So erkennt man einen Lügner

LANGE NACHT DER MATHEMATIK

BONN – MÜNSTER

MATTHIAS LÖWE

13.11.2020

Inhalt- So erkennt man einen Lügner

- Einleitendes
- Verführerische Statistiken
- Die unbekannte Zahl
- Der gefälschte Münzwurf
- Der irre Drucker
- Erfundene Daten
- Schlussbemerkungen

Einleitung

- So erkennt man einen Lügner

Einleitung

- Wer lügt hinterlässt Spuren
- Diese Spuren lassen sich in günstigen Fällen aufdecken durch
  - Wiederholte Kontrolle (Konsistenz)
  - Eigenschaften der „wahren Größen“
  - Clevere Ideen …
- Es geht also stets um die Wiederherstellung gestörter Information

Verführerische Statistiken

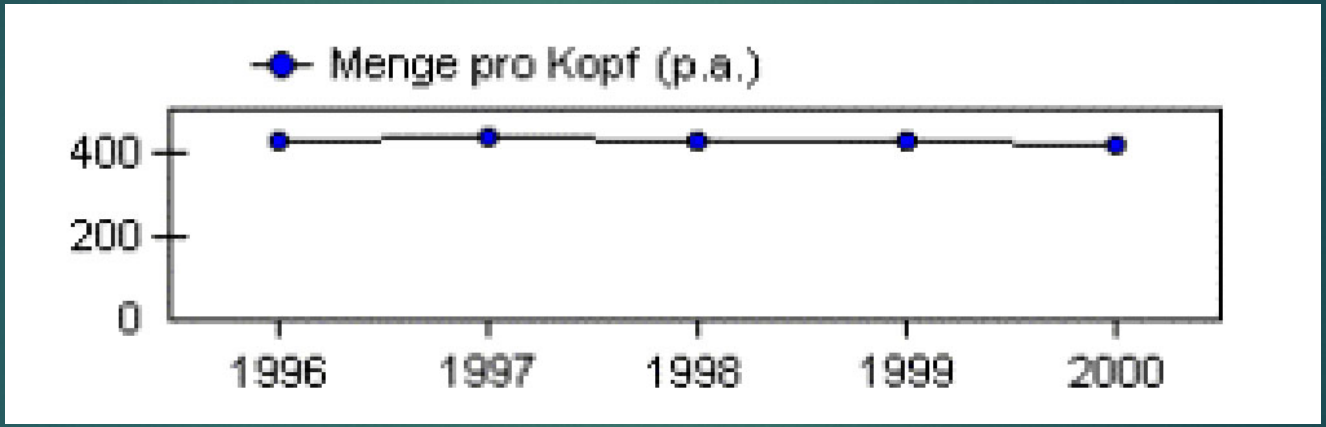
- Oftmals hinterlassen Grafiken – gewollt oder ungewollt – ein verzerrtes Bild dessen, was eigentlich gesagt werden soll.
- Bsp.: Wir wollen diese Tabelle grafisch darstellen:

Jahr	Hausmüll pro Kopf	Veränderung zum Vorjahr	
	in kg	abs.	in %
1996	429	-	-
1997	443	14	3 %
1998	437	-6	-1 %
1999	431	-6	-1 %
2000	425	-6	-1 %

Verführerische Statistiken

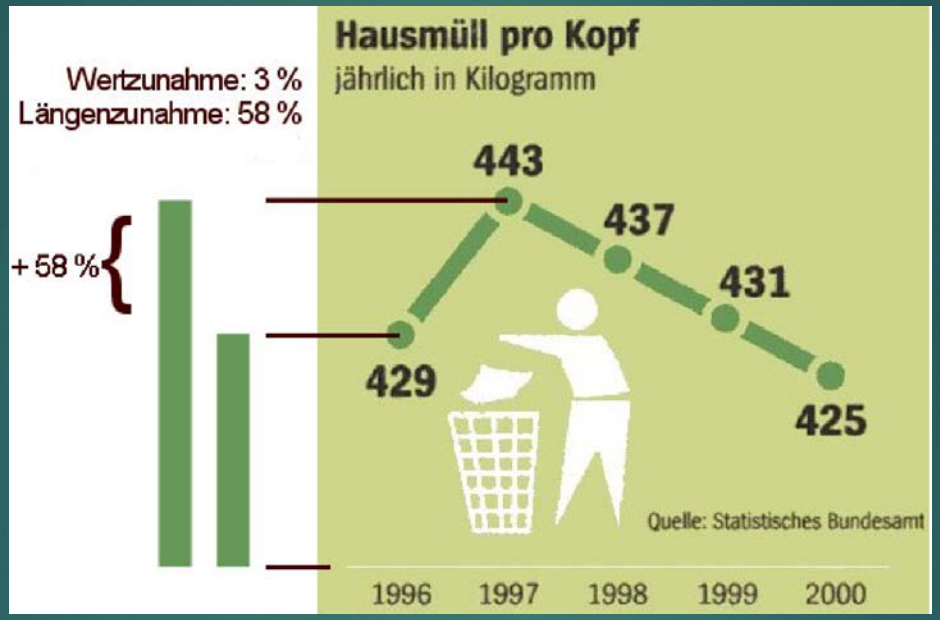
- Eine realistische Präsentation sähe so aus

(langweilig!):



Verführerische Statistiken

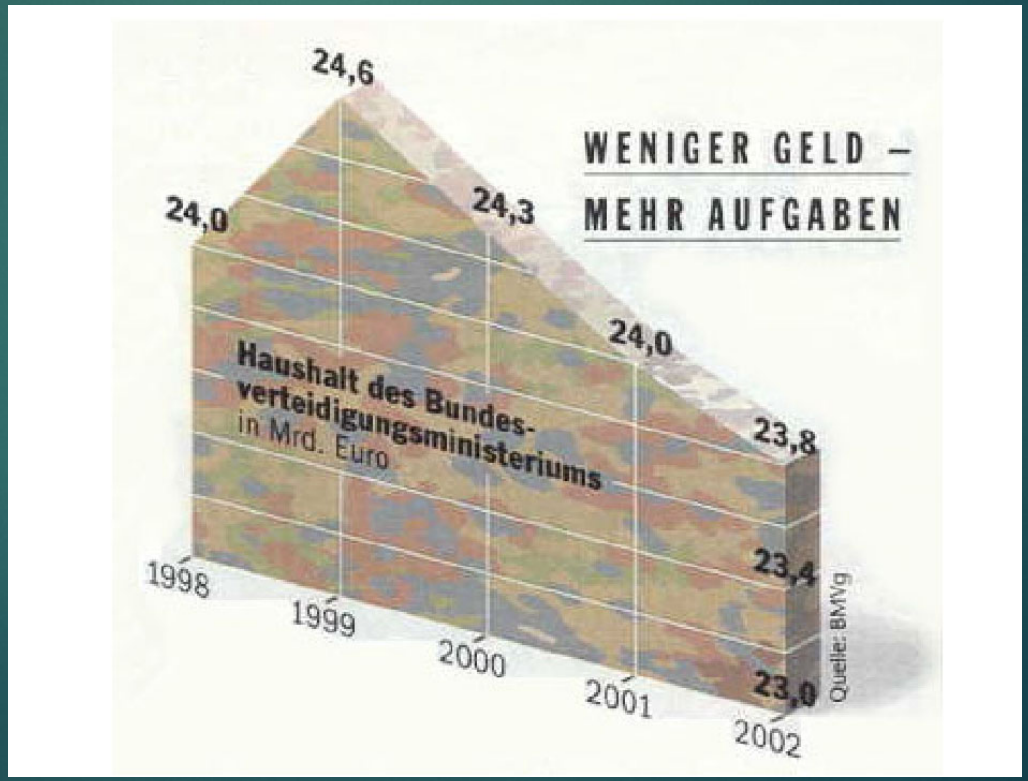
- Viel spannender aber ist es so:



- Abgeschnittene Achsen und verzerrte Größenverhältnisse gehören zu den Standardhilfsmitteln „statistischer Lügner“

Verführerische Statistiken

- Ähnliche Tricks finden sich auf jeder zweiten Seite eines beliebigen Nachrichtenmagazins:

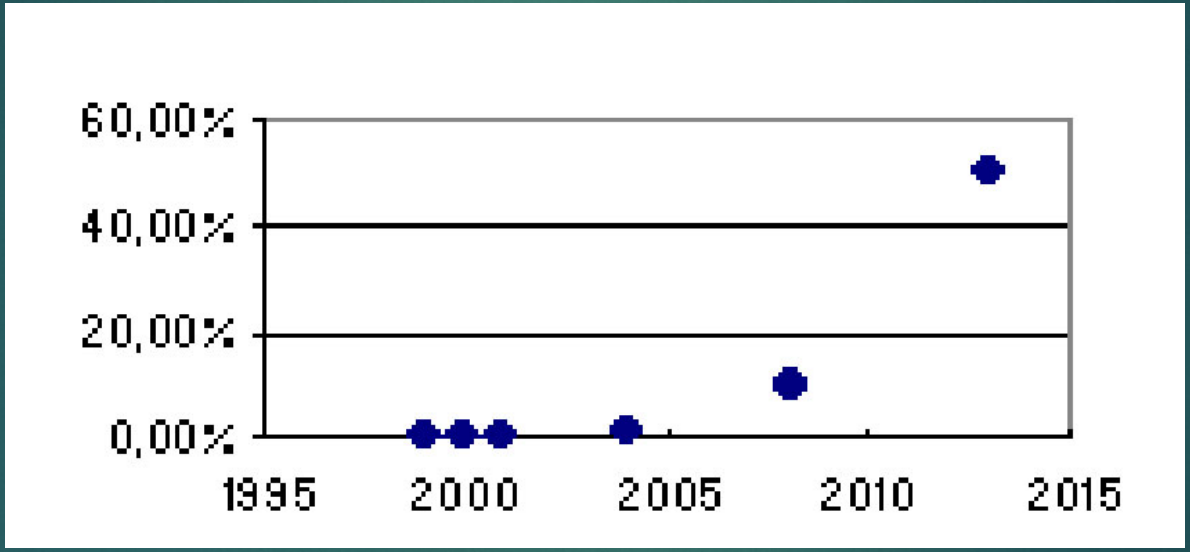


Verführerische Statistiken

- Beherrscht man sein Handwerk nicht, so können Grafiken sogar kontraproduktiv wirken:
- Bsp.: Die Firma Message Labs beobachtete eine Zunahme der Virenmails in den Jahren 1999-2001 von 1/1400 auf 1/300
- Sie prognostizierte daraus ein exponentielles Wachstum der Virengefahr.

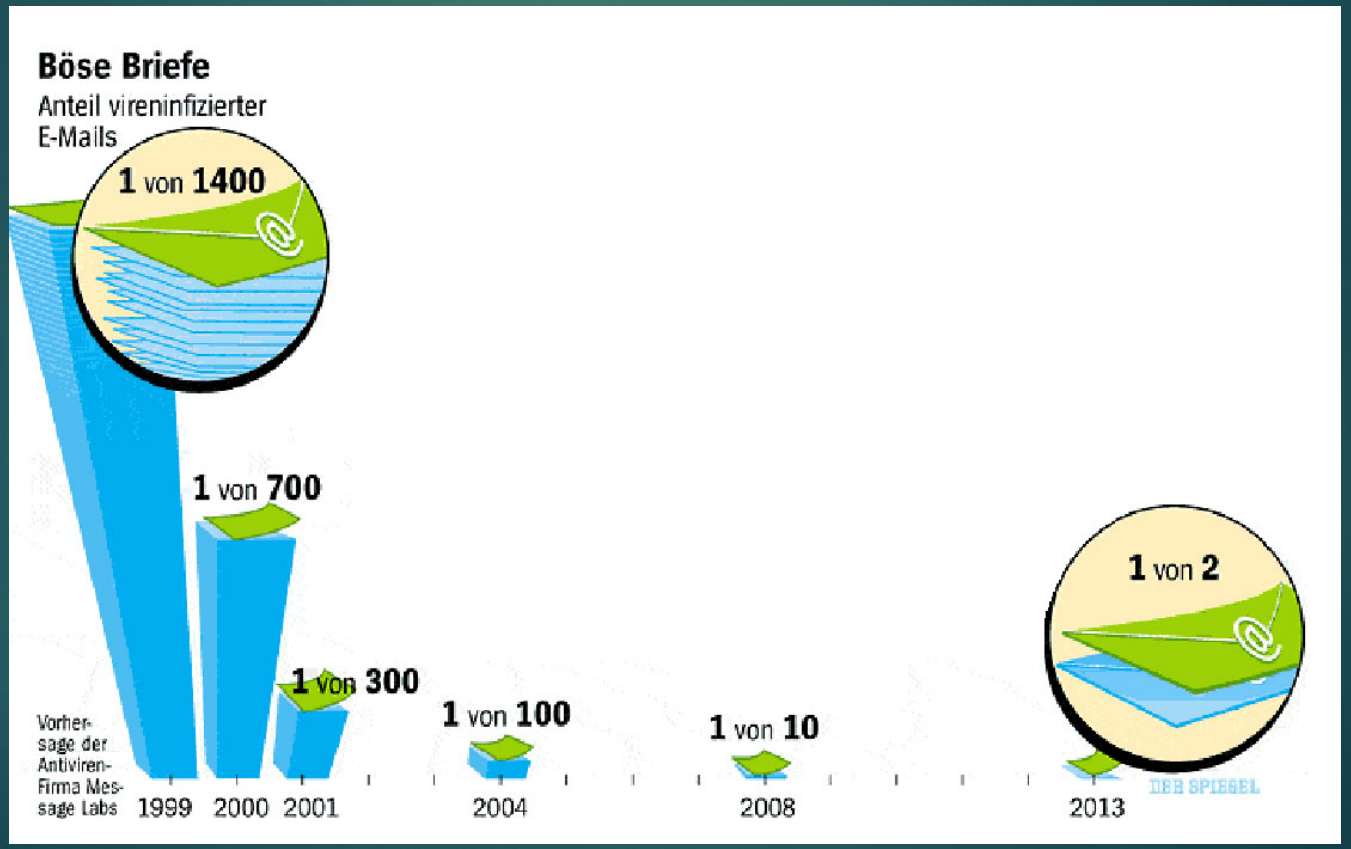
Verführerische Statistiken

- Gemeint war also ein Wachstum dieser Form:



Verführerische Statistiken

- Grafik des Spiegel



Verführerische Statistiken

- Probleme ergibt oft auch die richtige Interpretation der Ergebnisse:
- Habe ich die richtige Frage gestellt?
- Die Firma Durex veröffentlichte eine Studie zum Sexualverhalten heterosexueller Menschen in 41 Staaten.
- Danach haben Frauen im Schnitt 7 verschiedene Sexualpartner, Männer 10.

Verführerische Statistiken

- Korrelation versus Kausalität:
- Eine Studie zeigt, dass in Schweden die Anzahl der Neugeborenen positiv korreliert ist mit der Anzahl der Störche
- Schlussfolgerung: ???

Verführerische Statistiken

- Korrelation versus Kausalität:
- Ähnliche Beispiele:
- Verheiratete Menschen leben länger als Unverheiratete
- Aber Geschiedene leben noch länger
- Zweimal Geschiedene noch länger

Verführerische Statistiken

- Korrelation versus Kausalität:
- Menschen mit größeren Füßen haben ein höheres Einkommen
- Aber auch einen höheren IQ
- Umgekehrte Kausalität:
- Je schneller sich eine Windmühle dreht, desto mehr Wind gibt es.

Verführerische Statistiken

- Verstehe ich mein eigenes Experiment?
- In einer amerikanischen Studie wurden 10.000 Menschen auf einem 5% Niveau auf ASW getestet.
- Resultat: 500 davon hatten übersinnliche Fähigkeiten
- Bei einem erneuten Test zeigten sich diese nur bei 25 der 500.
- Resultat: Wenn man jemandem von seinen Fähigkeiten berichtet, verschwinden diese!

Verführerische Statistiken

- Simpsons Paradoxon
- Meine Prüfungsstatistik

	Männer			Frauen		
	Best.	ges.	D- Quote	Best.	ges.	D- Quote
1. Tag	1	1	0%	7	8	12,5%
2. Tag	2	3	33,3%	1	2	50%

Verführerische Statistiken

- Bevorzugen die Prüfungen Männer?

	Männer			Frauen		
	Best.	ges.	D- Quote	Best.	ges.	D- Quote
gesamt	3	4	25%	8	10	20%

- Der Grund ist die unterschiedliche Anzahl an Prüflingen (Gewichtung) an den einzelnen Tagen

Die unbekannte Zahl

- Aufgabe: Ich denke mir eine Zahl, versuchen Sie diese zu erraten!
- Das ist *so* unmöglich (jiddisch Poker)
- Variation: Ich denke mir eine Zahl zwischen 1 und N, versuchen Sie diese mit „Ja-Nein-Fragen“ zu erraten!
- Dazu benötigen Sie mit der richtigen Technik log(N) Fragen (Halbierungsstrategie)

Die unbekannte Zahl

- Was ändert sich, wenn ich lügen darf?
- Bei einer Lüge: Anzahl der notwendigen Fragen erhöht sich höchstens um den Faktor 3.
- Bei mehr als 50% Lügen: Das Auffinden der richtigen Zahl wird unmöglich.
- Dazwischen?

Die unbekannte Zahl

- Es gilt:

Satz: Wenn der Antwortende zu jedem Zeitpunkt höchstens in r % aller Antworten gelogen haben darf, so kann der Fragende gewinnen, wenn $r<50$ und der Antwortende gewinnt, wenn $r \geq 50$. Im Gewinnfalle braucht der Fragende weniger als Const. log(N) Fragen

Die unbekannte Zahl

- Veränderung der Spielregeln: Der Antwortende darf *insgesamt* nicht öfter als r% lügen, diese Lügen aber beliebig verteilen.
- Dies ist günstiger für den Antwortenden (Carol).
- Tatsächlich ändert sich nun auch der kritische Wert auf r<33,3%.
- Wieder gelingt ein Gewinn in Const. log(N) Schritten

Der gefälschte Münzwurf

- Wie erkennt man, dass ein Münzwurf „nicht echt“ ist?
- Für einen Wurf: Unmöglich!
- Bei mehreren Würfen hilft z.B. das **Gesetz der großen Zahlen**:
- Bei einer fairen Münze strebt die **relative Häufigkeit** der „Köpfe“ gegen den Wert ½.

Der gefälschte Münzwurf

- Mögliche Fehlinterpretationen:
- Die Wahrscheinlichkeit genau N/2 Köpfe zu werfen geht gegen 1. (Nein, die geht sogar gegen 0)
- Wenn man im Casino 10 Mal hintereinander „rot“ sieht, steigt die Chance für schwarz. (Nein, sie bleibt ½, vielleicht ist aber auch der Kessel kaputt)

Der gefälschte Münzwurf

- Rényis Experiment:
- Teile eine Vorlesung in zwei Gruppen:
- Gruppe 1 führt (pro Person) 200 Münzwürfe durch und notiert die Ergebnisse
- Gruppe 2 simuliert Münzwürfe, indem jeder 200 „gefakte“ Ergebnisse notiert.
- Wie kann man die echte von den falschen Münzwürfen unterscheiden?

Der gefälschte Münzwurf

- Auch die gefakten Münzwürfe weisen zumeist ca. 50% „Köpfe“ auf.
- Betrachte die längste Folge aufeinanderfolgender Köpfe (1-Run)
- Ihre Länge sei bei N Würfen L(N)
- Wie groß ist L(N) typischerweise?

Der gefälschte Münzwurf

- Heuristik
- Angenommen: der längste 1-Run ist eindeutig.
- Dann gibt es auch im Durchschnitt einen längsten 1-Run.
- Wahrscheinlichkeit für einen 1-Run der Länge L, der an fixer Position beginnt:

$$(1/2)^L$$

Der gefälschte Münzwurf

- Es gibt ca. (etwas weniger als) N Möglichkeiten einen längsten 1-Run zu beginnen
- Somit

$$N\,(1/2)^{L(N)}=1$$

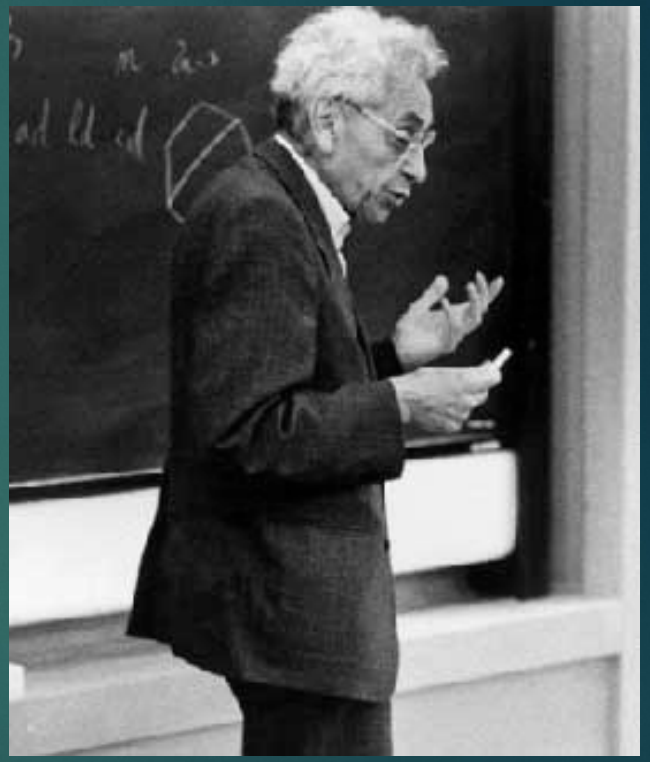
- Dies ergibt

$$L(N)= \log(N)/\log(2)$$

- Für N=200 ist L(N) ungefähr 7,64.
- Man muss daher in seiner gefälschten Folge ca. 7-8 aufeinanderfolgende Köpfe haben um realistisch zu sein.

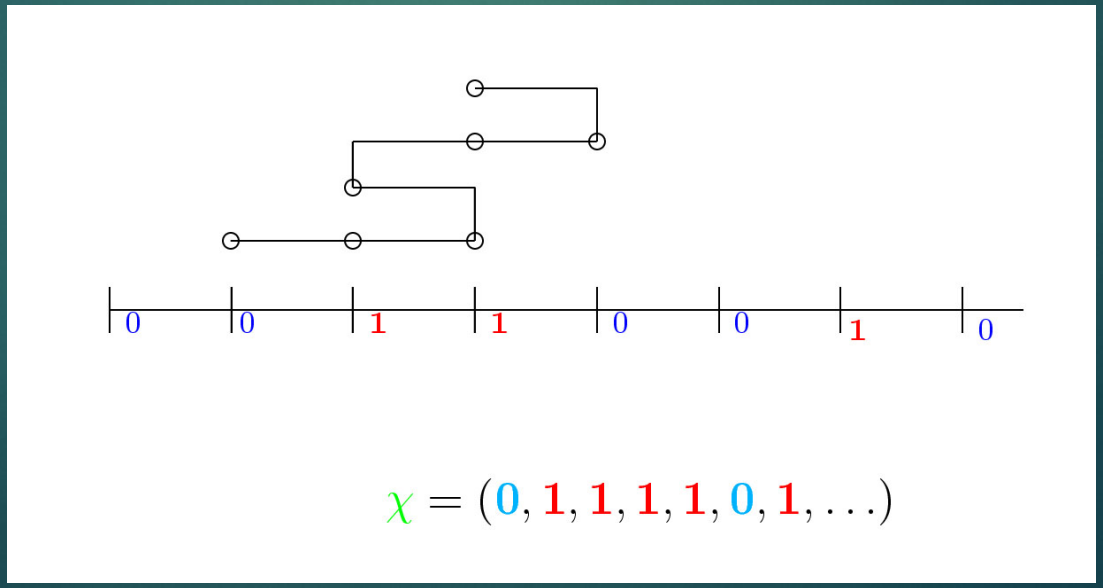
Der gefälschte Münzwurf

- Dahinter steckt der sogenannte Grenzwertsatz von Erdös und Rényi
- Paul Erdös (1913-1996)
- Mathematisches Wunderkind und einer der bemerkenswertesten Mathematiker des 20. Jahrhunderts



Der irre Drucker

- Eine „Geheimschrift“:
- Ein Drucker (Fax, o.ä.) erhält einen Text
- Der Drucker liest den Text nicht von links nach rechts, sondern springt bei jedem Schritt zufällig nach links oder rechts.



Der irre Drucker

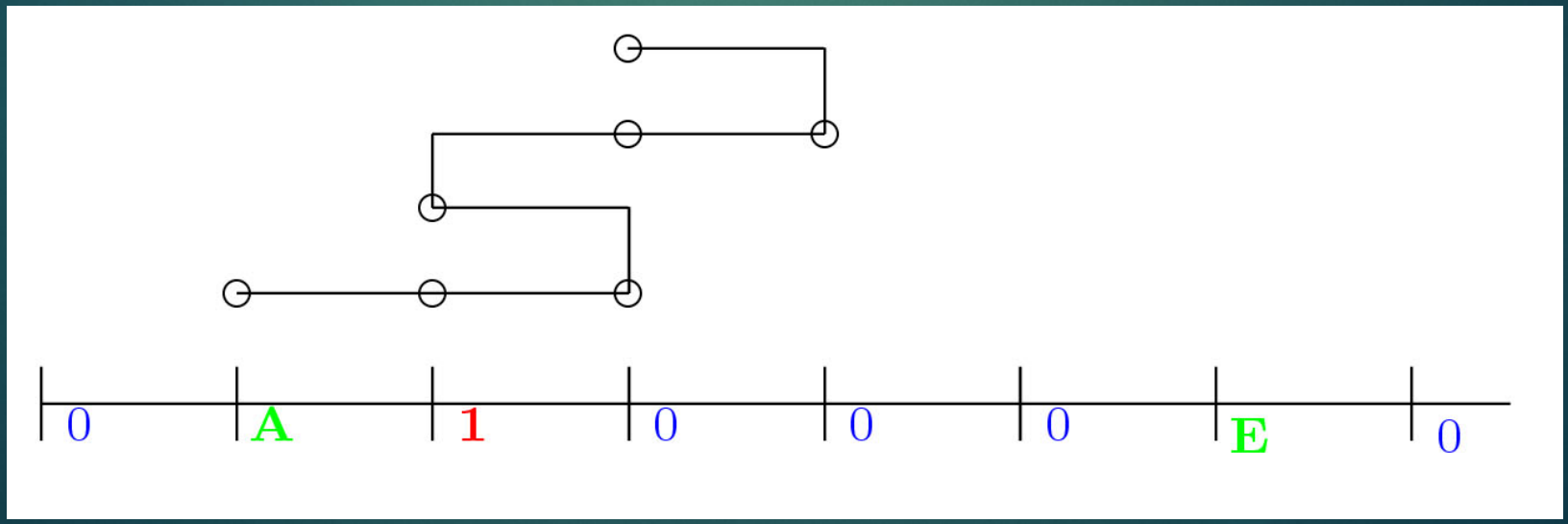
- Frage: Kann man aus dem „Ausdruck“ den ursprünglichen (unendlich langen) Text rekonstruieren, wenn man nicht weiß, wie der Drucker sich bewegt?
- Antwort: Nein, denn man weiß schon nicht, ob der Drucker zuerst nach links oder rechts geht.
- Ebenso verliert man die Information über den Startpunkt des Druckers.

Der irre Drucker

- Es gibt auch Texte, die nicht rekonstruiert werden können.
- Aber: Typische Texte (d.h. fast alle, die man erhält, wenn man die Buchstaben zufällig auswürfelt), können für typische Druckerbewegungen bis auf den Anfangspunkt und Spiegelsymmetrie rekonstruiert werden.

Der irre Drucker

- Wichtige Information: Bei typischen Druckerbewegungen liest der Drucker jede Textpassage unendlich oft.
- Ein Spielzeugbeispiel mit 2 Extrabuchstaben:



Der irre Drucker

- A und E kommen nur einmal im Text vor.
- Wenn der Drucker in A ist, hat er eine positive Wahrscheinlichkeit auf direktem Wege nach E zu gehen.
- Da er unendlich oft in A ist, wird er irgendwann direkt von A nach E gehen.
- Dies ist die kürzeste Art erst A und dann E zu sehen.
- Dabei liest der Drucker exakt den Text zwischen A und E.
- Wir schauen also in unserem Ausdruck auf den kürzesten Abstand zwischen A und E. Dazwischen steht der wahre Text.

Der irre Drucker

- Im Beweis besteht eine Hauptschwierigkeiten darin, diese besonderen Buchstaben zu konstruieren und aufzufinden.
- Der Beweis für ein 2-buchstabiges Alphabet ist aufwändig (ca. 60 Seiten)
- Für größere Alphabete ist er intuitiver und einfacher.

Erfundene Daten

- Niemand würde seine Steuererklärung mit folgenden Daten fälschen:
- 1 PC 512,00€
- 1 Bewirtung 53,00€
- 1 Büromaterial 59,00€
- Fortbildung 587,95€
- 1 Drucker 57,00€

Die 5‘en als Anfangsziffer machen skeptisch

Erfundene Daten

- Wir erwarten eine annähernde Gleichverteilung der Anfangsziffern.
- Ist das wahr?
- Nein, für sehr viele Datensätze
- Statistische Untersuchungen zeigen, dass die Anfangsziffern sehr vieler Datensätze dem Benford Gesetz gehorchen:

Erfundene Daten

- Je niedriger eine Anfangsziffer ist, umso häufiger tritt sie auf
- Genauer gilt die rechtsstehende Verteilung

Ziffer	W.keit
1	30,1%
2	17,6%
3	12,5%
4	9,7%
5	7,9%
6	6,7%
7	5,8%
8	5,1%
9	4,6%

Erfundene Daten

- Diese Verteilung gilt (annähernd) z.B. für
- Die Fibonacchi-Zahlen
- Eine Liste der Höhen der höchsten Berge
- Die Hausnummern der Amerikaner im „Who is who“
- Die Einwohnerzahlen von Städten
- Die Größe der Dateien in meiner Linuxpartition ….

Erfundene Daten

- Dies wurde 1881 von Simon Newcomb anhand der Tatsache entdeckt, dass Logarithmentafel auf den ersten Seiten abgenutzter sind als auf den hinteren
- 1938 wurde dieses Ergebnis erneut von dem Statistiker Frank Albert Benford publiziert.

Erfundene Daten

- Warum genügen viele Folgen dem Benford Gesetz (und welche)?
- Dies muss für die Folgen gesondert betrachtet werden, da sie oftmals sehr unterschiedliche Bildungsgesetze haben.
- Für einige der genannten Folgen ist analysiert, für andere nur empirisch bestätigt.

Erfundene Daten

- Warum taucht das Benford Gesetz auf (und nicht z.B. die Gleichverteilung)?
- Es ist plausibel, dass Folgen, die dem Benford Gesetz genügen, dies auf jeder Skala tun, d.h. unabhängig davon, ob ich eine Länge z.B. in Meter, Meilen oder Ellen messe.
- Man kann zeigen: Die Benford-Verteilung ist die einzige (auf {1, …9}) die skaleninvariant ist.
- Ihre genaue Form ist

$$p_k = \log_{10}(1+1/k)$$

Zusammenfassung

- Wir haben verschiedene Methoden kennen gelernt, „gefälschte Daten“ zu identifizieren. Z.B. durch
- Geschicktes Umformulieren der Frage
- Ausnutzung intrinsischer Eigenschaften des Zufallsmechanismus
- Geschickte Wiederholungen
- Einige dieser Methoden haben Anwendungen in Situationen des alltäglichen Lebens.