Vorlesung 1b

Wiederholte rein zufällige Wahl

(aus endlich vielen möglichen Ausgängen)

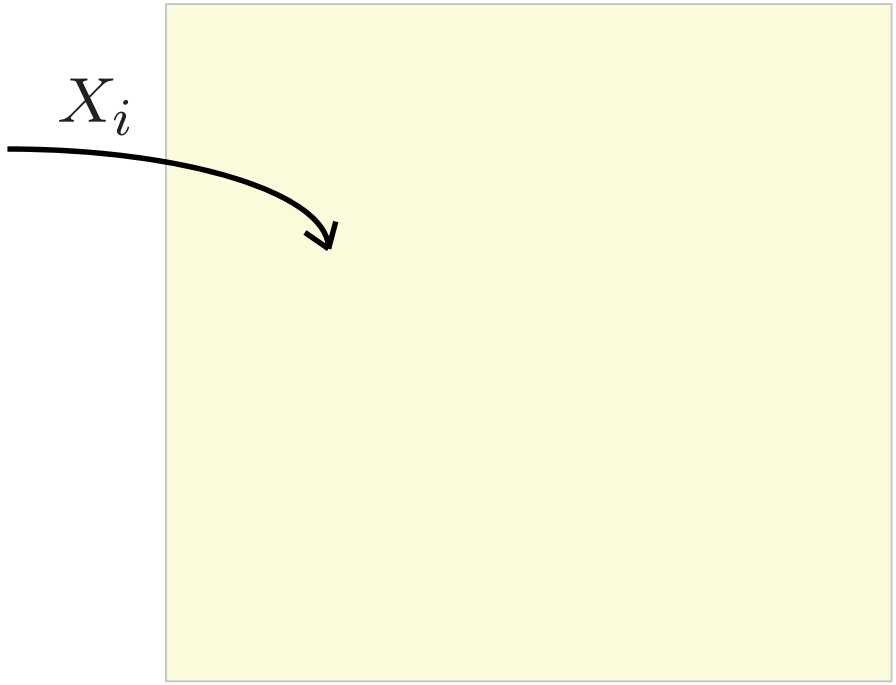
mit dem Beispiel

“Wahrscheinlichkeit von Kollisionen”

1. Erinnerung an die erste Vorlesung und Fragestellung

$n = 100$ mal wiederholt wird rein zufällig

ein Pixel aus $g = 10^6$ Pixeln gewählt.



$i = 1, \ldots, 100$

$n = 100$ 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.

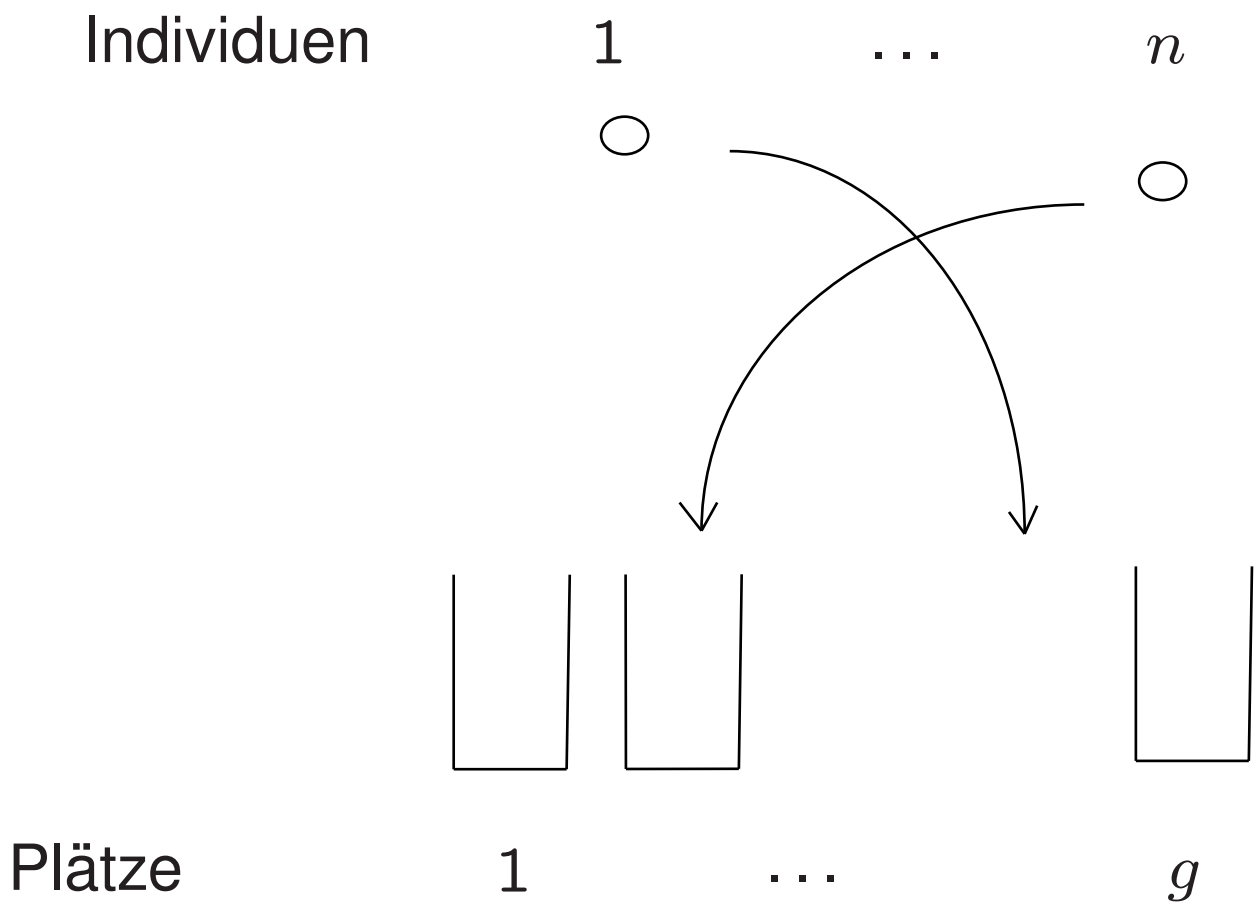
Jedes Individuum wird auf einen

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.

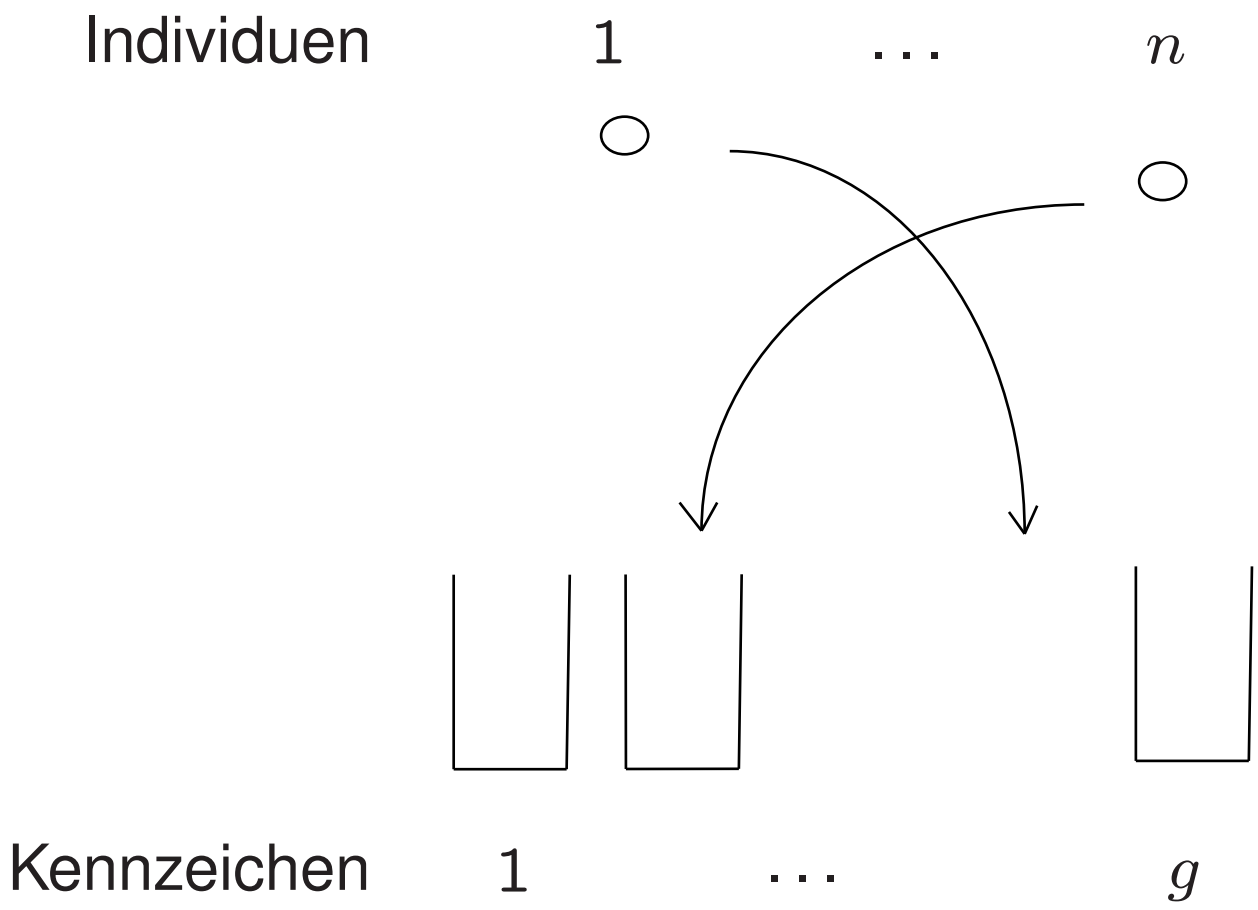
Wie stehen die Chancen,

dass alle Individuen verschieden gekennzeichnet sind?

Oder anders gesagt:

dass keine zwei der n Individuen gleich gekennzeichnet sind?

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


Individuen
1
...
n
Kennzeichen
1
...
g

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 $(\in G := \{1, 2, \ldots, 365\})$

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 g nummeriert.

Ein Ausgang der Kennzeichnung lässt sich beschreiben

durch das n-tupel

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

wobei a_i das Kennzeichen des i-ten Individuums bezeichnet

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

Die Menge der möglichen Ausgänge der Kennzeichnung ist

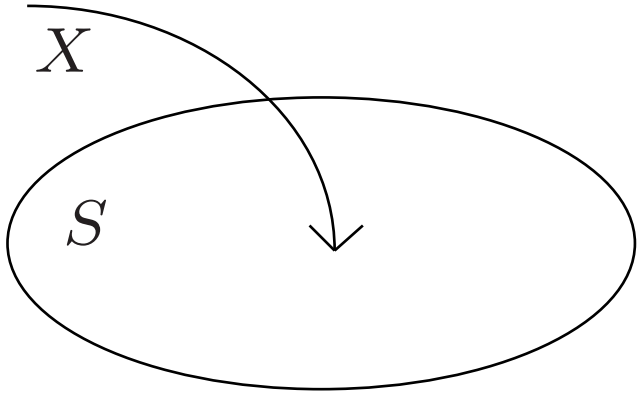
$$S := \{1, \ldots, g\}^n \,,$$

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

mit Einträgen (Komponenten)

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

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 $(a_1, \ldots, a_n)$ unserer Menge S

besteht auch die Zufallsvariable X aus n Komponenten:

$$X = (X_1, \ldots, X_n) \; .$$

Wir interessieren uns für das *Ereignis*,

dass keine zwei Komponenten von X gleich sind.

Dieses Ereignis schreiben wir als

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

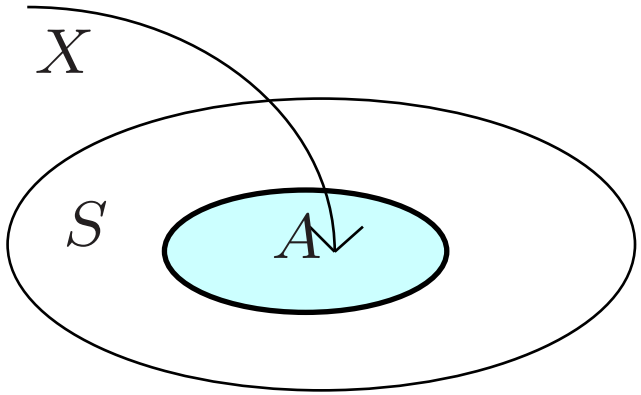
oder auch als

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

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:

$$\mathbf{P}(X \in A)$$

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

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

Wir haben nun die Aufgabe des Abzählens der zwei Mengen

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

und

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

$$\#S = g^n$$

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

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

$$\#A = ?$$

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 durch Linearisieren von exp

Eine Formel aus der AnaLinA:

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

Dabei ist der Term $o(h)$ *von kleinerer Ordnung als* h, d.h. $\lim\limits_{h\to 0} \frac{o(h)}{h} = 0$.

Salopp geschrieben:

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

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

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

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

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

Gibt es auch eine Näherungsformel,

die brauchbar ist für alle $n < g$?

Wohlan!

5. Die Stirling-Approximation

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

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

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

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

Die Stirling-Formel:

(de Moivre (1730) (noch ohne den Vorfaktor $\sqrt{2\pi}$),

dieser wurde gefunden von Stirling (1730))

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

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

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

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

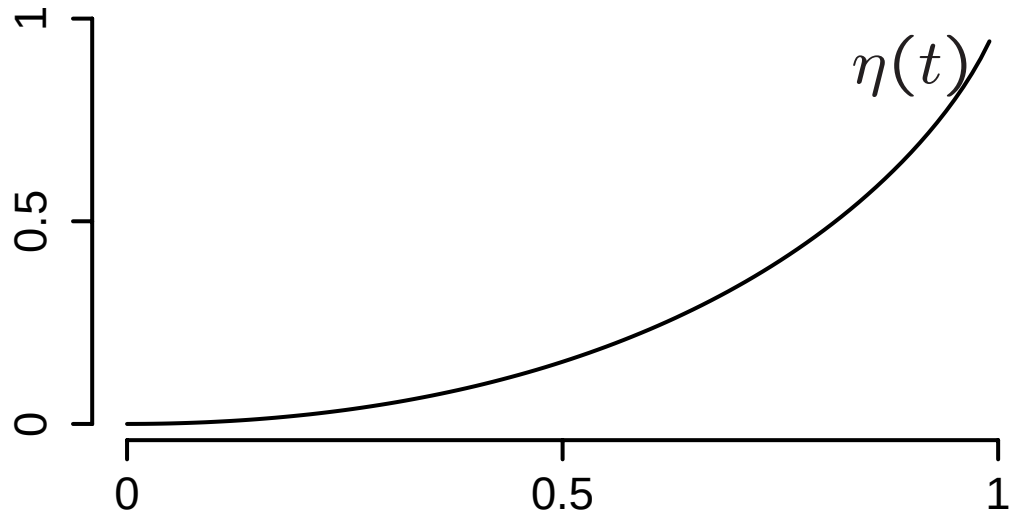
$$e^{-n}\Big(1-\frac{n}{g}\Big)^{n-g} = \exp\Big(-n+(n-g)\ln\big(1-\frac{n}{g}\big)\Big)$$

$$= \exp\Big(-g\,\big(\frac{n}{g}+\big(1-\frac{n}{g}\big)\ln\big(1-\frac{n}{g}\big)\big)\Big)$$

$$= \exp\Big(-g\,\eta\big(\frac{n}{g}\big)\Big)$$

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

$$\eta(t) := t + (1-t)\ln(1-t), \qquad 0 \le t \le 1.$$

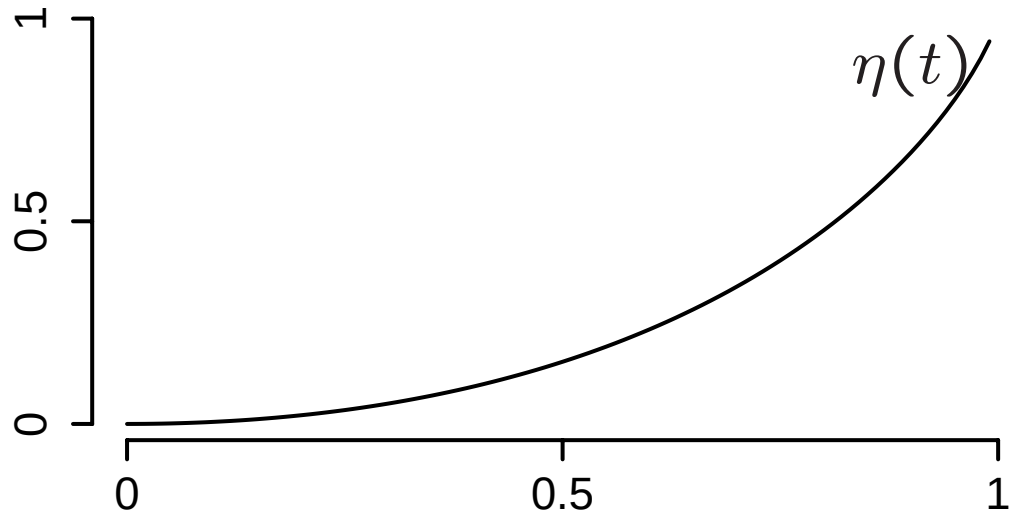


$$\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1-\frac{n}{g}}} \exp\left(-r\,\eta\left(\frac{n}{r}\right)\right)$$

Stirling-Approximation

der Wahrscheinlichkeit für Kollisionsfreiheit

$$\eta(t) := t + (1-t)\ln(1-t), \qquad 0 \le t \le 1.$$

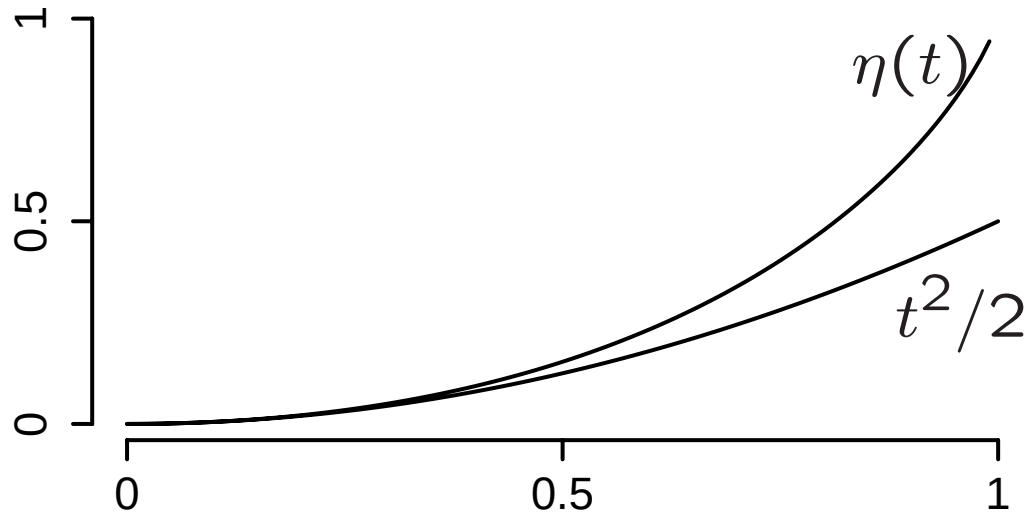


$$\mathbf{P}(X \in A) \approx \frac{1}{\sqrt{1-\frac{n}{g}}} \exp\left(-r\,\eta\Big(\frac{n}{g}\Big)\right)$$

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

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

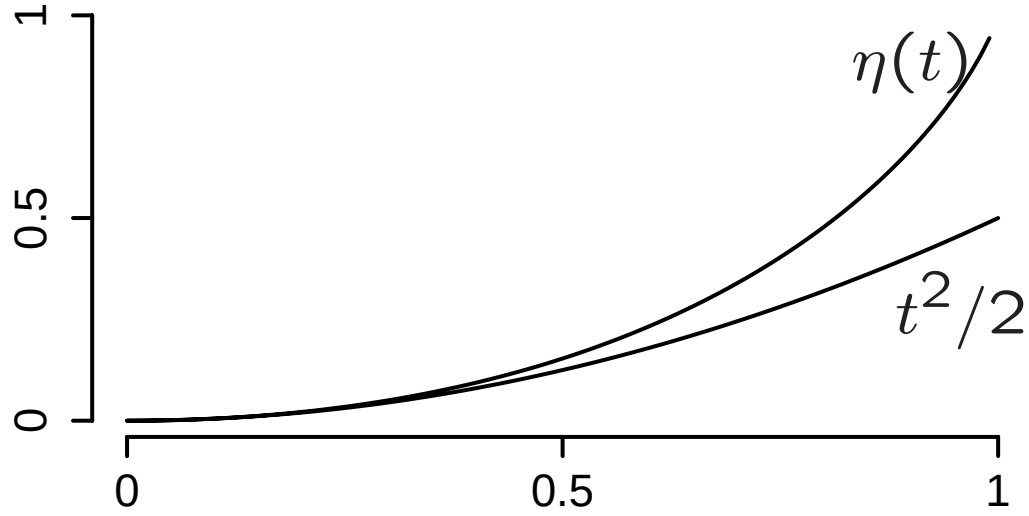
$$\eta(t) := t + (1-t)\ln(1-t), \qquad 0 \le t \le 1.$$



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

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

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



$\frac{1}{2}t^2$ 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 $\quad \mathbf{P}(X \in A) \approx \dfrac{1}{\sqrt{1-\frac{n}{g}}} \exp\left(-g\,\eta\left(\dfrac{n}{g}\right)\right)$

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

(Stirling-Approximation).

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

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

(Stirling+Taylor-Approximation)

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

Man beachte:

Die Stirling+Taylor Approximation

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

ist identisch mit der

Approximation durch Linearisieren von exp.

Die Größe der absoluten und relativen Fehler
in den beiden Approximationen (für $g = 365$ und variables n)
wird in dem über die Vorlesungsseite erhältlichen R-Programm
"Approximationen von w" illustriert.

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 ($n = 1, 2, ..$)

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

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

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

T:= Zeitpunkt der ersten Kollision

(ist ablesbar aus $(X_1, X_2, \ldots)$). 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.

Siehe dazu auch das R-Programm “Verteilung von T”
(Link über die Stofl-Seite)