

# Die Mathematik des Panini-Albums



Wie teuer ist/war es, das Panini-Sammelalbum zur WM 2026 voll zu bekommen?

**Die Mathematik und ihre Tücken hinter den Sammelbildchen...**

Prof. Dr. Peter Ruckdeschel  
Oldenburg, 03.09.2026

# Mein erstes Panini-Album – 1978



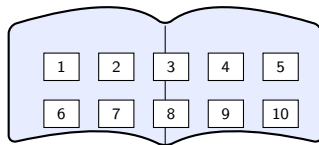
... habe ich auf dem Spielplatz liegen lassen!

Das war schlimmer als die Schmach von Cordoba / das 2:3 gegen Österreich!

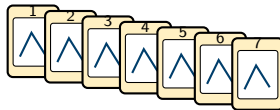
## Ausgangssituation: unser Panini-Modell

### Wir vereinfachen die Wirklichkeit

- ▶ Das Album enthält genau **980 verschiedene Bilder**.
- ▶ Jedes Bild ist bei jedem Kauf **gleich wahrscheinlich**.
- ▶ Die gezogenen Bilder sind **unabhängig** voneinander.
- ▶ **Tauschen ist nicht erlaubt**: Doppelte Bilder bleiben doppelt.
- ▶ klassisches Problem = **Coupon Collector**, [Baum & Billingsley (65)], [Erdős & Rényi (61)]



**980 Plätze**



1 Tütchen  
7 Bilder  
1,50 Euro

### Verkauf in Tütchen

- ▶ Ein Tütchen enthält **7 Bilder**.
- ▶ Ein Tütchen kostet **1,50 Euro**.

## Zwei Fragen, die wir beantworten wollen

(F1) Wann ist das Album voll?

Wie viele **Tütchen** muss man kaufen, bis alle 980 Bilder mindestens einmal vorhanden sind?

Wir bestimmen  
**die durchschnittlich nötige Zahl**  
der Tütchen

und eine Zahl, die  
**mit hoher Sicherheit** ausreicht.

(F2) Was kostet das?

Wie viel Geld müssen die Eltern ihren Kindern dafür geben?

Aus der Zahl der Tütchen wird  
**Geld: 1,50 Euro pro Tütchen**

wieder: **Durchschnitt**  
und **“mit hoher Sicherheit”**

### Wichtig

Ein fester Wert existiert nicht:

Manchmal hat man Glück, manchmal fehlt sehr lange genau ein einziges Bild.

## Wie wahrscheinlich ist der absolute Glücksfall?

Schon nach genau  $N = 980$  Bildern fertig heißt: *kein einziges Doppeltes*.

$$P(T_{980} = 980) = 1 \cdot \frac{979}{980} \cdot \frac{978}{980} \cdots \frac{1}{980} = \frac{980!}{980^{980}}.$$

### Das Ergebnis mit der Stirling-Formel (\*)

$$P(T_{980} = 980) \approx 1.93 \cdot 10^{-424}.$$

Praktisch ausgeschlossen! (\*: vgl. [Forster (04), §20, Satz 6])

### Wie klein ist das?

Im beobachtbaren Universum gibt es grob  $10^{80}$  Atome. Die fünfte Potenz davon ist etwa  $10^{400}$  – und selbst deren Kehrwert ist noch viel größer als unsere Wahrscheinlichkeit.

# Geometrische Wartezeiten: “Mensch-Ärgere-Dich-Nicht”

[Georgii (02), Abschn. 2.5.1, Bsp. (4.6)]

## Wann komme ich aufs Spielfeld?

Für eine Sechs gilt  $p = 1/6$ . Sei  $W$  die Zahl der Würfe bis zur ersten Sechs:

$$W \sim \text{Geom}(1/6) \iff$$

$$\mathbb{P}(W > k) = (1 - p)^k$$

Nach einem Fehlwurf stehen wir wieder vor demselben Problem  $\rightsquigarrow W' \sim \text{Geom}(1/6)$ .

## Der Rekursionstrick

Für  $W \sim \text{Geom}(p)$ :

$$W = \begin{cases} 1, & \text{mit Wkt. } p, \\ 1 + W', & \text{mit Wkt. } 1 - p. \end{cases}$$

Daher

$$\mu = p + (1 - p)(1 + \mu) \implies \boxed{\mathbb{E}[W] = 1/p}.$$

## Dieser Trick funktioniert auch für die Varianz

$$\mathbb{E}[W^2] = p + (1 - p)\mathbb{E}[(1 + W')^2] \implies \boxed{\text{Var}(W) = \frac{1 - p}{p^2}}.$$

## Beim “Mensch-Ärgere-Dich-Nicht”:

Für die erste Sechs:  $\mathbb{E}[W] = 6$  Würfe mit Varianz  $\text{Var}[W] = 30$ .

## Eine Abschätzung nach oben – und ein Quiz

Beim Panini-Album gilt für jeden festen Sticker (z.B. **Deniz Undav**) Wartezeit

$$p = \frac{1}{N}.$$

Für diese Ziehungswahrscheinlichkeit erhalten wir also die erwartete Wartezeit  $1/p = N$ ; wir haben  $N$  verschiedene Bilder, also

$$\mathbb{E}[T_N] \leq N \cdot N = N^2.$$

<https://1.uol.de/2026>

Für  $N = 980$ :

$$\mathbb{E}[T_{980}] \leq 960\,400 \text{ Bilder.}$$



Warum ist das sicher viel zu pessimistisch?

Die Trefferchance  $1/N$  gilt nur ganz am Ende; am Anfang ist fast jedes Bild neu.

**QUIZ:** Wie groß ist der wahre Erwartungswert?

**A** ca. 2 000

**B** ca. 5 000

**C** ca. 7 000

**D** ca. 960 000

# Einschluss–Ausschluss–Prinzip: — „zu wenig“ – „zu viel“

[Georgii (02), Aufg. 1.6]

## Fehl-Ereignisse

Für jede Stickerart  $i$  sei

$$A_i = \{\text{Sticker } i \text{ fehlt komplett}\}.$$

Vollständig ist das Album genau dann,  
wenn

kein  $A_i$  eintritt.

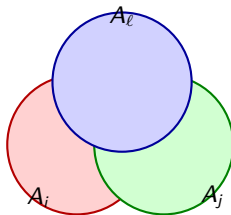
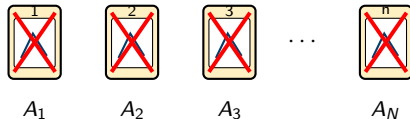
Also zählen wir das Komplement:

$$(A_1 \cup \dots \cup A_N)^c.$$

## Merksatz

Erst zu viel abziehen, dann Überlappungen  
wieder korrigieren.

## Einzelne fehlende Sticker



Überlappungen: zwei oder drei Sticker fehlen

# Einschluss–Ausschluss liefert die Formel

## Schrittweise korrigieren = die SIEBFORMEL

$$\begin{aligned} \#\{\text{vollständige Alben nach } m \text{ gekauften Stickern}\} &= N^m - \sum_i |A_i| + \sum_{i < j} |A_i \cap A_j| \\ &\quad - \sum_{i < j < \ell} |A_i \cap A_j \cap A_\ell| + \dots \end{aligned}$$

Wenn genau  $k$  bestimmte Sticker fehlen, bleiben nur noch  $N - k$  "erlaubte" Sticker. Also:

$$|A_{i_1} \cap \dots \cap A_{i_k}| = (N - k)^m.$$

Für die Auswahl dieser  $k$  fehlenden Sticker gibt es  $\binom{N}{k}$  Möglichkeiten.

Erinnerung:

$$\binom{N}{k} = \frac{N \cdot (N - 1) \cdot \dots \cdot (N - k + 2) \cdot (N - k + 1)}{k \cdot (k - 1) \cdot \dots \cdot 2 \cdot 1}$$

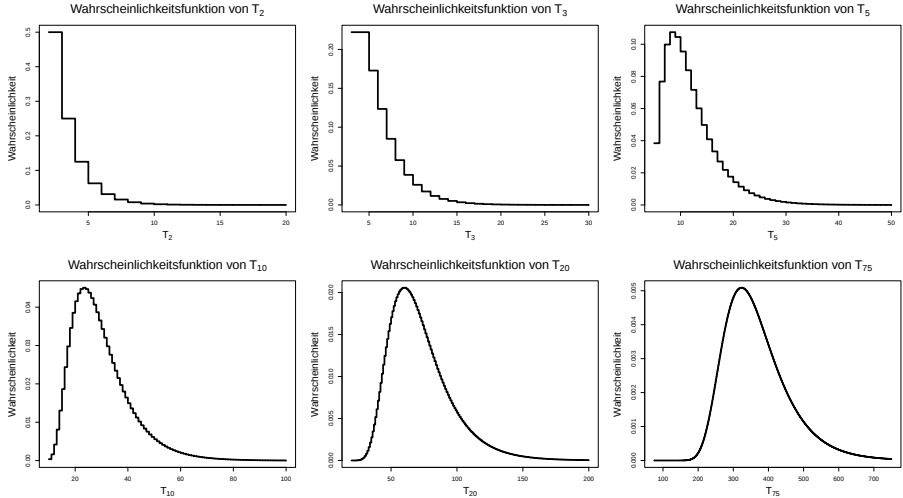
## Endformel

$$\#\text{vollständige Folgen} = \sum_{k=0}^N (-1)^k \binom{N}{k} (N - k)^m$$

$$\begin{aligned} \mathbb{P}(\text{Album voll nach } m \text{ Käufen}) &= \frac{\#\text{vollständige Folgen}}{N^m} \\ &= \sum_{k=0}^N (-1)^k \binom{N}{k} \left(1 - \frac{k}{N}\right)^m \end{aligned}$$

Das Vorzeichen wechselt:  $+, -, +, -, \dots$  damit jede Überlappung genau einmal zählt.

# Visualisierung der Wartezeitverteilung von $T_N$



... und was passiert für  $N > 75$ , insbesondere für  $N = 980$ ?

## Aus der Siebformel zum Erwartungswert

### Wir betrachten die Wartezeit

Sei  $T_N$  die Zahl der gekauften **Einzelbilder**, bis das Album mit  $N$  Bildern komplett ist. Dann gilt

$$\mathbb{E}[T_N] = \sum_{j=0}^{\infty} j \mathbb{P}(T_N = j) = \sum_{j,m=0}^{\infty} \mathbb{I}(0 \leq m < j) \mathbb{P}(T_N = j) = \sum_{m=0}^{\infty} \mathbb{P}(T_N > m).$$

Für  $m < N$  ist das Album sicher noch nicht vollständig und  $\mathbb{P}(T_N > m) = 1$ , also:

$$\sum_{m=0}^{N-1} \mathbb{P}(T_N > m) = N.$$

Für  $m \geq N$  haben wir die Formel auf der vorigen Folie gehabt:

$$\begin{aligned} \mathbb{P}(T_N \leq m) &= \sum_{k=0}^N (-1)^k \binom{N}{k} \left(1 - \frac{k}{N}\right)^m = \\ &= 1 - \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N}{k} \left(1 - \frac{k}{N}\right)^m. \end{aligned}$$

Also gilt für  $m \geq N$

$$\mathbb{P}(T_N > m) = \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N}{k} \left(1 - \frac{k}{N}\right)^m.$$

Aber für  $0 < q < 1$  gilt

$$\sum_{m=N}^{\infty} q^m = q^N / (1 - q),$$

sodass Summieren des Terms  $\sum_{m=N}^{\infty} \left(1 - \frac{k}{N}\right)^m$  liefert

### Erwartungswert mit Siebformel

$$\mathbb{E}[T_N] = N \left( 1 + \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N}{k} \frac{(1 - k/N)^N}{k} \right).$$

Direkt umgesetzt in  [R Core Team (25)]:

klappt erst einmal sehr gut, aber ...

### -Code: Siebformel

```
Siebformel.ewr <- function(n){
  k <- 1:(n-1)
  erg <- (-1)^(k-1) * choose(n,k) *
    (1-k/n)^n / k
  return(n * (sum(erg) + 1))
}
```

| N   | Siebformel     | wahrer Wert                |
|-----|----------------|----------------------------|
| 2   | 3              | 3                          |
| 3   | 5.5            | 5.5                        |
| 5   | 11.4167        | 11.4167                    |
| 10  | 29.2897        | 29.2897                    |
| 20  | 71.9548        | 71.9548                    |
| 50  | 224.9603       | 224.9603                   |
| 75  | 367.6017       | 367.6017                   |
| 100 | 518.7352       | 518.7378                   |
| 150 | -1719.871      | 838.6771                   |
| 200 | -37290172230   | 1175.606                   |
| 640 | 1.290374 e+63  | 4505.257                   |
| 980 | 9.127753 e+102 | wird (noch) nicht verraten |

### Interpretation

Das ist genau die gerade hergeleitete alternierende Summe.

### Beobachtung

Bis etwa  $N = 75$  sieht alles perfekt aus; danach wird es numerisch instabil.

⇒ Wie kommt das zustande?

# Was passiert numerisch? AUSLÖSCHUNGSEFFEKTE

## Mathematisch ist die Formel korrekt

Die einzelnen Summanden

$$(-1)^{k-1} \binom{n}{k} \frac{(1 - k/n)^n}{k}$$

werden für große  $n$  teilweise riesig im Betrag.  
Danach sollen sie sich fast vollständig wegekürzen.

## Computerproblem

Der Rechner speichert nur endlich viele Dezimalstellen. Beim Subtrahieren riesiger fast gleicher Zahlen gehen die entscheidenden Stellen verloren.

Diesen Effekt nennt man **AUSLÖSCHUNG**.

sehr große positive Beiträge



wahres Ergebnis  
 $\mathbb{E}[T_{200}] \approx 1176$



sehr große negative Beiträge

fast alles kürzt sich weg:  
übrig bleibt eine kleine Zahl

## Nächster Schritt

Äquivalente Ansätze / Formeln ohne katastrophale Auslöschung.

# 1. Ausweg: Simulation in – I: die Idee

Statt zu rechnen: viele Alben auf dem Rechner nachspielen/simulieren

Wir lassen den Computer sehr oft ein komplettes Album simulieren und messen jedes Mal:

$T_{980}$  = Wie viele Bilder wurden bis zur Vollständigkeit gebraucht?

**FAKT** (*Gesetz der großen Zahlen* [Georgii (02), Satz 5.7] bzw. [Bauer (02), Satz 12.1]):

Sei  $iter$  die Zahl der simulierten Alben. Dann gilt: (i) Für  $iter \rightarrow \infty$  konvergiert der Mittelwert  $\bar{T}_{iter,980}$  über die  $iter$  gemessenen Werte  $T_{980}$  (*fast sicher und in  $L_2$* ) gegen  $\mathbb{E} T_{980}$ . (ii)  $sd(\bar{T}_{iter,980}) = sd(T_{980})/\sqrt{iter}$ .

## Grundgerüst in :

```
N <- 980
set.seed(20260823)
iter <- 10000

counter <- rep(N, iter)
x <- 1:N
InzMatrix <- matrix(0, N, iter)
```

## Was wird gespeichert?

- ▶ `counter[i]`: benötigte Bilder für Album  $i$
- ▶ `InzMatrix[,i]`: wie oft jedes der 980 Bilder im fertigen Album  $i$  vorkommt
- ▶ `set.seed(...)`: macht die Simulation reproduzierbar

## Vorteil

Man bekommt nicht nur Information über die Verteilung von  $T_{980}$ , sondern auch Zusatzinformationen über den Zustand des fertigen Albums.

# 1. Ausweg: Simulation in – II: viele Alben simulieren

## Der eigentliche Simulationsschritt in :

```
for(i in 1:iter){  
  
  InzVektor <- numeric(N)  
  counter[i] <- N  
  
  # erste N Bilder auf einmal ziehen  
  y0 <- sample(x, N, replace = TRUE)  
  yt <- table(y0)  
  
  # erste N Bilder einsortieren  
  for(j in 1:length(yt)){  
    TabEintrag <- yt[j]  
    PositionInAlbum <- as.numeric(names(TabEintrag))  
    InzVektor[PositionInAlbum] <- TabEintrag  
  }  
  
  # danach einzeln weiterziehen  
  while(sum(InzVektor > 0) < N){  
    counter[i] <- counter[i] + 1  
    y0 <- sample(x, 1, replace = TRUE)  
    InzVektor[y0] <- InzVektor[y0] + 1  
  }  
  
  InzMatrix[,i] <- InzVektor  
}
```

## Warum erst $N$ Bilder auf einmal?

Mindestens  $N = 980$  Bilder braucht man sowieso.

Die ersten 980 Ziehungen können daher als Block simuliert werden; danach wird einzeln weitergezogen, bis wirklich alle Bilder mindestens einmal da sind.

## Nachteil

Die Schleife ist zwar einfach zu verstehen, aber für sehr viele Wiederholungen nicht besonders schnell.

Genauer wächst die Präzision von  $\bar{T}_{\text{iter};980}$  nur mit Rate  $1/\sqrt{\text{iter}}$ , also "kostet" eine weitere genaue Dezimalstelle für  $\mathbb{E} T_{980}$  einen um Faktor 100 größeren Simulationsumfang  $\text{iter}$ .

## 1. Ausweg: Simulation in – III: Ergebnisse und Einordnung

### 10 000 simulierte Alben

- ▶ Rechenzeit: ca. 15 Minuten
- ▶ Mittelwert / Schätzung  $\bar{T}_{\text{iter};980}$  für  $\mathbb{E} T_{980}$ :  
7 324.684
- ▶ 95%-Intervall für  $\mathbb{E} T_{980}$ :  
(7 300.247, 7 349.122)
- ▶ Minimum / Maximum:  
4 722, 15 771
- ▶ Empirische Standardabweichung:  
1 246.85

### Zusatzinformation über fertige Alben

Zum Beispiel: Maximale Häufigkeit  $M_{980}$  eines einzelnen Bildes im fertigen Album:

$$\mathbb{E}(M_{980}) \approx 17.73.$$

95%-Intervall für  $\mathbb{E} M_{980}$

$$(17.68, 17.77).$$

Minimum / Maximum:

$$13, 32.$$

Korrelation mit der Sammelzeit:

$$\text{Corr}(T_{980}, M_{980}) \approx 0.831.$$

### Fazit

Simulation ist anschaulich und flexibel, aber langsam, wenn man eine vorgegebene Genauigkeit garantieren möchte.

# 1. Ausweg: Simulation in – IV: „Glückskind“

Ergebnis-Album des Durchlaufs mit minimalem  $T_{980;\min} = 4722$

```

3 3 8 6 5 6 8 5 9 2 6 2 4 4 4 6 5 2 4 6 8 5 2 6 6 6 4 10 5 6 4 1 13 3 5
6 8 3 4 6 8 4 5 6 4 5 6 9 5 5 2 3 3 3 5 5 5 4 4 5 2 3 1 3 5 3 9 6 3 8
5 3 3 7 8 2 7 2 6 3 3 3 4 1 4 3 3 2 3 4 6 2 5 8 3 5 5 8 3 5 7 2 2 3 6
7 4 4 11 5 3 5 4 8 5 4 2 7 7 7 2 4 8 3 3 2 4 7 4 8 8 4 5 3 7 9 7 7 3 10 8
4 4 4 6 6 6 4 5 4 2 4 5 4 3 4 3 3 4 7 1 9 7 5 4 9 7 3 5 2 7 4 4 6 3 7
6 7 8 9 4 5 9 6 5 2 2 1 3 6 7 3 6 4 3 4 4 3 3 8 1 1 2 5 8 5 2 5 1 7 4
5 6 7 5 1 7 4 8 4 6 6 5 5 3 3 9 3 1 5 4 5 4 6 5 4 4 3 4 4 5 6 5 4 9 3
3 3 5 4 5 3 3 6 3 10 4 2 8 4 2 2 5 6 3 7 9 2 10 12 6 7 4 5 9 3 3 4 6 4 7
4 5 6 6 8 2 4 3 4 4 4 5 3 7 7 5 4 4 2 1 6 8 3 5 6 5 4 4 10 6 5 5 4 3 10
6 2 2 5 7 5 6 5 4 2 3 6 6 4 6 6 5 6 1 7 6 4 5 3 4 2 2 7 3 5 3 6 4 3 3
6 6 4 5 2 6 5 5 4 5 4 6 7 3 3 5 7 6 6 8 6 5 3 5 5 5 7 8 5 2 4 8 6 5
2 4 5 6 6 4 3 8 5 4 2 7 3 5 6 3 5 3 7 6 2 3 7 3 1 3 7 2 2 6 3 5 1 5
6 7 4 6 7 7 8 3 5 4 4 2 8 5 2 1 7 7 5 2 3 6 2 6 7 4 5 5 2 5 1 12 2 4 1
2 4 3 5 1 7 5 3 2 5 4 3 2 3 4 4 6 6 4 4 3 8 6 5 4 7 5 5 3 8 1 3 2 5 1
4 4 3 4 3 5 4 7 5 6 4 5 6 6 4 3 2 5 1 5 8 3 10 6 4 6 5 3 4 5 5 3 9 2 5
1 1 6 10 2 7 5 6 8 6 10 5 7 6 3 1 6 7 4 7 3 3 6 3 9 5 2 3 4 4 3 5 6 4 8
7 5 7 3 4 7 4 5 5 6 6 7 3 5 6 7 5 5 4 4 6 5 5 2 5 4 5 5 4 6 3 4 7 5 8
6 6 3 5 4 4 1 10 9 3 4 1 6 3 6 10 6 3 7 3 3 2 5 6 7 3 6 3 1 4 2 7 4 7 6
2 6 4 8 5 3 7 5 6 5 7 8 4 3 1 2 1 4 5 12 1 7 6 5 5 4 6 6 7 6 4 8 4 5 6
7 6 10 6 8 3 3 3 5 5 2 2 6 6 3 10 7 8 1 2 3 6 4 3 5 4 4 6 4 2 5 8 4 6 1
5 8 7 4 5 8 3 8 3 4 4 6 3 5 6 2 8 3 4 8 5 6 2 1 4 2 2 11 7 4 2 8 6 7 5
4 7 4 2 5 7 11 5 8 3 7 3 2 2 3 8 6 4 3 5 4 8 3 4 9 3 6 6 4 6 5 4 4 3 4
1 5 3 7 4 5 7 2 4 4 6 3 4 5 5 4 1 8 7 5 4 8 6 6 3 1 5 5 5 4 4 3 2 7 2
3 7 4 2 6 7 13 6 5 5 5 2 3 3 6 7 1 4 5 3 5 8 2 1 4 4 1 3 5 3 6 1 3 6
8 8 4 5 5 4 6 8 4 7 5 7 6 10 6 8 3 8 8 3 7 7 4 4 3 4 3 6 3 5 6 5 3 5 6
5 2 2 2 3 4 6 4 6 6 5 2 5 6 5 6 6 1 4 4 5 6 7 4 7 1 6 6 7 4 3 3 7 4 3
4 10 4 2 8 6 4 4 4 5 5 8 5 2 4 6 6 3 6 6 8 3 5 6 6 1 4 3 4 2 7 2 5 6 6
4 5 11 2 4 4 6 4 4 5 4 9 10 3 2 6 9 7 8 4 6 7 7 5 4 8 5 2 2 4 7 3 5 5 7

```

# 1. Ausweg: Simulation in – V: „Pechvogel“

Ergebnis-Album des Durchlaufs mit maximalem  $T_{980;\max} = 15\,771$

14 12 20 19 21 9 17 11 21 13 23 11 13 18 1 19 25 15 17 14 17 19 17 15 19 9 13 11 13 19 7 19 17 10 12  
16 16 16 18 10 13 19 20 12 20 22 14 21 8 14 16 14 16 18 15 23 20 23 14 18 14 18 12 26 15 15 12 6 22 19  
16 20 14 17 21 28 20 21 16 22 12 16 21 20 11 10 14 13 18 12 25 14 18 13 15 14 14 16 23 19 16 15 18 18 10  
17 12 9 10 18 22 19 19 11 20 14 18 18 13 20 14 21 26 13 15 18 19 20 16 16 9 20 14 13 14 24 7 15 13 14  
21 15 18 19 15 20 11 9 11 13 13 18 12 17 18 7 30 13 15 24 19 20 13 19 22 16 16 14 14 13 18 11 22 19 14  
15 17 15 17 18 13 18 14 15 17 19 14 20 19 18 15 14 14 15 17 20 9 19 12 18 11 14 17 16 18 10 14 19 15 12  
15 29 9 12 21 18 11 24 19 9 17 14 16 11 19 21 21 14 15 17 15 13 16 11 14 15 25 12 14 18 13 18 17 18 17  
22 24 20 14 15 17 16 12 21 19 19 14 18 19 12 17 18 15 21 11 19 11 12 17 17 18 22 20 12 19 10 21 17 12 17  
21 24 10 16 16 23 13 18 19 15 14 15 13 12 18 12 16 12 15 14 18 13 12 13 21 12 20 16 22 17 12 18 20 17 19  
19 16 15 14 13 13 9 12 18 17 21 20 16 11 19 8 25 17 15 18 9 16 12 8 13 15 21 14 13 20 20 13 14 14 12  
19 15 17 19 13 19 18 13 19 14 15 19 22 16 11 19 11 18 17 14 21 20 19 17 17 12 16 18 17 20 17 15 13 21 7  
13 23 15 18 18 22 15 15 19 10 14 12 20 17 13 12 14 21 16 20 23 17 23 13 15 14 19 16 20 12 17 13 13 19 16  
19 18 14 21 17 17 14 12 14 25 8 14 16 19 12 13 15 16 15 14 19 13 20 14 17 22 12 18 11 11 14 16 13 13 12  
14 18 21 13 11 21 14 18 12 14 24 15 16 12 12 16 9 17 19 25 18 21 12 18 13 13 16 17 20 14 26 16 19 17 20  
14 12 15 10 15 15 21 13 10 16 15 18 15 13 20 15 13 8 17 17 9 11 14 16 21 19 7 13 11 19 24 15 19 23 13  
17 19 16 16 16 13 16 17 17 15 18 13 15 23 15 12 13 20 14 10 15 14 15 12 16 16 15 17 21 19 16 15 15 19  
19 19 19 15 21 15 14 13 13 18 18 19 14 11 12 16 13 22 19 11 19 18 21 13 19 17 22 20 17 15 13 9 16 21 14  
12 20 16 15 20 12 14 21 16 11 15 18 9 19 12 11 21 15 10 11 8 15 18 20 14 12 16 12 22 15 21 21 14 14 17  
21 14 14 12 13 11 13 25 22 16 25 18 15 13 19 17 17 25 21 20 11 13 14 11 18 9 18 13 13 19 16 12 17 7  
16 22 10 20 23 12 10 16 13 13 16 17 8 23 19 13 20 22 17 11 21 23 18 20 13 16 4 21 12 17 12 10 18 18 12  
13 16 12 20 17 15 19 16 17 18 17 11 13 13 17 22 13 21 17 14 19 19 16 14 15 17 20 16 17 15 18 13 18 16 18  
18 14 18 19 15 13 16 15 18 19 11 13 14 12 15 17 17 18 18 19 15 14 14 16 14 16 15 22 16 15 16 23 15 17 16  
19 12 13 4 14 17 15 19 26 14 16 16 15 20 22 14 20 17 16 20 13 18 12 15 10 13 21 18 12 11 14 19 22 20 16  
18 15 23 22 19 14 12 14 19 18 25 12 21 22 10 15 23 23 14 29 21 19 10 23 11 13 19 15 14 13 15 15 16 14  
12 16 21 14 19 23 23 15 22 17 10 13 20 13 9 18 14 25 19 11 18 11 21 16 21 14 25 20 27 13 14 10 19 8 18  
21 16 12 15 21 23 12 13 16 18 18 9 13 18 11 17 13 21 19 15 18 13 9 18 18 13 14 16 19 14 10 17 11 13 21  
11 21 19 14 12 19 14 16 10 13 22 15 18 14 16 15 16 18 22 19 15 16 13 20 22 13 17 14 18 12 17 16 19 21 20  
16 12 14 16 9 13 14 24 19 12 13 19 17 15 17 19 18 16 15 12 13 21 16 15 10 14 27 11 16 13 11 27 15 17 21

## 2. Ausweg: Der clevere Blick: Nur die Anzahl zählt [Georgii (02), Aufg. 4.19]

Statt uns zu merken, **welche** Bilder schon da sind, merken wir uns nur

$i = \text{Anzahl}$  verschiedener Bilder, die wir schon haben.

Wenn schon  $i$  verschiedene Bilder da sind ...

... dann fehlen noch  $N - i$  verschiedene Bilder.  
Beim nächsten Bild ist daher

$$p_i = \mathbb{P}(\text{neues Bild}) = \frac{N - i}{N} = 1 - \frac{i}{N}.$$

Wartezeit auf das nächste neue Bild

Sei  $W_i$  die Zahl der weiteren Bilder, bis ein neues kommt. Dann

$$W_i \sim \text{Geom}\left(1 - \frac{i}{N}\right), \quad i = 0, \dots, N - 1.$$

und  $W_i$  stochastisch unabhängig.

Damit zerfällt die gesamte Sammelzeit in einfache Wartezeiten:

$$T_N = W_0 + W_1 + \dots + W_{N-1}.$$

Idee:

Am Anfang ist jedes Bild neu; ganz am Ende beträgt die Trefferchance nur noch  $1/N$ ; das ist leicht zu simulieren, aber man kann die Verteilung damit in  auch exakt berechnen.

## Numerisch exakt: Faltung der Wartezeiten in

### -Code mit distrEx (\*)

```
library(distrEx) # Rechnen mit Verteilungen

N <- 980

# R zählt bei Geom(p) nur Fehlversuche.
# Beim ersten Zug haben wir sicher Wartezeit N:
XN <- as(Dirac(N), "LatticeDistribution")
# "LatticeDistribution": Die Verteilung hat
# Massepunkte auf Gitter {0,1,2,...}

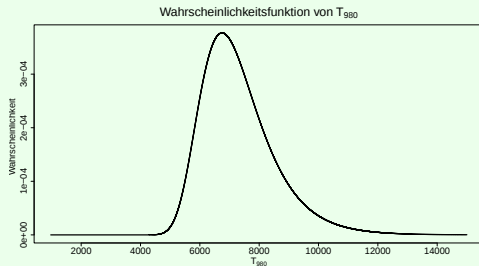
for(i in 1:(N-1)) {
  # Y = W_{N-i}-1 ^= Fehlversuche in W_{N-i}
  Y <- as(Geom(i/N), "LatticeDistribution")
  XN <- XN + Y
}

## XN ist nun die Verteilung von T_{980}
(*: siehe [Kohl & R. (25), R. et al. (06)])
```

### Idee

Die Verteilung von  $T_{980}$  entsteht als Verteilung der Summe der geometrischen Wartezeiten. Das nennt man **FALTUNG**. Damit erhält man die Wahrscheinlichkeitsfunktion numerisch sehr genau.

### Visualisierung von $T_{980}$



### Nutzen

Man bekommt nicht nur Mittelwert und Varianz, sondern die ganze Verteilung: z.B. Quantile und exakte obere 95%-Grenzen.

# Auflösung des Quiz': der wahre Erwartungswert (analytisch / ohne )

Für die Wartezeit  $W_i \sim \text{Geom}(p_i)$  ist

$$p_i = 1 - \frac{i}{N} = \frac{N-i}{N}.$$

Mit den Formeln für Erwartungswert und Varianz für  $\text{Geom}(p)$  erhalten wir

$$\mathbb{E}[W_i] = \frac{N}{N-i}, \quad \text{Var}(W_i) = \frac{iN}{(N-i)^2}.$$

## Erwartungswert

$$\mathbb{E}[T_N] = \sum_{i=0}^{N-1} \mathbb{E}[W_i] = \sum_{i=0}^{N-1} \frac{N}{N-i} = \sum_{j=1}^N \frac{N}{j}.$$

Sei  $H_n$  der Beginn der **harmonischen Reihe**, also

$$H_n := \sum_{j=1}^n 1/j \approx \log(n) + 0,5772, \quad [\text{Cramér (46), p.127}]$$

so ist

$$\mathbb{E}[T_N] = N H_N$$

## Varianz

Die Wartezeiten  $W_i$  sind unabhängig. Also

$$\text{Var}[T_N] = \sum_{i=0}^{N-1} \text{Var}[W_i] = \sum_{i=0}^{N-1} \frac{iN}{(N-i)^2} = N^2 \sum_{j=1}^N \frac{1}{j^2} - N H_N.$$

Also ist mit  $H_n^{(2)} = \sum_{j=1}^n 1/j^2 \approx \frac{\pi^2}{6}$  [Forster (04), (21.8)]

$$\text{Var}[T_N] = N^2 H_N^{(2)} - N H_N$$

## Auflösung des Quiz': der wahre Erwartungswert in zwei Formeln

### Zum Vergleich

Wir haben also hergeleitet, dass

$$\begin{aligned}\mathbb{E}[T_N] &= N \sum_{j=1}^N 1/j \\ &= N \left( 1 + \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N}{k} \frac{(1 - k/N)^N}{k} \right)\end{aligned}$$

Mathematisch sind beide Formeln äquivalent; die neue Formel ist aber viel einfacher und darüber hinaus numerisch stabil.

### Auflösung des Quiz'

Für  $N = 980$ :

$$\mathbb{E}[T_{980}] \approx 7.315,97$$

, also Antwort

**C**

, mit  $\text{sd}(T_{980}) \approx 1.253,60$ .

## Wie sicher ist unsere Antwort?

Aus der vorigen Folie kennen wir Erwartungswert und Varianz für  $N = 980$ :

$$\mu = \mathbb{E}[T_{980}] \approx 7.315,97, \quad \sigma = \text{sd}(T_{980}) \approx 1.253,60.$$

Ziel ist nun eine (möglichst kleine) Zahl  $m$  als **Garantieschranke**, sodass  $\mathbb{P}(T_{980} > m) \leq 5\%$  ist.

### Chebyshev für Garantieschranken

Sei  $Z$  eine positive Zufallsvariable,  $c > 0$  und  $g$  eine monoton wachsende Funktion mit  $g(c) > 0$ . Dann gilt

$$\begin{aligned} \mathbb{P}(Z \geq c) &= \mathbb{E}[\mathbb{I}(Z \geq c)] \leq \mathbb{E}[\mathbb{I}(g(Z) \geq g(c))] \\ &= \mathbb{E}[\mathbb{I}(\frac{g(Z)}{g(c)} \geq 1)] \leq \mathbb{E}[\frac{g(Z)}{g(c)} \mathbb{I}(\frac{g(Z)}{g(c)} \geq 1)] \\ &\leq \mathbb{E}[g(Z)]/g(c) \end{aligned}$$

Für  $Z = |T_{980} - \mu|$  und  $g(x) = x^2$  erhalten wir die **Chebyshev-Ungleichung**, also

$$\mathbb{P}(|T_{980} - \mu| \geq a\sigma) \leq \frac{\text{Var } T_{980}}{\sigma^2 a^2} = \frac{1}{a^2}.$$

### Alternative obere Garantieschranken

- ▶ **Chebyshev:** immer gültig, aber vorsichtig.
- ▶ **Normalapproximation:** schärfer, denn statt Faktor  $\sqrt{20} \approx 4.47$  erhält man  $\Phi^{-1}(.975) \approx 1.96$ ; aber nur Näherung (und man muss Voraussetzungen für Näherung prüfen [Bauer (02), Satz 28.3]).
- ▶ **Exakt:** numerisch als 95%-Stelle der Verteilungsfunktion von  $T_{980} = W_0 + \dots + W_{979}$ , (in  mit `distrEx`).

### Obere Chebyshev-95%-Schranke für $T_{980}$

Wählen wir  $\frac{1}{a^2} = 0,05 \iff a = \sqrt{20}$ , so erhalten wir mit der Chebyshev-Ungleichung als obere 95%-Garantieschranke

$$\mu + \sqrt{20} \sigma \approx 12.923.$$

## Antworten auf unsere beiden Fragen

|                                          | $E(\cdot)$        | sd       | obere 95%-Garantieschranke |                |                | min                                 | max                                           |
|------------------------------------------|-------------------|----------|----------------------------|----------------|----------------|-------------------------------------|-----------------------------------------------|
|                                          |                   |          | Chebyshev                  | Normalapprox.  | exakt          |                                     |                                               |
| $T_{980}$ Bilder<br>$P(T_{980} < \cdot)$ | 7.315,97<br>57,1% | 1.253,59 | 12.923<br>99,8%            | 9.378<br>93,1% | 9.656<br>95,0% | 980<br>$\approx \varepsilon_1$ 100% | 960.400<br>$\approx (1 - \varepsilon_2)$ 100% |
| Tütchen                                  | 1.045             | 179      | 1.846                      | 1.339          | 1.379          | 140                                 | 137.200                                       |
| Preis                                    | 1.567,50 €        | 269,50 € | 2.769,00 €                 | 2.008,50 €     | 2.068,50 €     | 210,00 €                            | 205.800,00 €                                  |

für  $\varepsilon_1 = 1.9 \cdot 10^{-422}$  (mit [Forster (04), §20, Satz 6]),  $\varepsilon_2 = 2.4 \cdot 10^{-423}$  (mit [Baum & Billingsley (65), Thm. 4])

### Antwort auf (F1)

Im Mittel braucht man etwa **1.045** Tütchen.

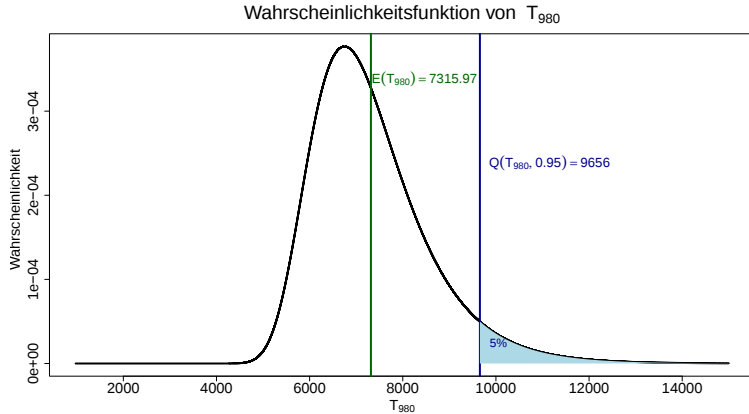
Eine realistische obere 95%-Grenze liegt bei etwa **1.379** Tütchen.

### Antwort auf (F2)

Das volle Album kostet im Mittel **1.567,50 €**.

Mit 95% Sicherheit reichen nach distrEx **2.068,50 €**.

# Visualisierung der Wartezeit $T_{980}$ mit Erwartungswert und Garantieschranke



weitere Kenngrößen (abgelesen aus  mit `distrEx`):

$\text{Median}(T_{980}) = 7\,109$ ,  $\text{Modus}(T_{980}) = 6\,749$ ,  $Q(T_{980}, 0.25) = 6\,432$ ,  $Q(T_{980}, 0.75) = 7\,969$ ,  $Q(T_{980}, 0.99) = 11\,253$ ,

$Q(T_{980}, 0.999) = 13\,510$ ,  $\text{IQR}(T_{980}) = 1\,537$ ,  $\text{mad}(T_{980}) = 749$ ,  $\text{Schiefe}(T_{980}) = 1.1439$ ,  $\text{Kurtosis}(T_{980}) = 2.3929$

Die Mathematik des Panini-Albums — Prof. Dr. Peter Ruckdeschel

## Ausblick: Was bringt Tauschen?

Bisher galt die harte Regel:

Kein Tauschen!

Beim echten Panini-Sammeln kann man doppelte Bilder aber mit anderen Sammlern tauschen.

### Achtung: Das ist ein neues Modell

Man muss genauer festlegen, *wie* getauscht wird: Wie viele Tauschpartner gibt es? Wie oft trifft man sie? Welche Bilder haben sie? Das modellieren wir heute nicht im Detail.

Wichtige Botschaft aber: Durch Tauschen wird das Sammeln *asymptotisch deutlich schneller!*  
Statt einer Größenordnung

$$N \log N$$

benötigt man bei  $m$  poolenden Sammlern unter geeigneten Tauschmodellen nur noch ungefähr

$$N \left( (\log N)/m + \log \log N \right)$$

Bilder pro Sammler, vgl. [Newman & Shepp (60), Thm. 2].

### Für unser Album mit $N = 980$

Grob angesetzt ergibt das für  $m = 30$  im Erwartungswert pro Sammler etwa eine Reduktion um 70% bzgl. der Situation ohne Tauschen, genauer circa

$$2200 \text{ Bilder} \iff 315 \text{ Tütchen} \iff \boxed{\text{ca. 475 €}}.$$

**BOTSCHAFT:** Tauschen hilft enorm – aber billig wird das Album trotzdem nicht.

# Ausblick: Mehr als Panini

## Die allgemeinere Frage

Wir beobachten ein zunächst unbekanntes Terrain mit vielen verschiedenen Phänomenen, Arten, Fehlern, Ereignistypen, ...

**Frage aus dem Active Learning:**

**Wie lange müssen wir weiter beobachten, bis wir glauben können,  
dass wir das Terrain hinreichend gut erforscht haben?**

## Die Panini-Idee

Beim Album haben wir beobachtet:

Je mehr Bilder bekannt sind, desto länger wartet man auf ein neues.

Die Wartezeit auf das *nächste noch unbekannte* Bild enthält also Information darüber, wie viel noch unentdeckt sein könnte.

## Im unbekannten Terrain

Dasselbe Prinzip kann man nutzen:

- ▶ neue Phänomene fortlaufend registrieren,
- ▶ Wartezeiten zwischen Neuentdeckungen beobachten,
- ▶ daraus abschätzen, wie viele Phänomene noch fehlen könnten,
- ▶ und eine obere Schranke für die nötige weitere Beobachtungszeit bestimmen.

## Eine zusätzliche Schwierigkeit

Beim Panini-Album kannten wir die Zahl der Bilder  $N = 980$ . In einem unbekannten Terrain kennen wir die Gesamtzahl  $N$  gerade *nicht*:

Wir müssen  $N$  parallel aus den Beobachtungen mitschätzen.

**Coupon Collector wird damit zu einer Methode für die Frage:  
„Wann haben wir wahrscheinlich genug gesehen?“**

## Was haben wir gelernt?

- ▶ Hinter einem Panini-Album steckt überraschend viel **Mathematik** – schon wenn wir das Tauschen zunächst weglassen.
- ▶ Ein Album **ohne Dubletten** zu füllen, ist praktisch unmöglich.
- ▶ **Einschluss–Ausschluss**, **Simulation** und **geometrische Wartezeiten** sind drei verschiedene Wege, das Sammelproblem mathematisch zu untersuchen.
- ▶ Der clevere Blick auf die **Wartezeit bis zum nächsten neuen Bild** führt zu

$$E[T_N] = NH_N \approx N \log N.$$

- ▶ Für unser Album: **knapp 1.600 Euro im Mittel** und mit 95 % Sicherheit **nicht mehr als etwa 2.100 Euro**. Tauschen in einer ganzen Klasse senkt die Kosten drastisch auf grob **500 Euro**.
- ▶ Die Idee geht über Panini hinaus: **Wie lange müssen wir beobachten, bis kaum noch Neues kommt?** Das ist auch beim Erforschen unbekannter Phänomene eine wichtige Frage.

## Literatur

- [1] BAUER, H. (2002). *Wahrscheinlichkeitstheorie*. 5., durchgesehene Auflage. Walter de Gruyter.
- [2] BAUM, L.E. UND BILLINGSLEY, P. (1965): Asymptotic Distributions for the Coupon Collector's Problem. *The Annals of Mathematical Statistics*. **36**(6): 1835–1839.
- [3] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- [4] ERDŐS, P. UND RÉNYI, A. (1961): On a classical problem of probability theory, *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **6**: 203–214.
- [5] FORSTER, O. (2004). *Analysis I*. 7., verbesserte Auflage. Vieweg.
- [6] GEORGI, H.-G. (2002): *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter.
- [7] KOHL, M. UND RUCKDESCHEL, P. (2025): distrEx. Extensions of Package 'distr'. Version 2.9.6. <https://CRAN.R-project.org/package=distrEx>, DOI: 10.32614/CRAN.package.distrEx
- [8] NEWMAN, D.J. UND SHEPP, L. (1960): The Double Dixie Cup Problem, *American Mathematical Monthly* **67**: 58–61.
- [9] R CORE TEAM (2025). : *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- [10] RUCKDESCHEL, P., KOHL, M., STABLA, T. UND CAMPHAUSEN, F. (2006): S4 classes for distributions. *R News*, **6**(2): 2–6. [https://CRAN.R-project.org/doc/Rnews/Rnews\\_2006-2.pdf](https://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf)

## VIELEN DANK FÜR DIE AUFMERKSAMKEIT!