Statistik und Wahrscheinlichkeitsrechnung

Dr. Benjamin Adrian

Algorithmen und Technologien für die Betrugserkennung

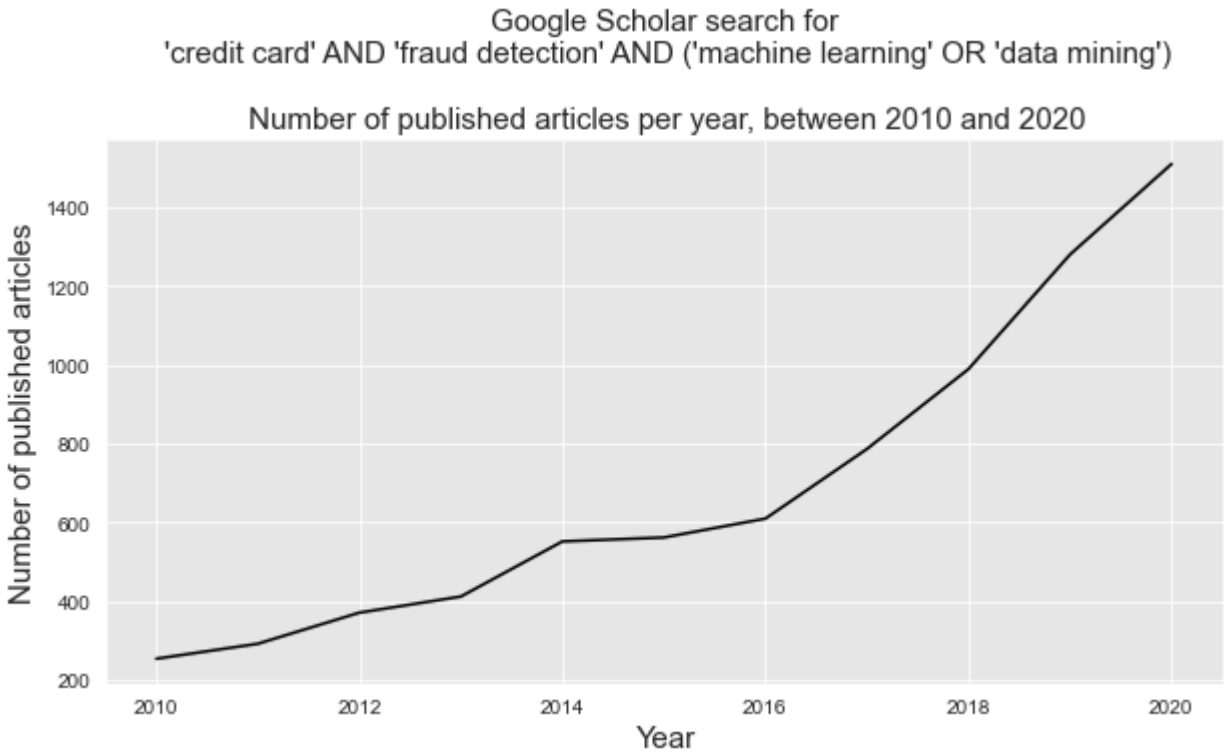
Gliederung
Betrugserkennung

Problemstellung

- Was zeichnet Betrug aus?
- Beispiel
- Wie erkennt man Betrug?

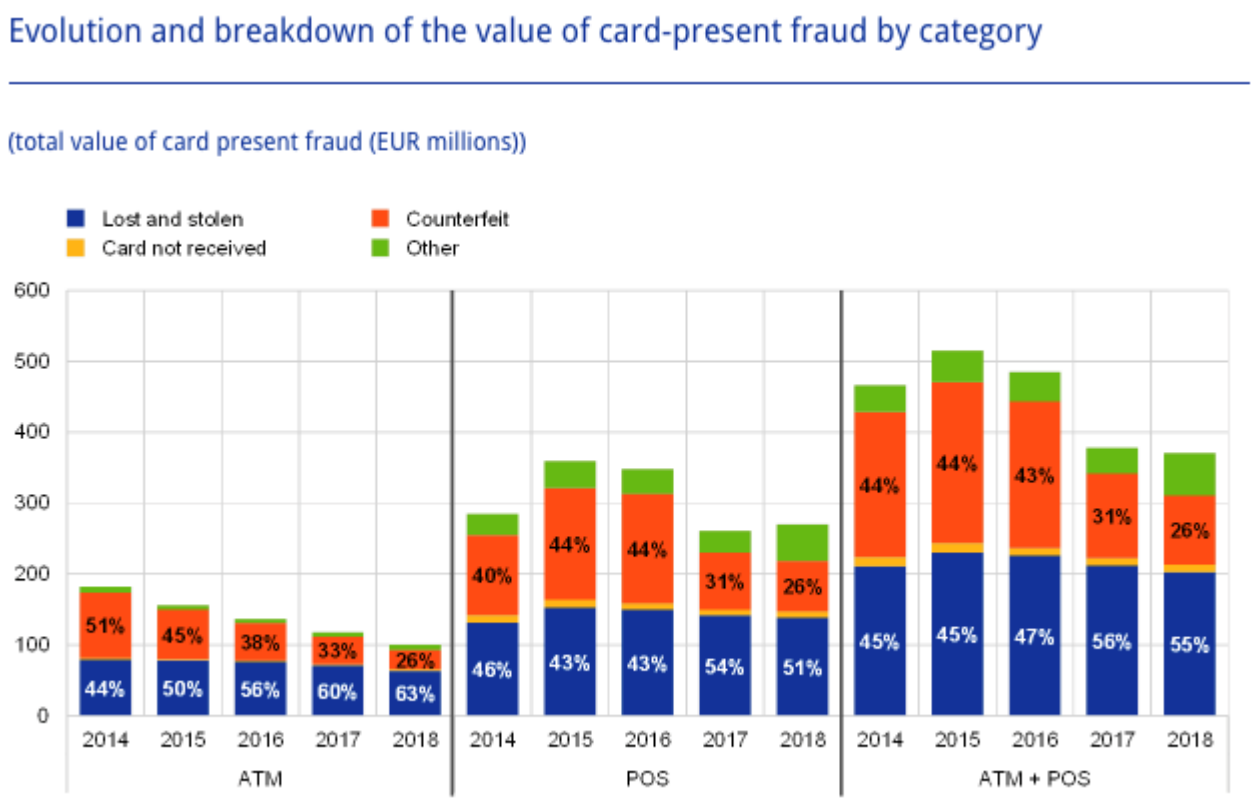
Algorithmen und Technologien

- Descriptive Analytics
- Predictive Analytics



https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_2_Background/MachineLearningForFraudDetection.html

Problemstellung



https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_2_Background/CreditCardFraud.html

Was zeichnet Betrug und Betrüger aus?

Definition und Fakten

Betrug

Betrug ist ein ungewöhnliches, gut durchdachtes, unmerklich verborgenes, sich mit der Zeit entwickelndes und oft sorgfältig organisiertes Verbrechen, das in den unterschiedlichsten Formen auftritt.

Betrüger

… lernen bestehende Unternehmensregeln, z.B. über Trial & Error.

… teilen ihr Wissen über erfolgreiche betrügerische Handlungen

- Baesens, B., Van Vlasselaer, V., Verbeke, W. (2015). Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Vereinigtes Königreich: Wiley.

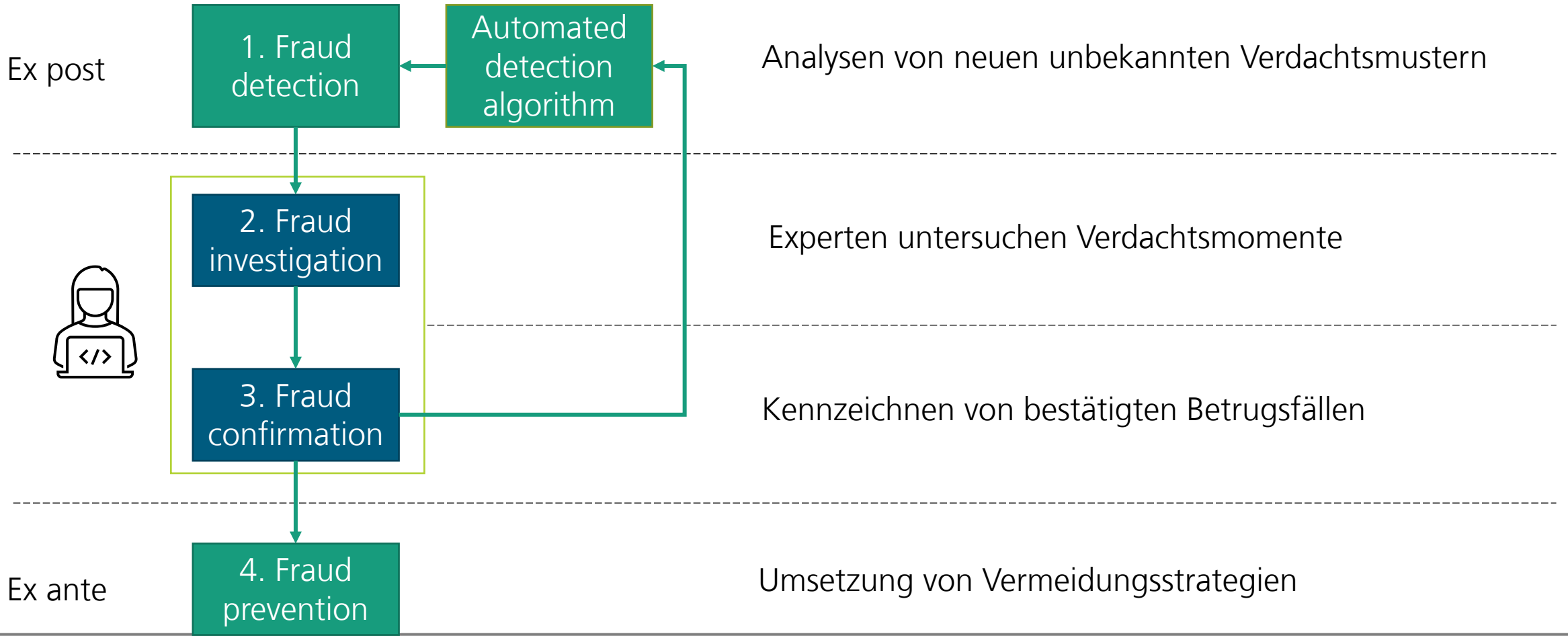
Fakten

- Durch Betrug verlieren durchschnittliche Firmen 5% ihres Ertrags.
- 0.5% aller Kreditkartentransaktionen gelten als betrügerisch
- Die US kostet Versicherungsbetrug 40 Mrd. $ pro Jahr

Einige Betrugskategorien

- Kreditkartenbetrug: z.B. Identitätsdiebstahl
- Versicherungsbetrug: z.B. gefälschte Rezepte
- Steuerhinterziehung: z.B. Cum-Ex
- Geldwäsche: z.B. gefälschte Abrechnungen
- Bilanzfälschung: z.B. Wirecard

Betrugserkennungszyklus

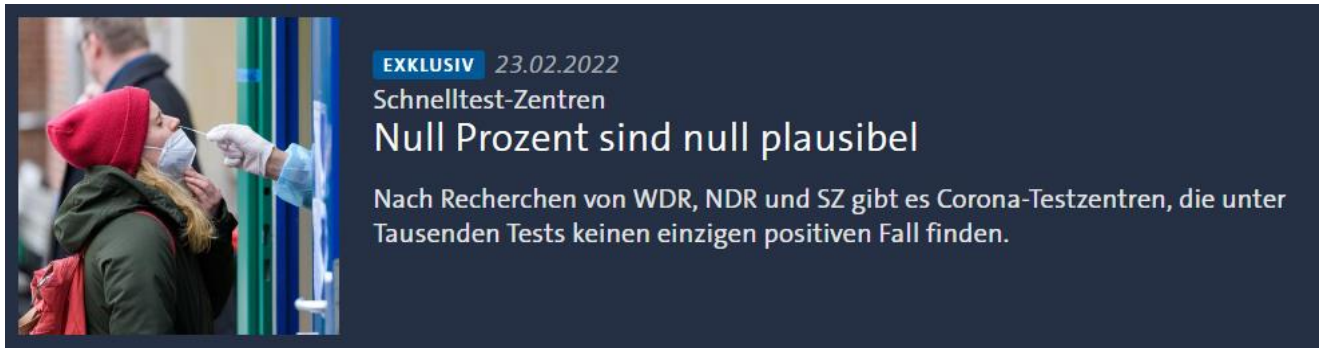


RKI soll Coronatest-Betrug aufdecken

Beispiel

Stand: 17.08.2022 17:59 Uhr

[…] Demnach soll das RKI künftig die Abrechnungsdaten analysieren, "statistische Ausreißer" identifizieren, […] wie die Positivrate der Schnelltests. Entdecke das RKI Unregelmäßigkeiten, solle es die örtlichen Gesundheitsämter und die zuständige Kassenärztliche Vereinigung unterrichten, so der Plan des Ministeriums.



Corona Schnelltest-Center

Konkrete Zahlen

Szenario:

In einem Kreis wurde eine Inzidenz von
990 (Infizierte pro 100.000 Einwohner) festgestellt.

Von mehreren Schnelltest-stationen wurden in einem Zeitraum die Zahlen von 10.000 Testresultate gesichtet.

Gegeben:

$$\text{Inzidenz} = 990$$

$$p_{\text{test=positiv}} = 0.0099$$

$$n_{\text{Anzahl Tests}} = 10000$$

$$k_{\text{Anzahl positive Tests}}$$

Corona Schnelltest-Center

Modellierung von Coronatests mit Binomialverteilung

Gegeben

$$\text{Inzidenz} = 990$$

$$p_{\text{test=positiv}} = 0.0099$$

$$n_{\text{Anzahl Tests}} = 10000$$

$$k_{\text{Anzahl positive Tests}}$$

Definition: Bernoulli-Versuch

Ein Bernoulli-Versuch (oder Binomialversuch) ist ein Zufallsexperiment mit genau zwei möglichen Ergebnissen, bei dem die Erfolgswahrscheinlichkeit bei jeder Durchführung des Experiments gleich ist.

Definition: Binomialverteilung

Die Binomialverteilung $B(n,p)$ ist eine diskrete Verteilung und beschreibt die Anzahl der Erfolge k in einer Serie von gleichartigen und unabhängigen Zufallsexperimenten n mit Erfolgswahrscheinlichkeit p, die jeweils genau zwei mögliche Ergebnisse haben.

Corona Schnelltest-Center

Modellierung von Coronatests mit Binomialverteilung

Gegeben:

$$\text{Inzidenz} = 990$$

$$p_{\text{test=positiv}} = 0.0099$$

$$n_{\text{Anzahl Tests}} = 10000$$

$$k_{\text{Anzahl positive Tests}}$$

Binomialverteilung:

Verteilung: $B(n, p)$

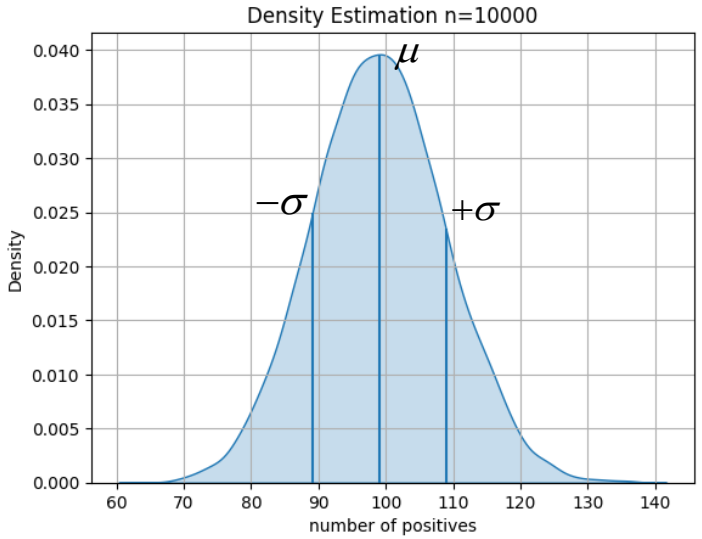
Wahrscheinlichkeit: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Erwartungswert: $\mu = np = 99$

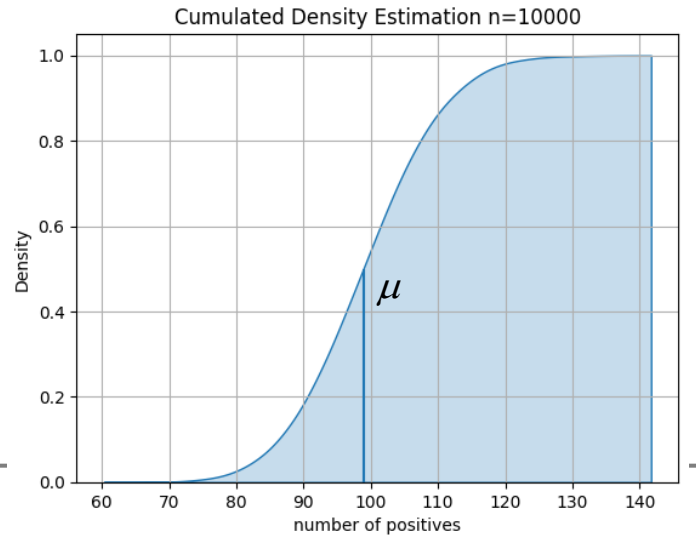
Varianz: $\sigma^2 = np(1-p+np) = 98.0199$

Standardabweichung: $\sigma = 9.9$

Dichte:



Kumulierte Dichte:



Datenlage

Zeitreihen

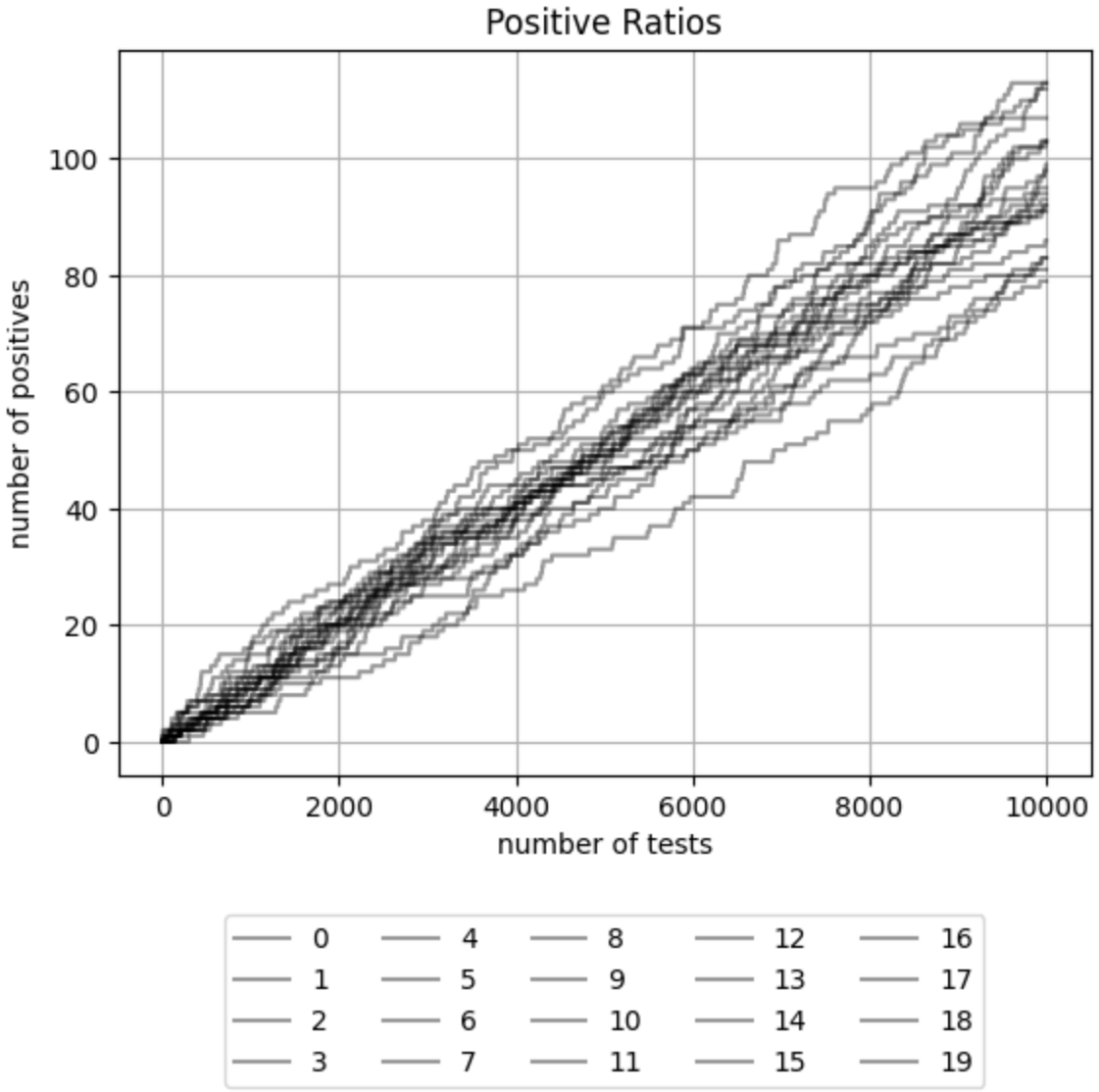
Aufsummierte Ergebnisse der Bernoulli Zufallsexperimente innerhalb eines Bernoulli-Prozesses (random walk):

```
from scipy.stats import binom

p_i=0.0099
number_tests=10000
good_stations = 20
positive_counts_good_station = np.array([0 for good_station in
range(good_stations)])


for i in range(1, number_tests+1):
   positive_counts_good_station += binom.rvs(1, p_i, size=good_stations)


cases_sum.append([i] + positive_counts_good_station.tolist())
```



Corona Schnelltest-Center

Betrugstrategien

1. Der naïve Betrüger

Die Auftrittswahrscheinlichkeit von 0.0099 ist sehr niedrig. Ich teste nicht wirklich und gebe alle Ergebnisse als negative an. Das merkt doch keiner.

2. Der etwas klügere Betrüger

Wenn geprüft wird, dann bestimmt nur der Mittelwert, deshalb besorge ich mir positive Ergebnisse und gebe sie nach jedem 1/0.0099 = 101 Mal an.

3. Der smarte Betrüger

Ich habe eine laufende Teststation in einer anderen Stadt mit einer Inzidenz von 800. Ich nehme diese Daten einfach und reiche die Ergebnisse hier nochmal zur Abrechnung ein.

4. Der sehr smarte Betrüger

Ich teste wirklich. Aber nach 1000 Tests gebe ich 200 Tests mit negativen Ergebnissen dazu und teste dann wieder 800 Personen richtig.

Algorithmen und Technologien

Deskriptive Analytik

Prädiktive Analytik

Betrugserkennung

Deskriptive Analytik

Deskriptive Analytik

Findet anormale Verhaltensweisen in Beobachtungen , die vom normalen Verhalten der Grundgesamtheit abweichen.

- Erkennt bislang unbekannte Verdachtsmomente
- Anfällig für Täuschung

Anomalieerkennung:

- Erkennung von statistischen Ausreissern
  Eine „außen liegende“ Beobachtung oder Ausreißer ist eine, die deutlich von anderen Mitgliedern der Stichprobe abzuweichen scheint, in der sie auftritt.
- Erkennung von Strukturbrüchen
  Strukturbrüche in Zeitreihen sind dadurch gekennzeichnet, dass sich das stationäre Verhalten über die Zeit durch eine Veränderung der Varianz oder des Mittelwertes ändert.

Vorgestellte Methoden:

1. Z-Score
2. Statistische Tests
3. Binomial Proportion Tests
4. Strukturbruch-Analysen
5. Red Flags

Weitere Methoden:

1. Cluster-Analysen
2. Benfords Law

Z-Score

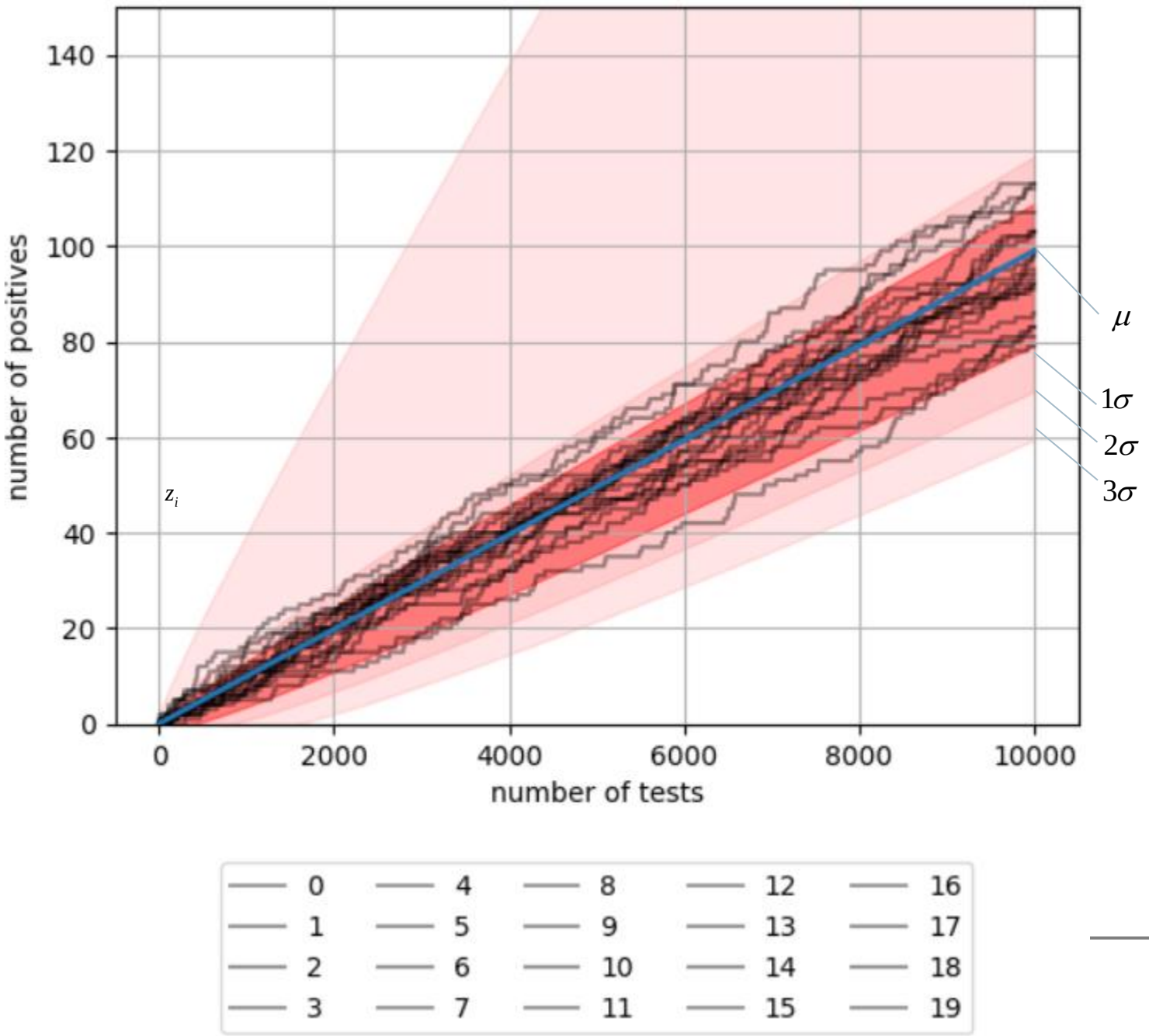
Messung der Abweichung von der Norm

Z-Score

Wenn der Erwartungswert μ und die Standardabweichung σ einer Zufallsvariablen bekannt sind, wird ein Rohwert x_i durch folgende Z-Transformation nach z_i normalisiert.

$$z_i = \frac{x_i - \mu}{\sigma}$$

z_i gibt an, die angibt welche Anzahl mal der Standardabweichung eine Beobachtung vom Erwartungswert entfernt ist.



Z-Score

Messung der Abweichung von der Norm

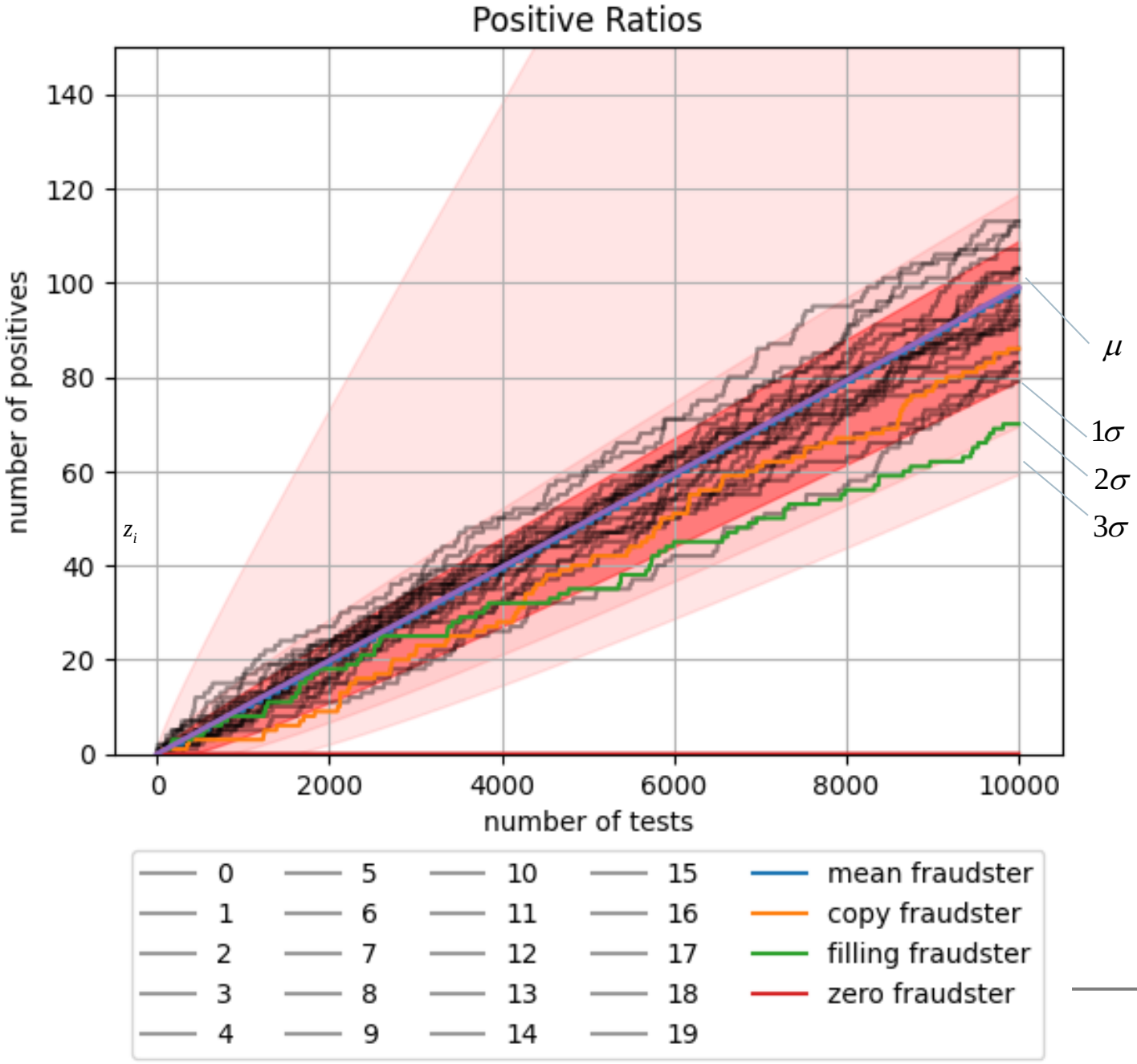
Z-Score

Wenn der Erwartungswert μ und die Standardabweichung σ einer Zufallsvariablen bekannt sind, wird ein Rohwert x_i durch folgende Z-Transformation nach z_i normalisiert.

$$z_i = \frac{x_i - \mu}{\sigma}$$

z_i gibt an, die angibt welche Anzahl mal der Standardabweichung eine Beobachtung vom Erwartungswert entfernt ist.

Z-Scores größer als 3 gelten als statistische Ausreißer!



Test auf Signifikanz

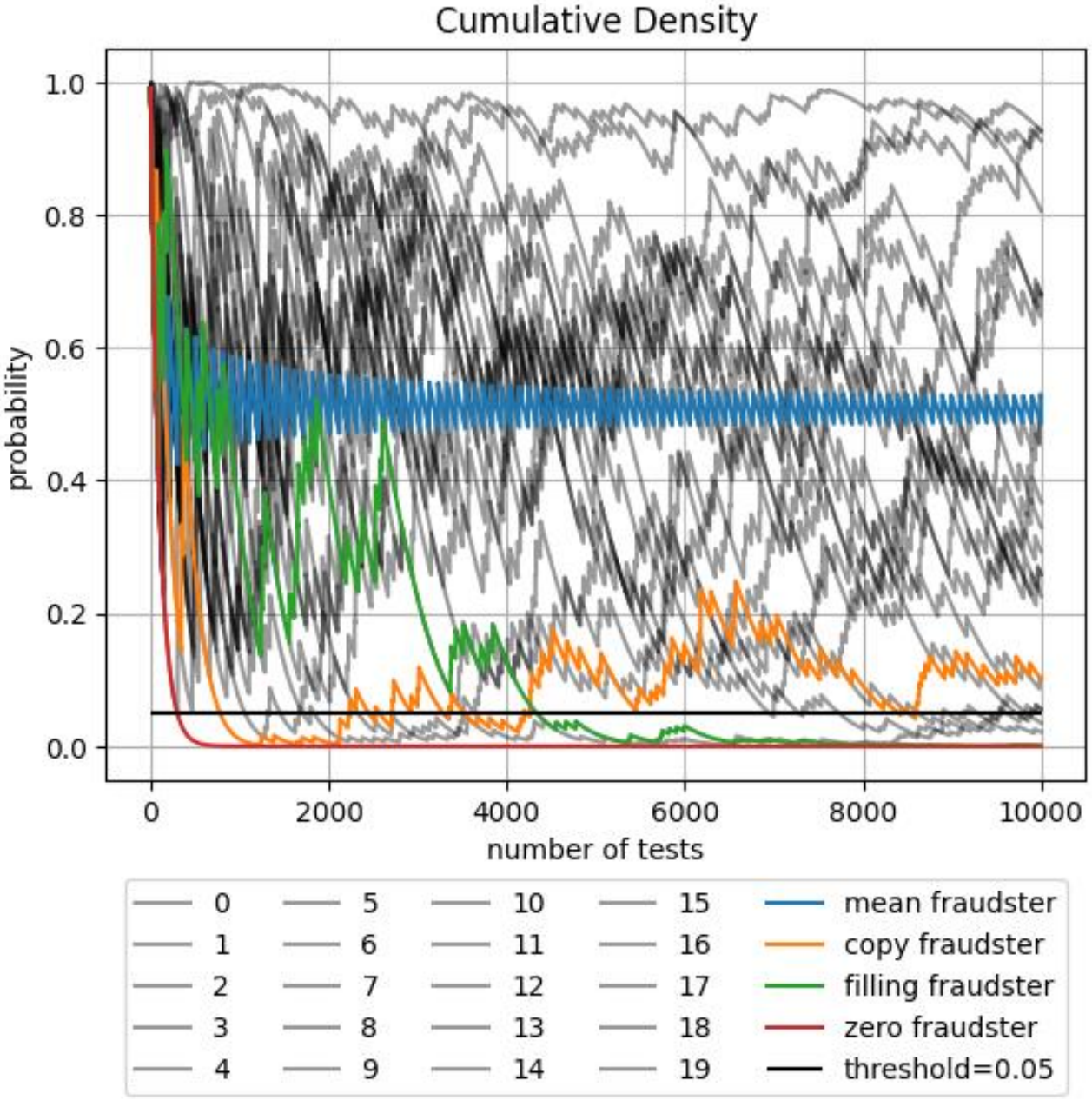
Wie wahrscheinlich ist die Testverteilung?

Aufsummierte Wahrscheinlichkeiten der Binomialverteilung

$$P(X \leq k) = \sum_{0}^{k} \binom{n}{k} p^k (1-p)^{n-k}$$

Einsatz von Konfidenzintervallen

$$P(X \leq k) \geq t$$



Konvergenz zum Erwartungswert

Binomial Proportion Tests

Two-Sample Binomial Proportion Test

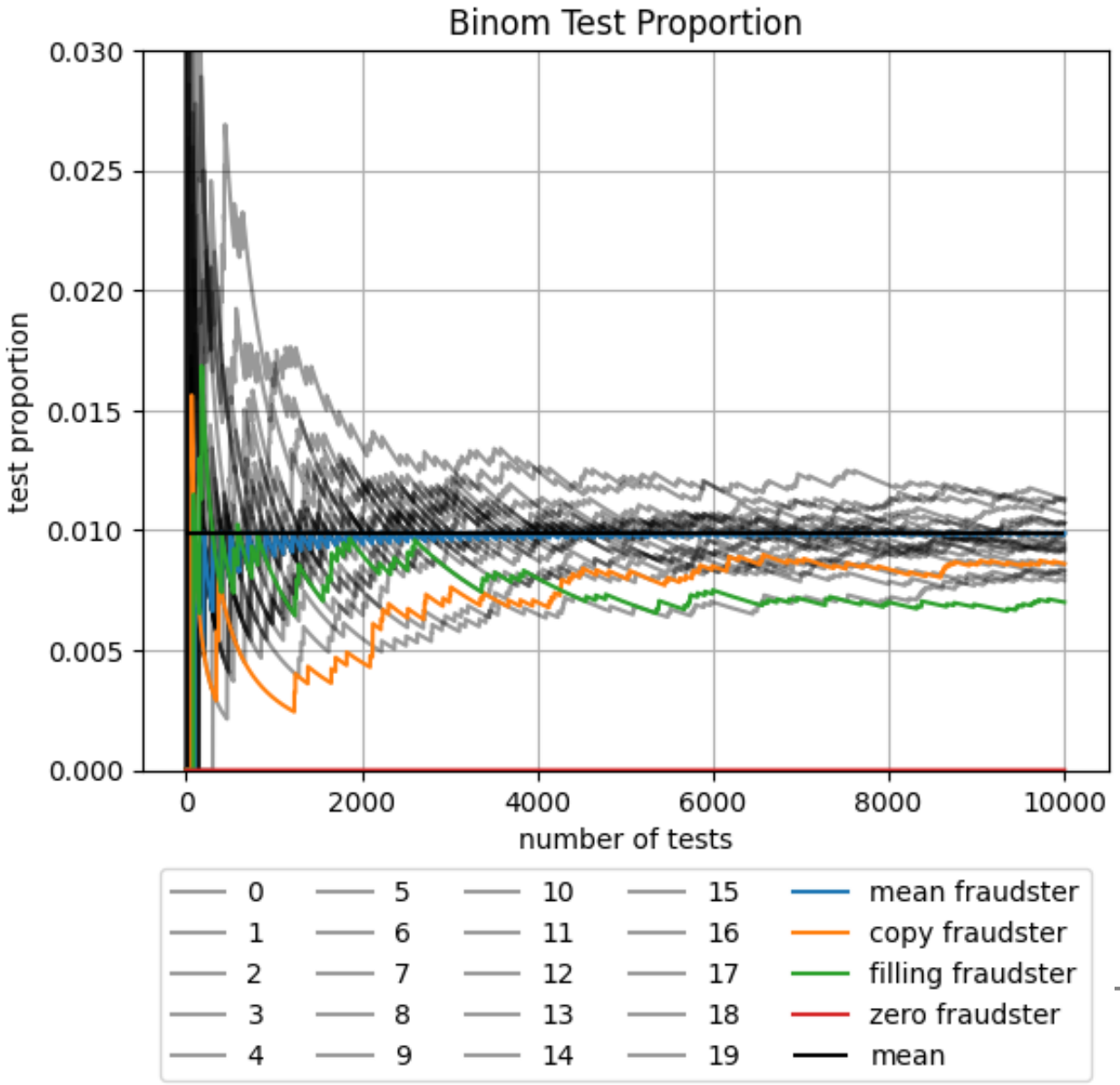
Angenommen, wir haben zwei oder mehr Beobachtungen einer gewissen Länge bzw. Zeitbereich.

Wir berechnen die positiven Anteile aus diesen Beobachtungen und prüfen, ob beide Beobachtungen die gleichen Anteile haben oder nicht

$$\text{Proportion}_t = \frac{\text{Anzahl positive Ergebnisse}}{\text{Anzahl Tests}}$$

Gesetz der großen Zahlen

Die Häufigkeit mit der ein Zufallsereignis eintritt, nähert sich seiner rechnerischen Wahrscheinlichkeit immer weiter an, je häufiger ein Zufallsexperiment durchgeführt wird.



Strukturbruchanalysen

Erkennung von verdächtigen Veränderungen im Verhalten

Stationarität

Stationarität beschreibt Eigenschaften von Zeitreihen, die invariant über die Zeit hinweg gültig sind. Eine stationäre Zeitreihe hat zu allen Zeitpunkten den *gleichen Erwartungswert* und die *gleiche Varianz*.

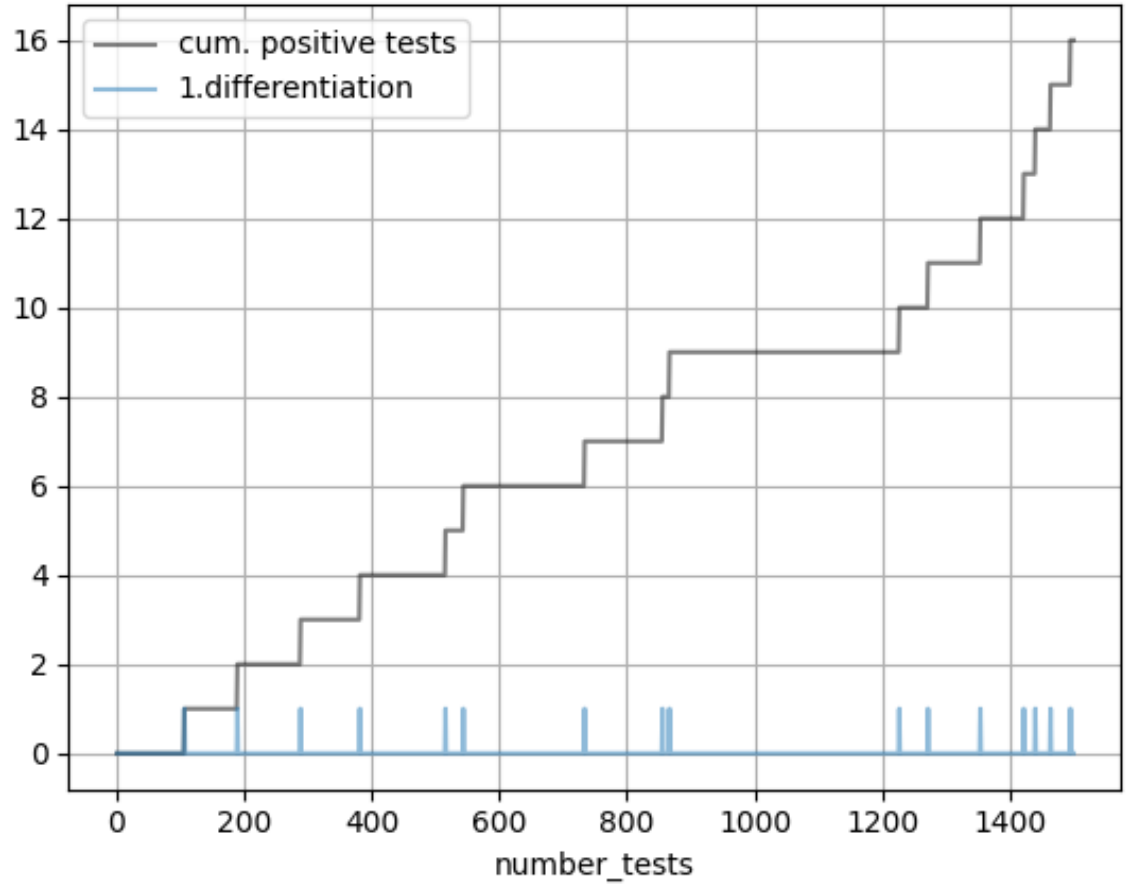
Numerische Differentiation

Um von der Zeitreihe f der stetig wachsenden aufsummierten positiven Testresultaten eine Zeitreihe f' ohne ansteigenden Trend zu erhalten, wird die erste numerische Differentiation gebildet.

$$f'(t) = f(t) - f(t+1)$$

Auf f' kann nun der Test auf Stationarität erfolgen.

z.B. Dickey-Fuller-Test



Strukturbruchanalysen

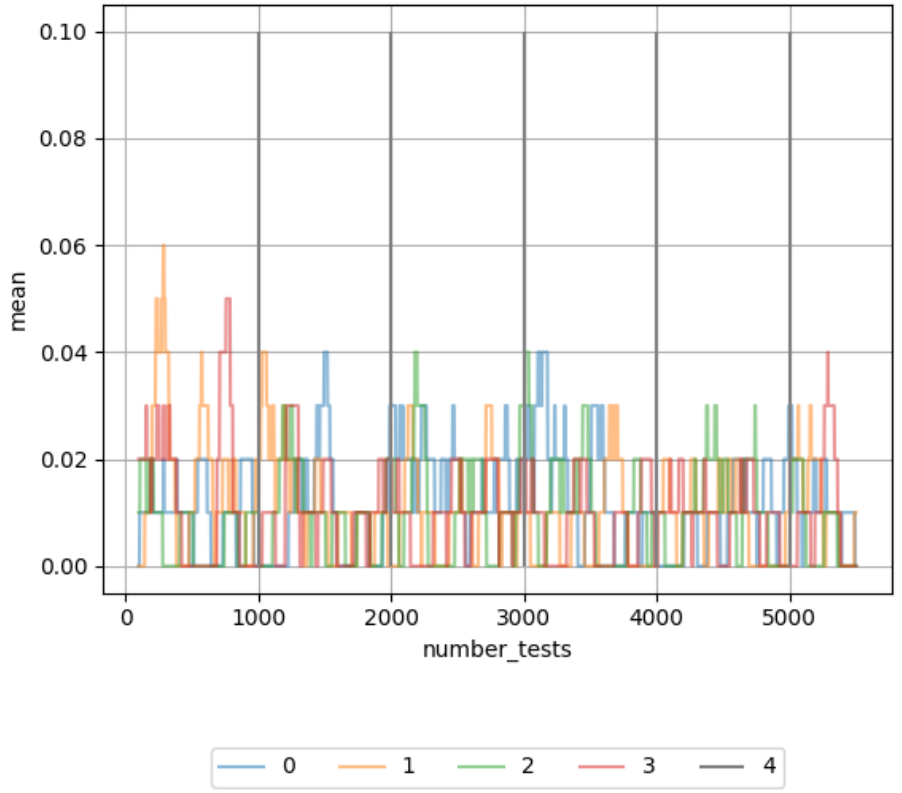
Erkennung von verdächtigen Veränderungen im Verhalten

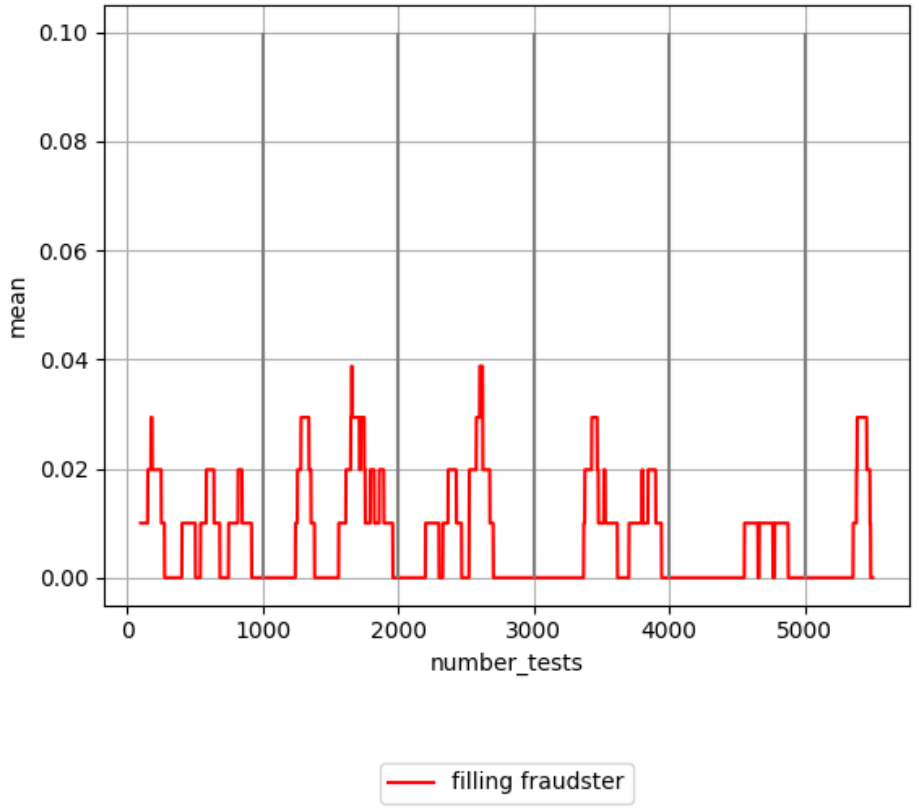
Gleitender Mittelwert

Gegeben:

- Fensterlänge n
- Zeitreihe f'

$$\text{mean}_{\text{MA}}^{\text{n}}(\text{t}) = \frac{1}{\text{n}} \sum_{i=0}^{n-t} f(t-i)$$





Red Flags

Regeln zur Erkennung von verdächtigen Ausreißern

Eine Red Flag ist eine Reihe von Umständen, die ungewöhnlich sind oder von der normalen Aktivität abweichen. Es ist ein Signal, dass etwas ungewöhnlich ist und möglicherweise weiter untersucht werden muss.

Red Flags zeigen weder Schuld oder Unschuld an, sondern stellen lediglich mögliche Warnsignale für Betrug dar.

Regeln:

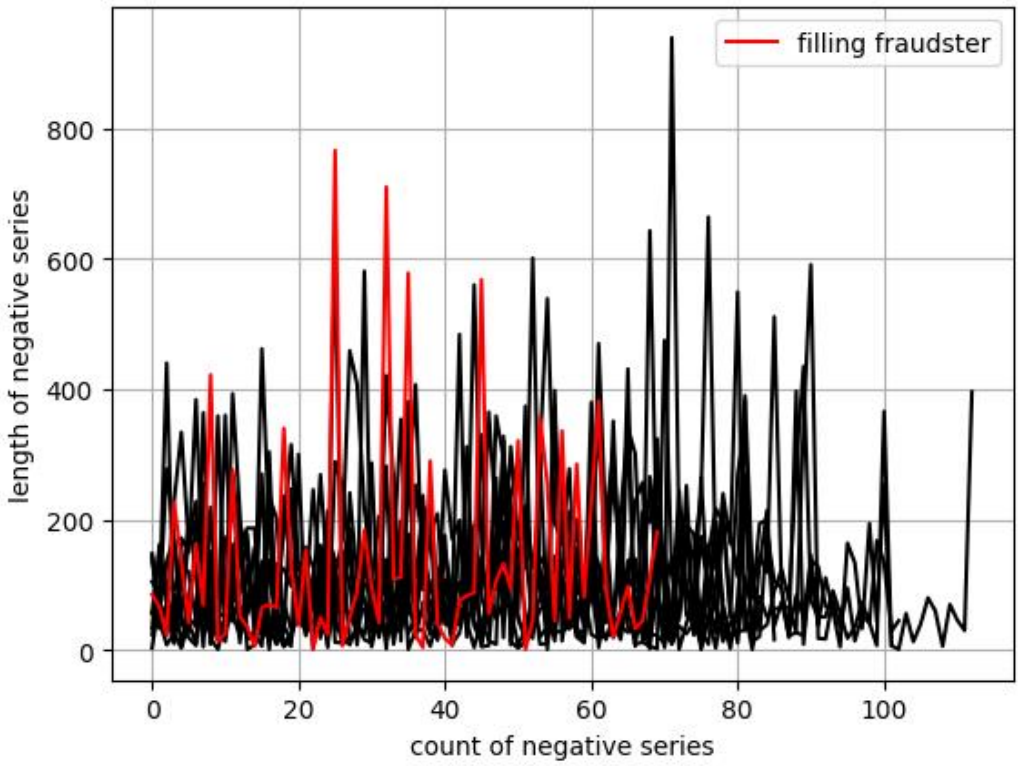
1. „Do not ignore a red flag“
2. “Sometimes an error is just an error”

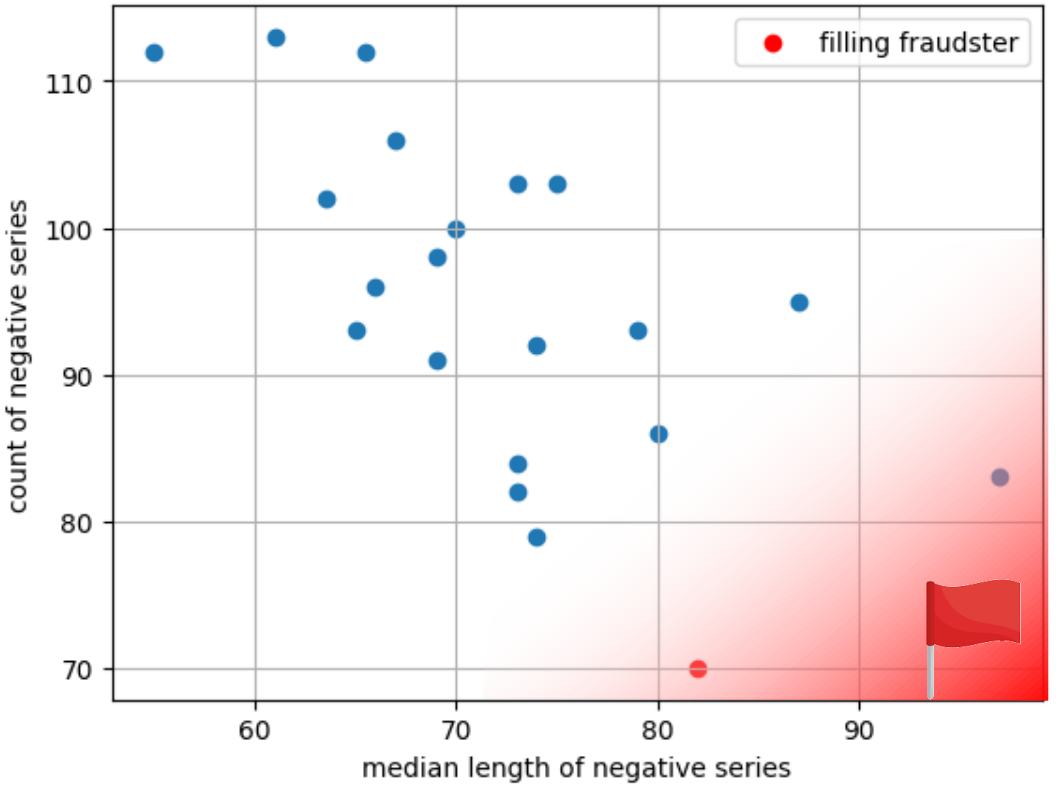
Red Flags:

- Anzahl in einer Serie aufeinander folgender negativer Testresultaten darf nicht zu groß werden.
  - [0,0,0,0,1,0,0,0,0,0,0,0]
  - [4 , 7]
- Anzahl Serien aufeinander folgender negativer Testresultaten darf in einem Zeitintervall nicht zu klein werden.
  - [0,0,0,0,1,0,0,0,0,0,0,0,1,0,0]
  - [1 ,2 ,3]

Red Flags

Regeln zur Erkennung von verdächtigen Ausreißern





Clusteranalysen

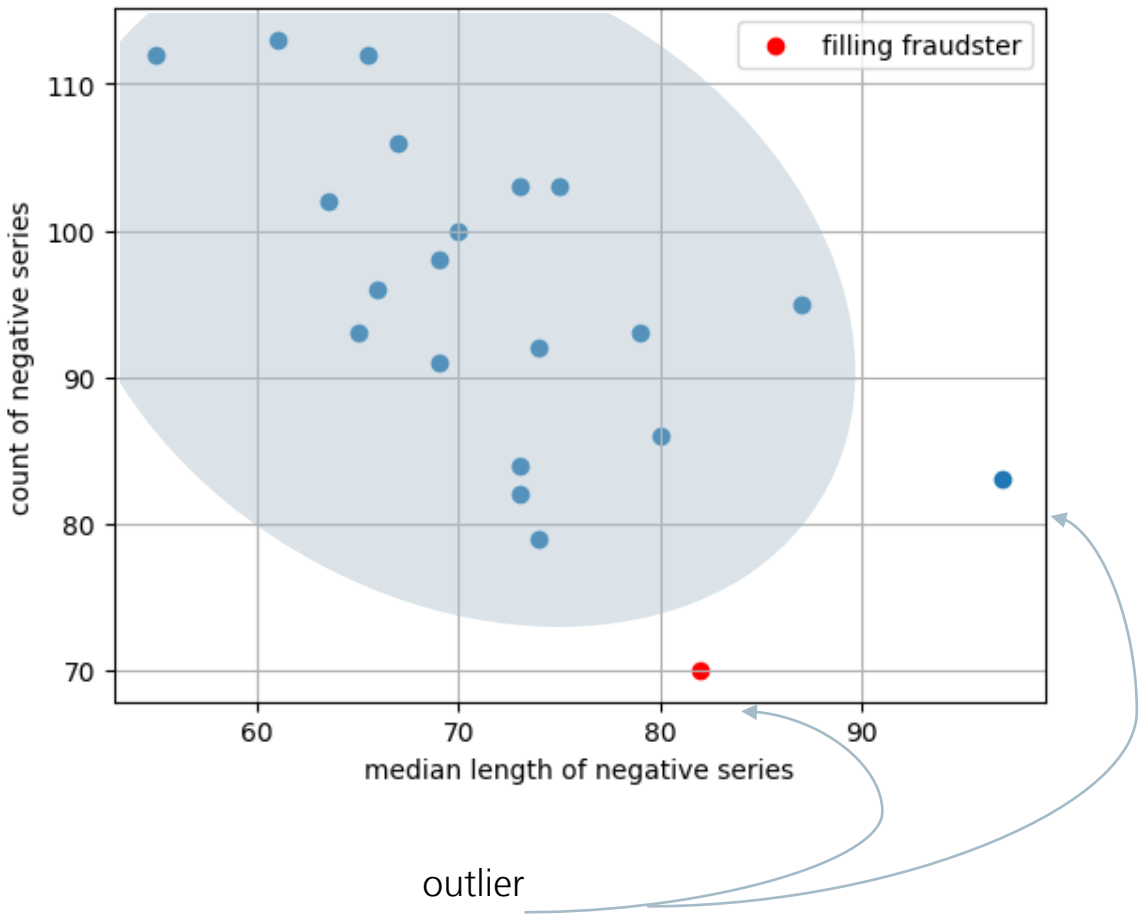
Gruppierung unverdächtiger Beobachtungen

Das Ziel des Clusterings ist es, Beobachtungen in Segmente aufzuteilen, so dass die Homogenität innerhalb des Segments maximiert (kohäsiv) und die Heterogenität zwischen den Segmenten maximiert (Kopplung) wird.

Ein mögliches Ziel des Clusterings in der Betrugserkennung ist es, Anomalien in kleine, spärliche Cluster zu gruppieren.

Typische Verfahren:

- K-Means Clustering
- Hierarchisches Clustering
- DBSCAN



Benford’s Law

Verteilung der führenden Ziffern von Zahlen in empirischen Datensätzen

Das Benfordsche Gesetz beschreibt die Häufigkeitsverteilung der ersten Ziffer in Grundgesamtheiten mit logarithmischen Normalverteilungen

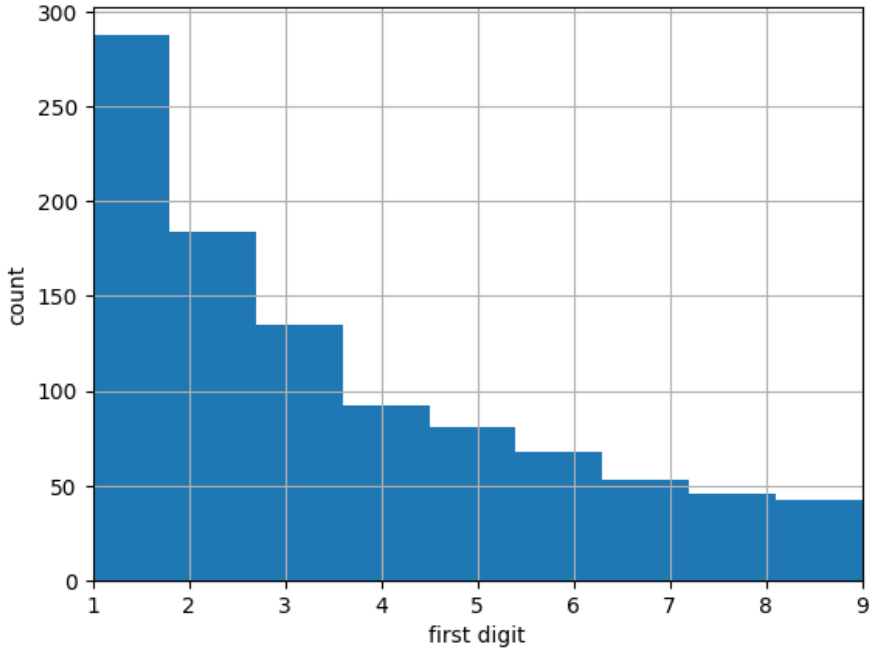
Logarithmische Normalverteilungen treten typischerweise in mechanischen, medizinischen oder ökonomischen Systemen mit exponentiellem Wachstums auf.

Benfordsche Gesetz: $$P(d) = log_{10}(1+\frac{1}{d})$$

Je niedriger der zahlenmäßige Wert einer Ziffernsequenz an einer bestimmten Stelle einer Zahl ist, desto wahrscheinlicher ist ihr Auftreten.

Weichen die Ziffernverteilungen von z.B. Bilanzen von der Benford-Verteilung ab, kann dies als Red Flag definiert werden.

```
benford = [s[0] for s in np.array(
    [
       10**(np.random.normal(0.5,0.5)+np.random.randint(0,20))
           for i in range(1000)
    ]
  ).astype(str)]
```



Betrugserkennung

Prädiktive Analytik

Prädiktive Analytik

Lernt aus den Beobachtungen historischer Betrugsmuster prädiktive Modelle, um zwischen normalen und betrügerischen Verhaltensweisen zu unterscheiden.

- Benötigt historische Beispiele
- Erkennt lediglich bekannte Betrugsmuster
- Robuster gegenüber Täuschung

Vorgestellte Methoden:

1. Modellansatz
2. Lineare Regression
3. Log. Regression

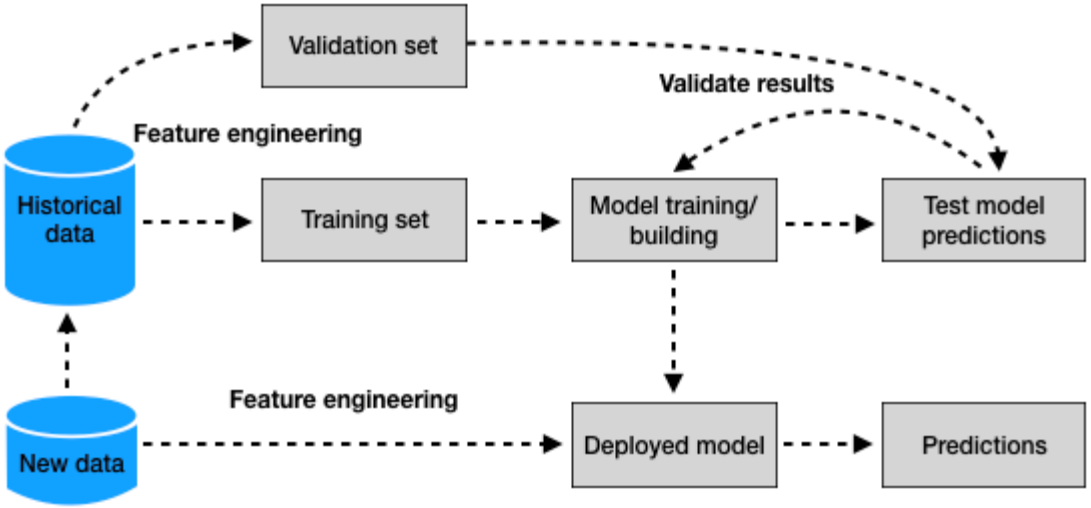
Datenbasierte Modellierung

Allgemeine Vorgehensweise

Modellansatz

Formal ist ein Vorhersagemodell eine parametrische Funktion h, auch Hypothese genannt, die eine Eingabe aus einer Eingabedomäne X nimmt und eine Vorhersage y ausgibt.

- Parameter: θ
- Erklärende Variablen: $X \subset \Re^n$
- Zielvariable: $y \subset \Re$
- Hypothese: $h(x, \theta) : X \rightarrow y$
- Prädiktion: $\hat{y} = h(x, \theta)$



https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_2_Background/MachineLearningForFraudDetection.html#

Nimmt die Zielvariable lediglich binäre Werte (0,1) an, spricht man von Klassifikation, sonst von Regression.

Korrelation und Regression

Korrelation

Die Korrelation R (Korrelationskoeffizient von Pearson) beschreibt die lineare Abhängigkeit zwischen zwei Merkmalen x_i und y_i .

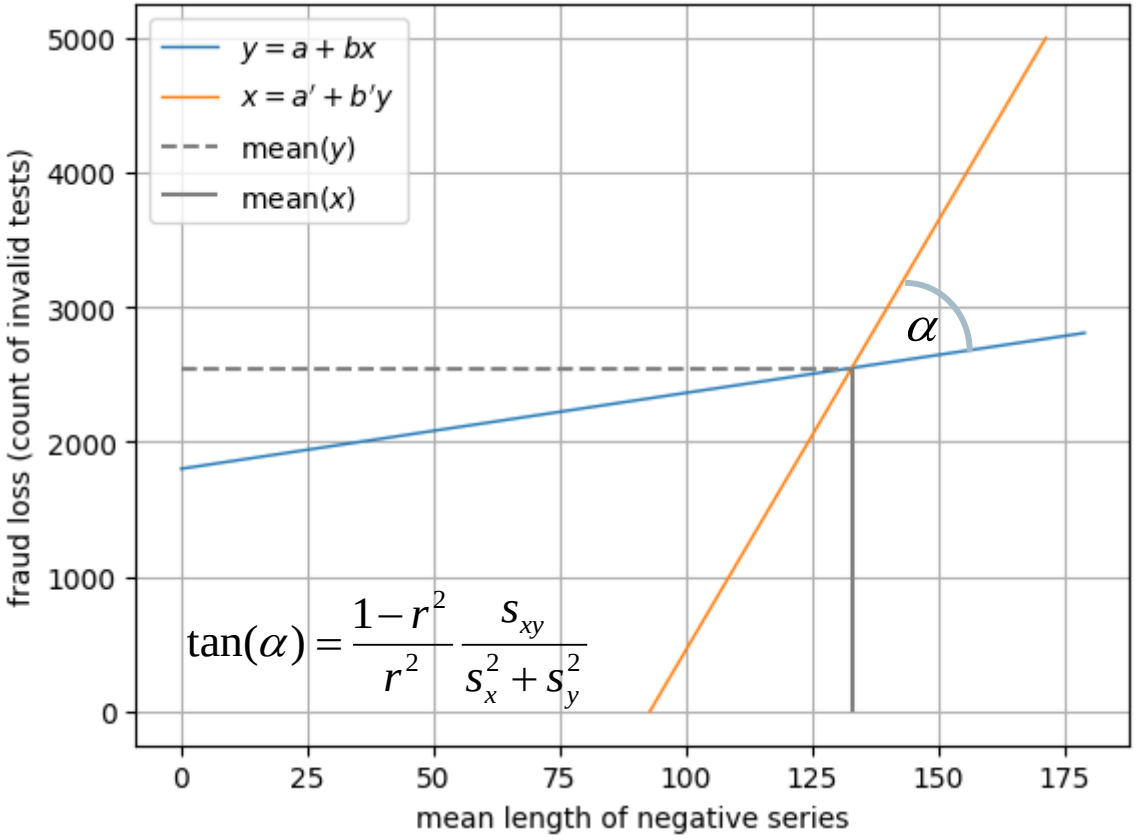
$$R_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Regression

Die Lineare Regression beschreibt die funktionale Beziehung zwischen zwei Merkmalen x_i und y_i über eine Geradengleichung.

$$y_i = a + b_i x_i$$

Achsenabschnitt (Intercept, Offset) a
Regressionskoeffizient (Slope, Steigung) b_i



Klassifikation

Klassifikation über Logistische Regression

Die Logistische Regression beschränkt die funktionale Beziehung zwischen den Merkmalen x_i und y_i über die logistische Funktion.

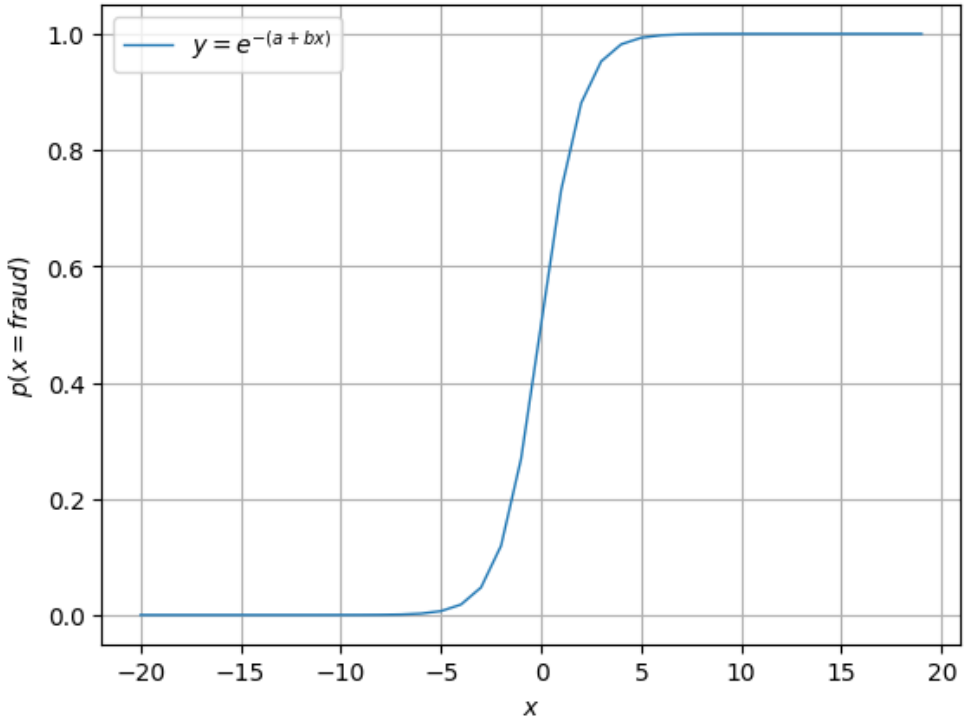
$$y_{(x=fraud)} = 1/(1+e^{-a+b_i x_i})$$

Achsenabschnitt (Intercept, Offset) a
Regressionskoeffizient (Slope, Steigung) b_i

Linearer Klassifikator

Die logistische Regression beschreibt einen linearen Klassifikator, da :

$$\log it(y_{(x=fraud)}) = \log(y_{(x=fraud)}/(1-y_{(x=fraud)}) = a + b_i x_i$$



Logistische Funktion

Modellansätze

Betrugserkennung

Vorhersage des potentiellen Verlusts

Wenn zu einer Beobachtung X eine Wahrscheinlichkeit P, dass X ein Betrugsfall ist und eine Verlust L durch den Betrug geschätzt werden kann, dann ist E der erwarteter Verlust durch Betrug:

$$E(x) = P(x = \text{Fraud}) \times L(x = \text{Fraud})$$

Schätzung der potentiellen Verlusts durch betrügerisches Potential in den Beobachtungen.

Modellansätze

$L(x)$ Schätzung der Höhe des Verlusts durch Regression

$P(x)$ Erkennung von Betrug duch Klassifikationsansatz

Historische Daten

Aufbau eines Trainingsdatensatzes

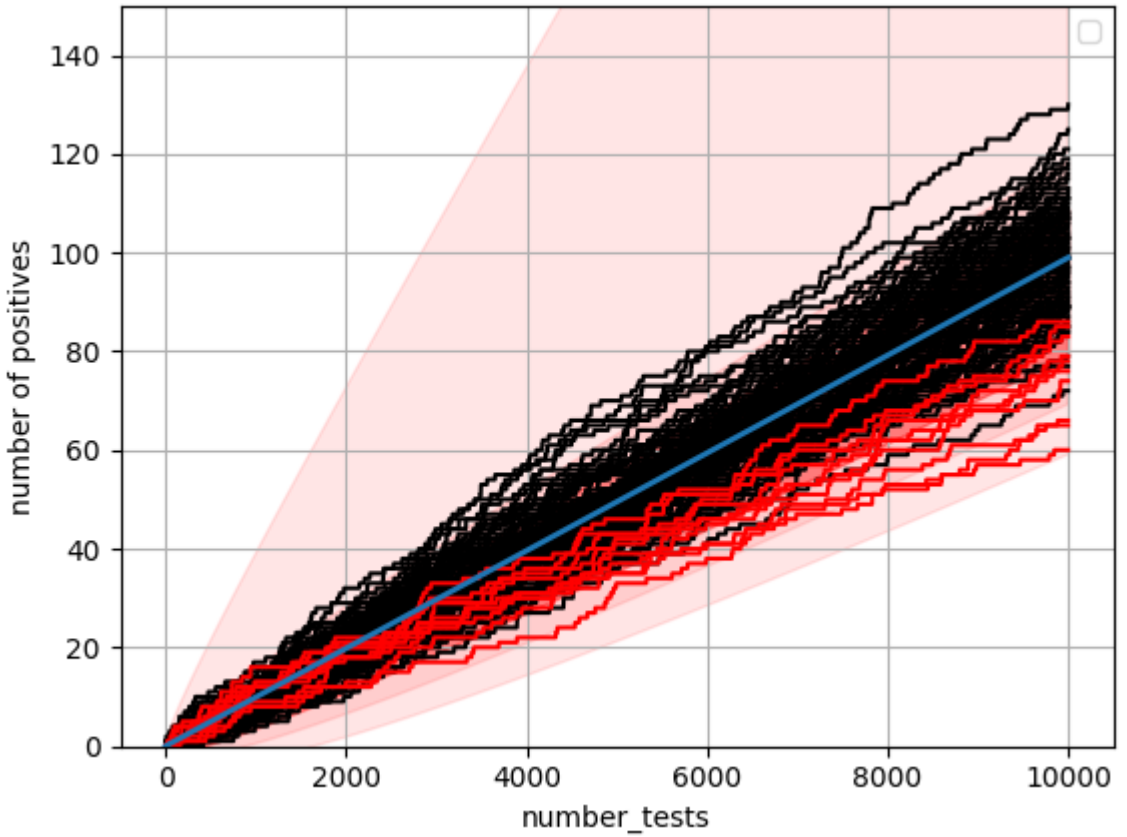
Weitere Trainingsdaten:

100 ehrliche Teststationen
Grundgesamtheit: 100000 Tests

10 bekannte betrügerische Teststationen
Nach 1000 Tests werden 210-300 erfundene Tests mit negativen Ergebnissen dazu gemischt, danach werdenwieder 700-790 Personen richtig getestet.

Höhe des Verlusts über Anzahl erfundener Tests:

L = [2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000]



Aufbau eines Trainingsdatensatzes

Merkmalskonstruktion

Die Konstruktion der Red Flags zeigte zwei Merkmale mit hinreichend guter Beschreibung zur Erkennung von Betrügern.

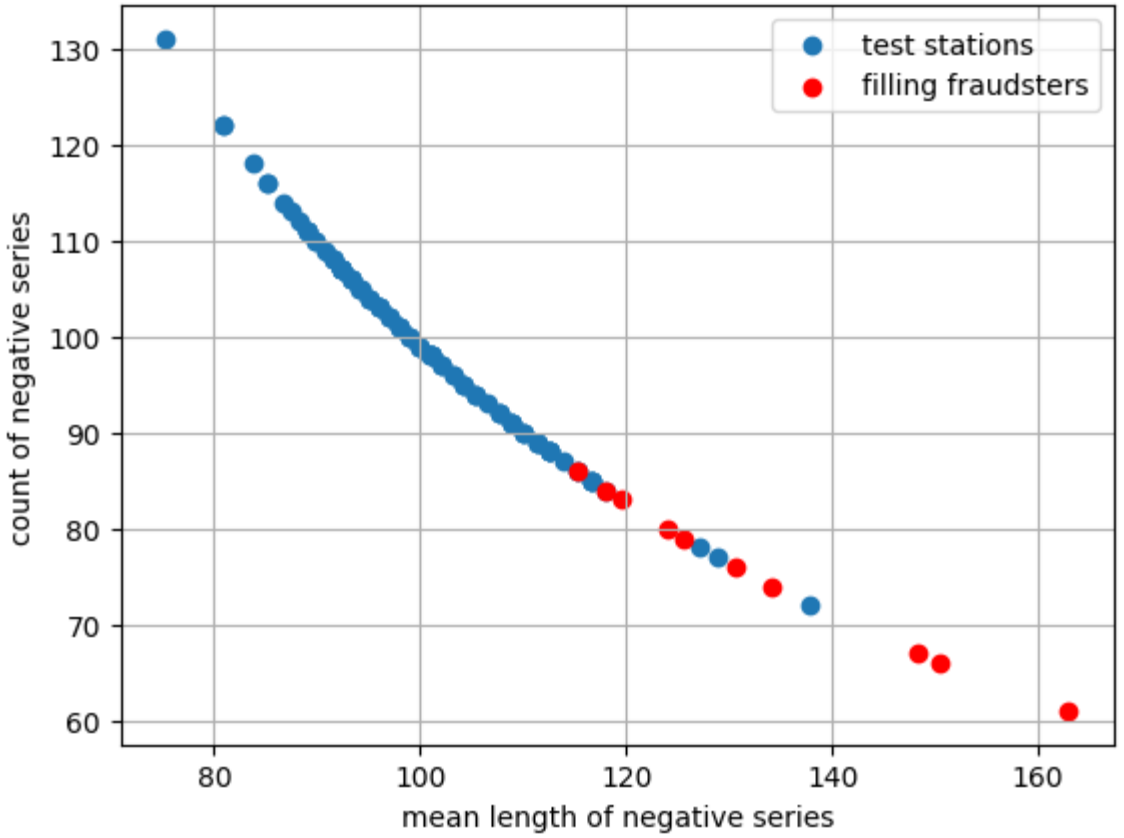
- Anzahl der Serien mit negative Testergebnissen
- Mittlere Länge der negative Serien

Betrachtung der Abhängigkeitsstruktur

correlation	*mean length*	*count*	*loss*
mean length	1.0	-0.996	0.3
count	-0.996	1.0	-0.27
loss	0.3	-0.27	1.0

Achtung! Die Merkmale „mean length“ und „count“ sind linear voneinander abhängig.

Es liegt eine Ko-Linearität vor. Beide Merkmale beeinflussen sich gegenseitig.



Aufbau eines Trainingsdatensatzes

Merkmalskonstruktion

Die Konstruktion der Red Flags zeigte zwei Merkmale mit hinreichend guter Beschreibung Erkennung von Betrügern.

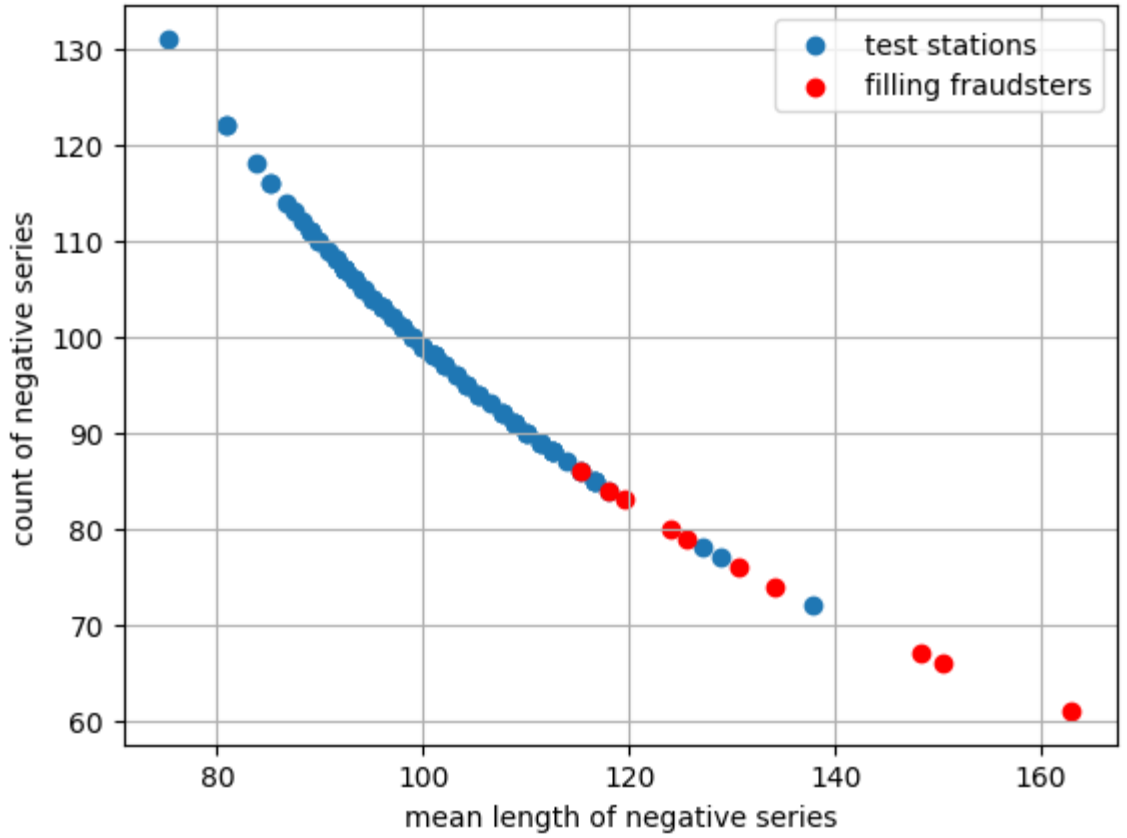
- Anzahl der Serien mit negative Testergebnissen
- Mittlere Länge der negative Serien

Modellierung des erwarteten Verlusts durch Betrugspotential:

$$E(x) = P(x = \text{Fraud}) \times L(x = \text{Fraud})$$

L: Regression der Verlusthöhe durch x als Mittlere Länge der negative Serien

P Klassifikation durch
x1: Mittlere Länge der negative Serien
x2: Anzahl der Serien mit negative Testergebnissen



Training

Regression der Verlusthöhe

Regressionsgleichung der Verlusthöhe L durch x als mittlere Länge der negative Serien:

$$\text{loss} \qquad L(x) = a + bx^{\text{mean length}}$$

Training (Schätzung der Parameter a,b) erfolgt mit der Methode der kleinsten Quadrate (ordinary least squares, OLS) über die Minimierung der Summe aller Fehlerquadrate.

$$:\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y_i\right)^2$$

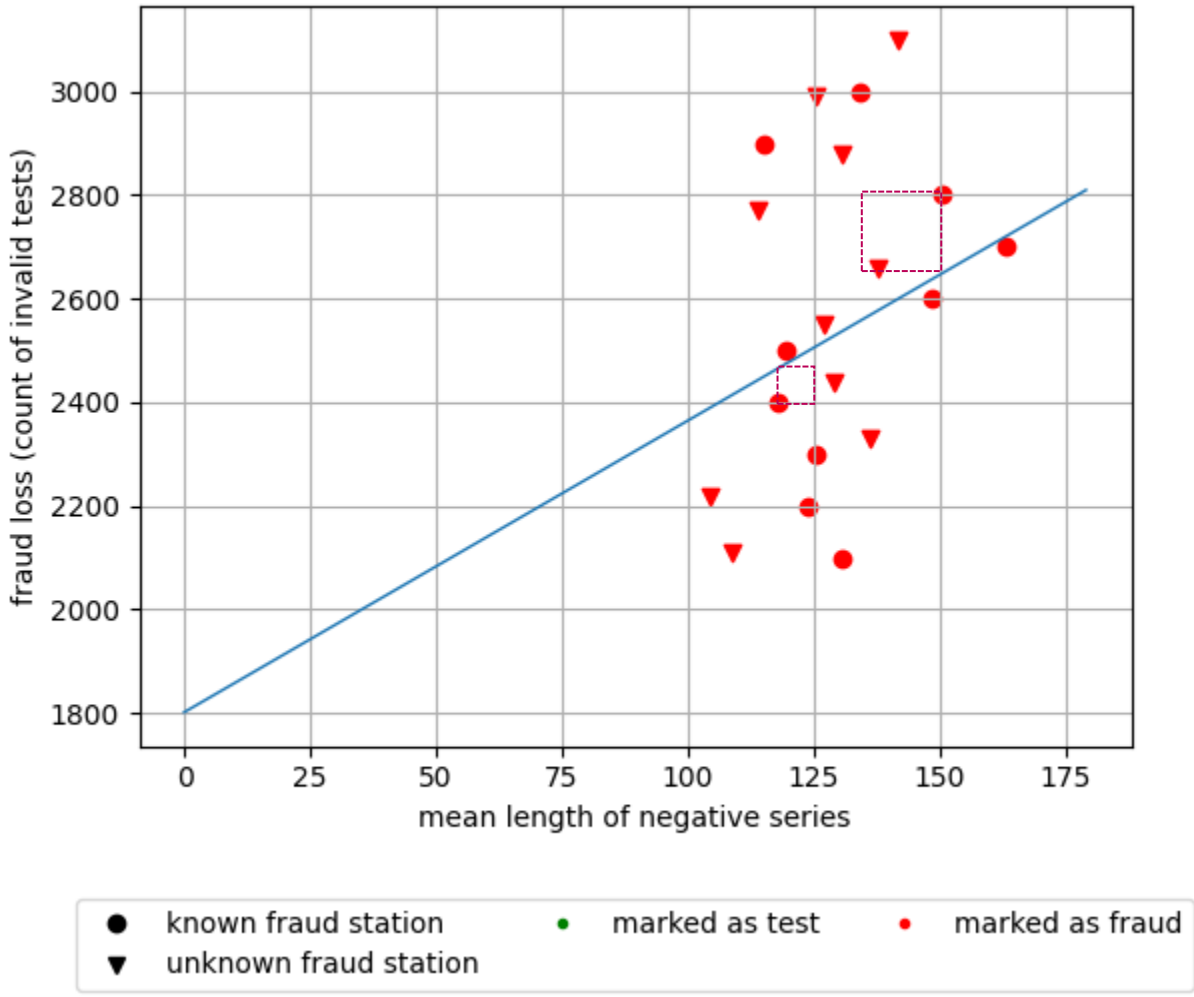
Resultat

intercept $a = 1801.8$

slope $b = 5.6$

$\sqrt{\text{MSE}}$(Training) = 274.2

$\sqrt{\text{MSE}}$(Test) = 299.7



Training

Klassifikation der Betrugserkennung

Klassifikation der Betrugswahrscheinlichkeit durch x als mittlere Länge der negative Serien:

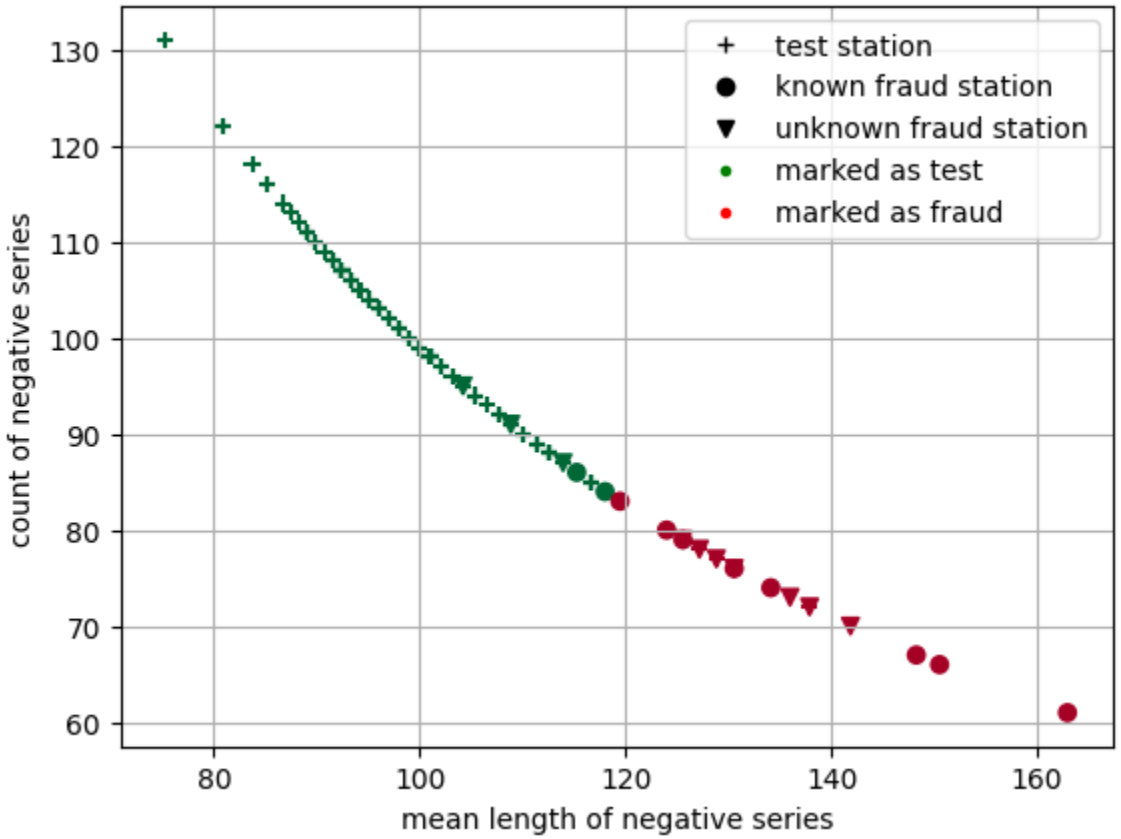
$$P_{(x=fraud)} = 1/(1+e^{-a+b_1x_1+b_2x_2})$$

Training (Schätzung der Parameter a, b_1, b_2) erfolgt mit der Methode der Maximum Likelihood.

intercept	$a = -0.00057$
mean length	$b_1 = 0.14$
count	$b_2 = -0.2$

Genauigkeiten

- Gut = 0.97
- Bekannte Betrüger = 0.7
- Unbekannte Betrüger = 0.8



Anwendung

Erkennung und Bewertung von Betrug in Corona-Teststationen

Schätzung des erwarteten Verlusts durch Betrugspotential in beobachteten Teststationen:

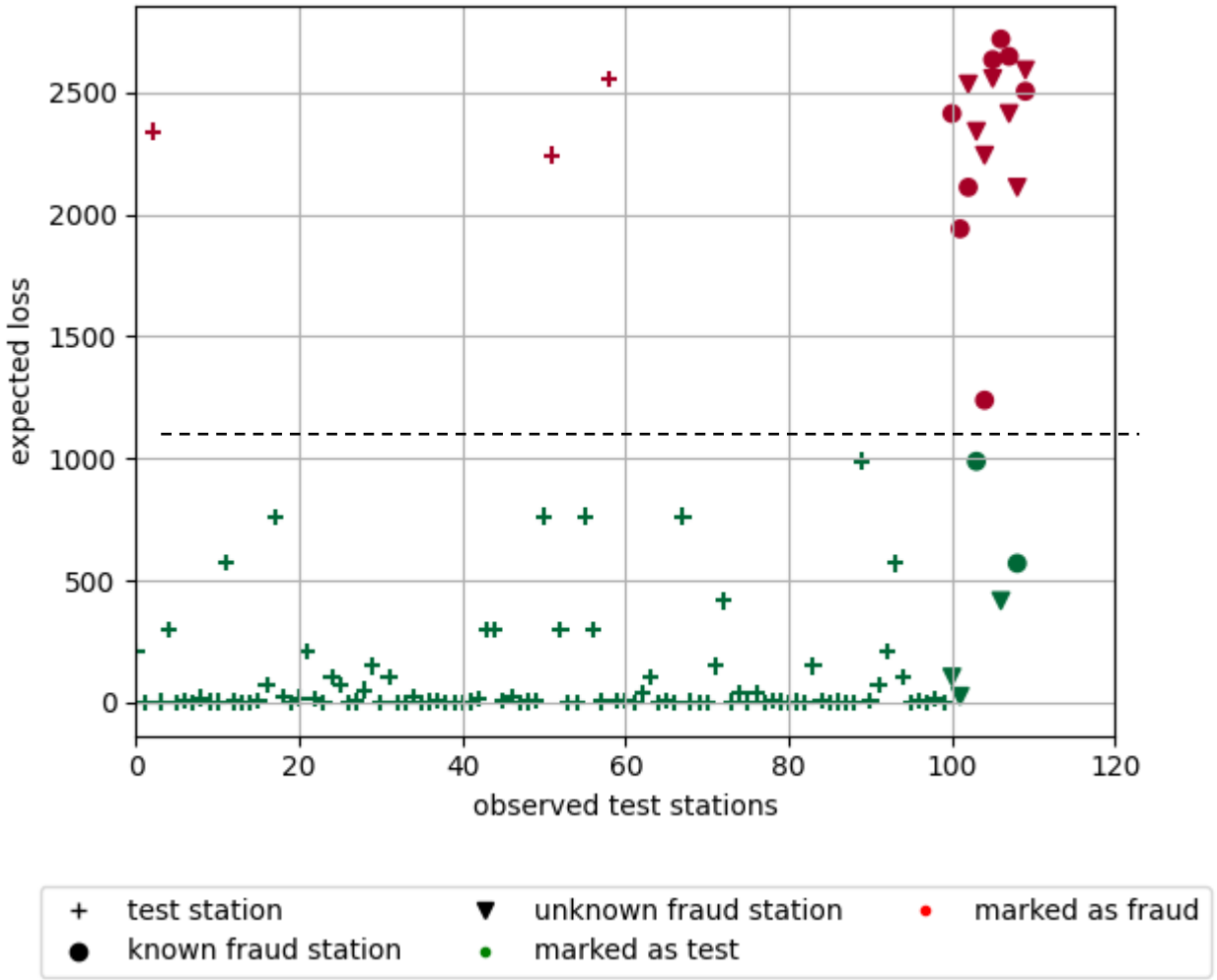
$$E(x) = P(x = \text{Fraud}) \times L(x = \text{Fraud})$$

Erkenntnisse eines Testlaufs:

- Nicht alle betrügerischen Stationen werden erkannt
- Nicht alle nicht betrügerischen Stationen werden erkannt

Nächste Schritte

- Hinzugabe weiterer Trainingsdaten
- Modellierung besserer Merkmale
- Evaluation komplexerer Modellklassen

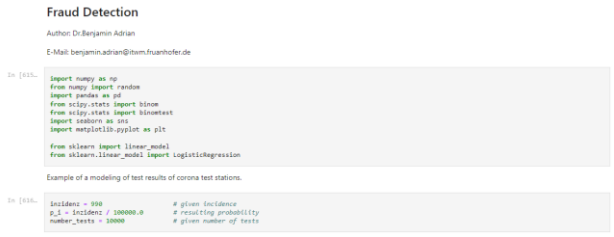


Literatur

Quellen und weitere Ressourcen

Interaktive Codebase

https://github.com/benjamin-adrian/fraud_detection_example/blob/main/Fraud_Workbench.ipynb


Fraud Detection

Author: Dr.Benjamin Adrian

E-Mail: benjamin.adrian@iitwm.fruanhofer.de

```
import numpy as np
from numpy import random
import pandas as pd
from scipy.stats import binom
from scipy.stats import binomtest
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn import linear_model
from sklearn.linear_model import LogisticRegression
```

Example of a modeling of test results of corona test stations.

```
Inzidenz = 990                  # given incidence
p_I = Inzidenz / 100000.0       # resulting probability
number_tests = 10000            # given number of tests
```

Weiterer Datensatz

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Weiterer Fraud-Simulator

https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook

Literatur

- Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook; Le Borgne, Yann-Ael and Siblini, Wissam and Lebichot, Bertrand and Bontempi, Gianluca, https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook, 2022, Université Libre de Bruxelles
- Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques : A Guide to Data Science for Fraud Detection; Baesens, Bart and Van Vlasselaer, Veronique and Verbeke, Wouter, ISBN 10: 1119133122 / ISBN 13: 9781119133124, 2015, Wiley
- Angewandte Statistik: Methodensammlung mit R. Sachs, Lothar and Hedderich, Jürgen; ISBN 10: 9783540321613 / ISBN 13:3540321616Springer Berlin Heidelberg, 2006.
- Fraud and Fraud Detection, + Website: A Data Analytics Approach; Gee, Sunder, 2014; Wiley

Kontakt

Dr. Benjamin Adrian
Systemanalyse, Prognose und Regelung
benjamin.adrian@itwm.fraunhofer.de

Fraunhofer ITWM
Fraunhofer-Platz 1
67663 Kaiserslautern
www.itwm.fraunhofer.de

Vielen Dank für Ihre Aufmerksamkeit