$

Stirling-Approximation

der Wahrscheinlichkeit für Kollisionsfreiheit

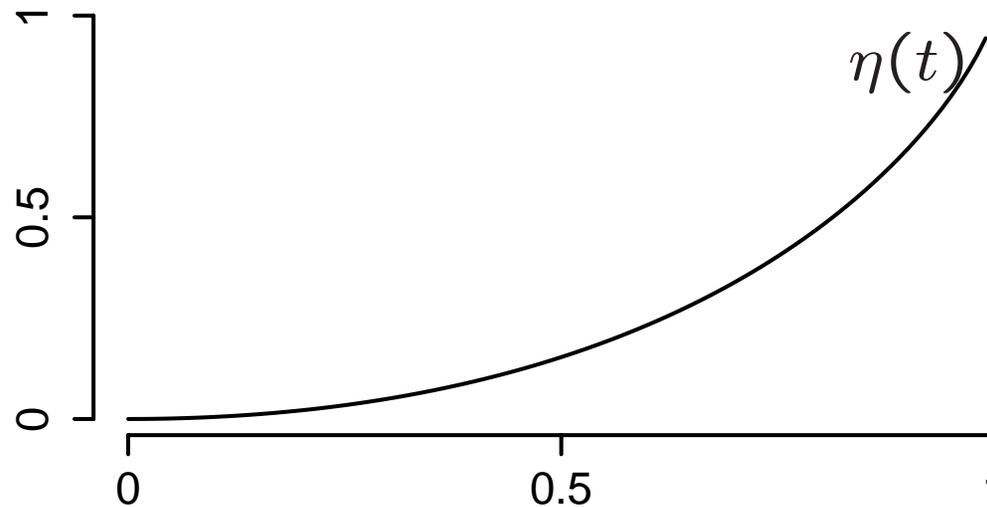
$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$



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

Für  $n \ll g$ , also  $t := \frac{n}{g} \ll 1$ ,  
können wir  $\eta(t)$  quadratisch approximieren:

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

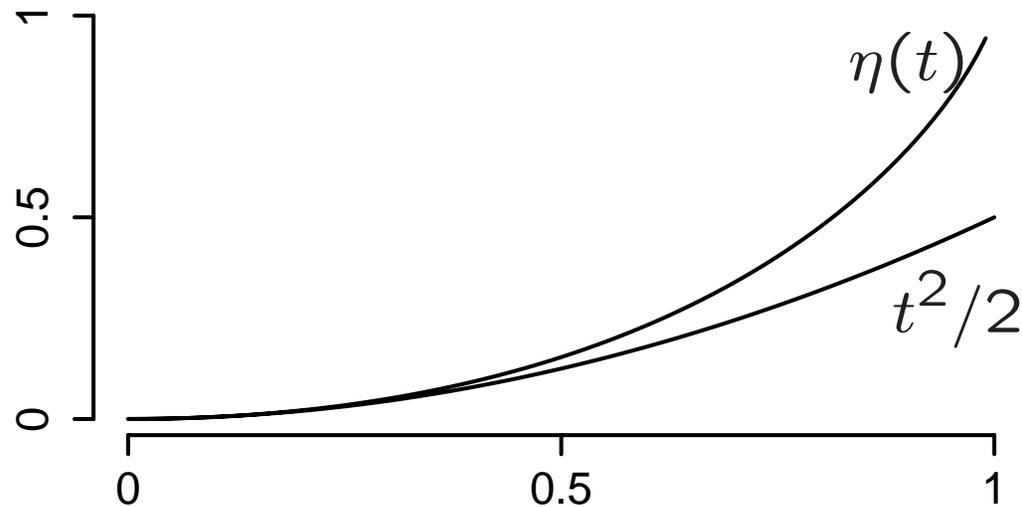


$$\ln(1 - t) = -t - \frac{t^2}{2} - \frac{t^3}{3} - \dots$$

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

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

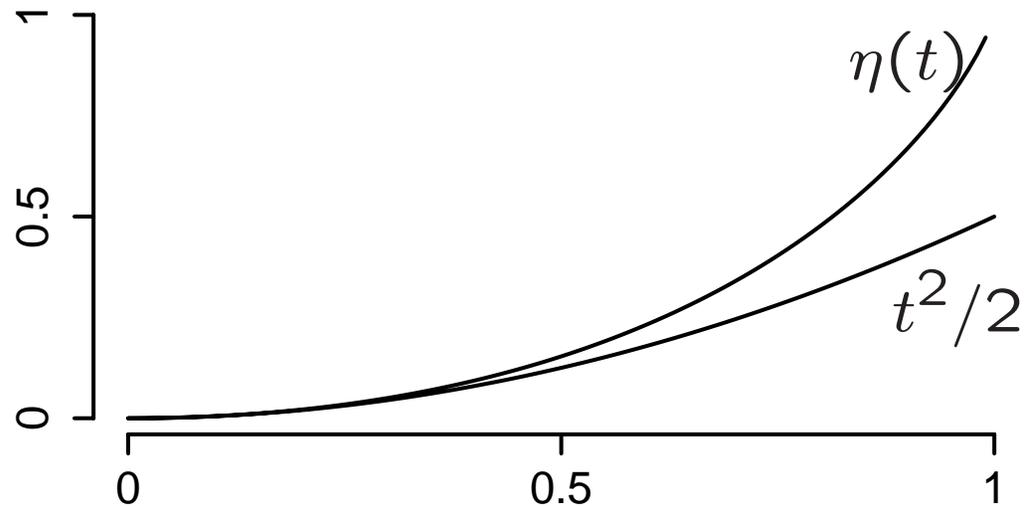
$$\eta(t) := t + (1 - t) \ln(1 - t), \quad 0 \leq t \leq 1.$$



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

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

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



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

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

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

Fazit:

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

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

(Stirling-Approximation).

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

(Stirling+Taylor-Approximation)

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

Man beachte:

Die Stirling+Taylor Approximation

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

ist identisch mit der

Approximation über die Linearisierung von exp.

Im Buch haben wir hier den Faktor  $\sqrt{1 - n/g}$  in der Stirling-Approximation vernachlässigt, d.h. gleich 1 gesetzt. Dadurch erhielten wir dort die etwas weniger genaue Näherung

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

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

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

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

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

## 6. Der Zeitpunkt der ersten Kollision.

## Ein dynamisches Bild:

Wir denken uns  $g$  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  $1 \dots g$ -Folge

$E_n :=$  sei das Ereignis “ keine Kollision bis (einschließlich)  $n$   
”

Die Zufallsvariable  $T :=$  Zeitpunkt der ersten Kollision  
(ist ablesbar aus  $(X_1, X_2, \dots)$ ). Es gilt

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

also insbesondere auch

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

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

Die Ferebee-Wandtner'schen Programme zur Vorlesung 1b (Links auf der Stoff-Seite) geben eine erhellende Illustration

- 1) zum Abnehmen von  $w_n := \mathbf{P}(E_n)$  mit wachsendem  $n$ ,
- 2) zur Verteilung von  $T_n$ ,
- 3) zur Güte der beiden Approximationen
  - (i) über die Linearsierung von  $\exp$
  - (ii) mit Stirling.

Die Voreinstellung\* ist jeweils  $g = 365$ , sie kann im Programm geändert werden.

\*Im Programm wird  $g$  mit  $r$  bezeichnet