Maximum Likelihood Modellierung von Zähldaten

Student's Jour Fixe

Martin Schmettow

Lst. WI2, Uni Passau

Student's Jour Fixe, 23. Juni 2008

Gliederung

Beispiel: Mehrfache Ziehung aus einer Urne

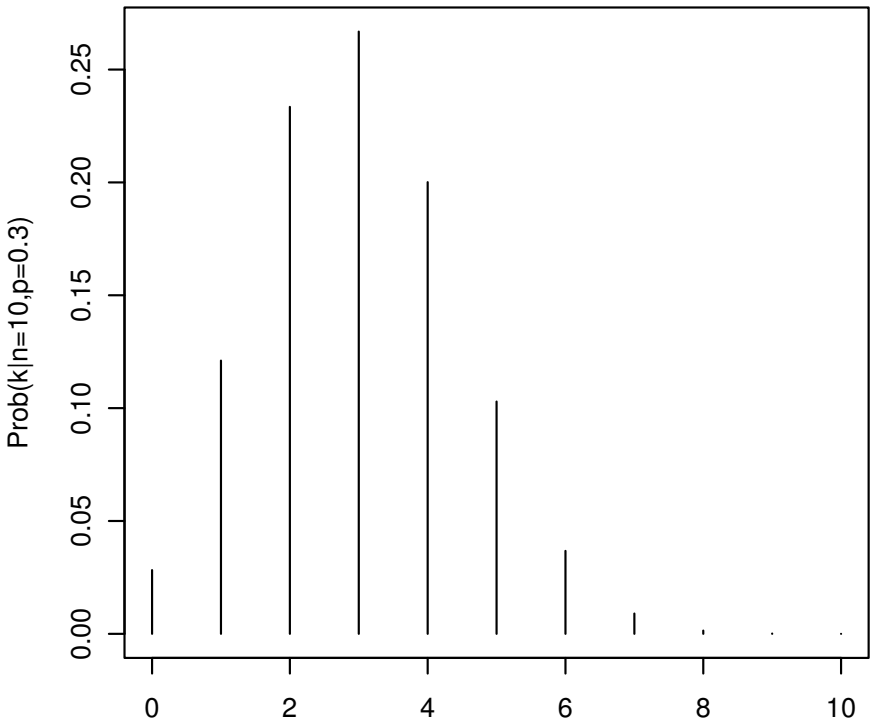
- Unsere Urne: n Kugeln, davon s schwarze
- Ziehen mit Zurücklegen
- Wie hoch ist die Wahrscheinlichkeit bei z Ziehungen x schwarze Kugeln zu ziehen?

Die Bernoulli Verteilung

- Parameter: $p = {}^{n}/_{k}$
- Bei einer Ziehung gilt:
  - Prob("schwarze Kugel"$|p) = p$
  - Prob("weisse Kugel"$|p) = 1 - p$

Die Binomialverteilung

- Bei n Ziehungen gilt:
  Prob("genau k schwarze Kugeln"$|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\text{Prob}(k = 5|n = 10, p = 0.3) = 0.10$



Welche Urne war es?

Angenommen wir haben drei Urnen:

- Urne 1: $p_1 = 0.3$
- Urne 2: $p_2 = 0.5$
- Urne 3: $p_3 = 0.7$

Und wir haben ein Ergebnis $k = 4$ nach $n = 10$ Ziehungen
Frage: Aus welcher Urne wurde gezogen?

Die Likelihoodfunktion

Bisher hatten wir die Wahrscheinlichkeitsfunktion Prob($k|n,p$)
Oder allgemeiner: Prob(Daten|Parameter)
Die Likelihoodfunktion sucht nach dem *Mutmaßlichkeit des Parameter* unter der Bedingung beobachteter Daten

$$L(\text{Parameter}) = \text{Prob}(\text{Parameter}|\text{Daten})$$

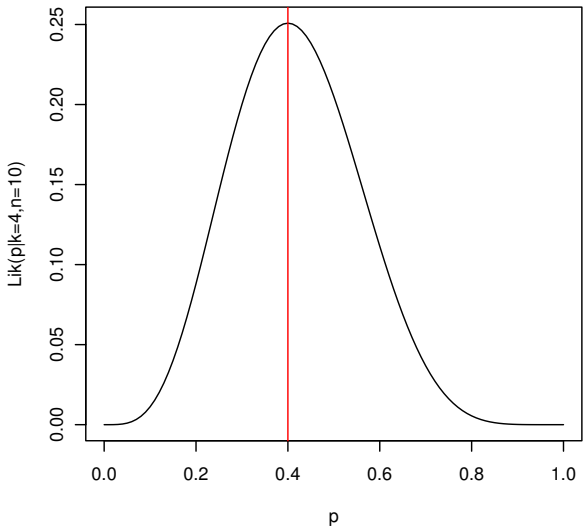
Es war Urne 2!

Welche Urne war es?

- $L(p_1 = 0.3|k = 4, n = 10) = 0.20$
- $\underline{L(p_2 = 0.5|k = 4, n = 10) = 0.21}$
- $L(p_3 = 0.7|k = 4, n = 10) = 0.04$

Die Binomiale Likelihoodfunktion

Wir betrachten zunächst den Fall eines einzelnen Datensatzes $k = 4$ und plotten die Likelihoodfunktion $L(p|4,10) = \binom{10}{4} p^4 (1-p)^{10-4}$



$\max L(p,4,10) = 0.4$
Höchstwahrscheinlich stammt das Ergebnis aus einer Urne mit $p = 0.4$

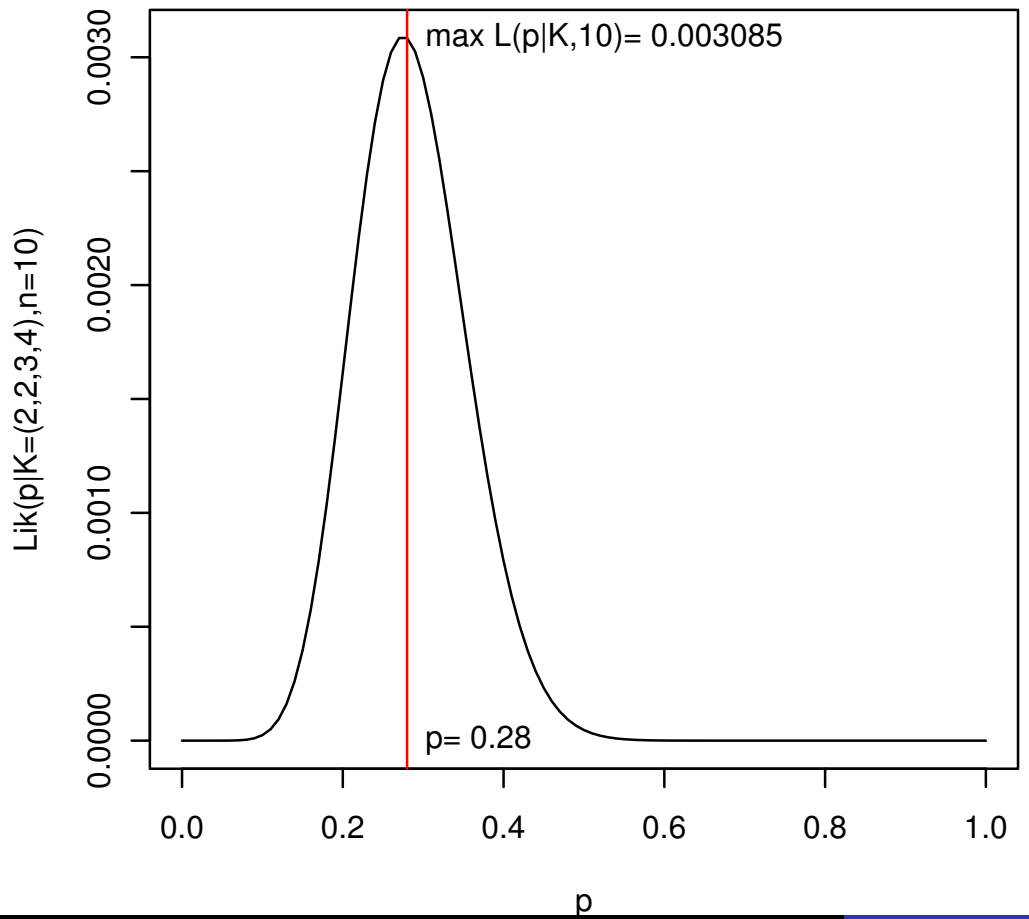
Erweiterung auf Stichproben N>1

- Bisher betrachtet: Stichprobe $N = 1$ $(k = 4)$
- Frage: Was passiert bei $N > 1$? Z.B. $K = [2, 2, 3, 4]$
- Antwort: Unabhängige Ziehungen, daher Multiplikation

$$L(p|K = (x_1, \ldots, x_N), n) = \prod_{i=1}^{N} L(p|x_i, n)$$

Beispiel: ML Schätzung von K=[2,2,3,4]

$$L(p|K=(2,2,3,4),10) = L(p|2,10) \cdot L(p|2,10) \cdot L(p|3,10) \cdot L(p|4,10)$$



Zwischenstand ML

- wir haben eine Stichprobe gezogen
- ... wollen einen Parameter der Population schätzen
- ... setzen eine bestimmte Verteilung (allgemeiner: Modell) voraus
- ... leiten die Likelihoodfunktion aus der Dichtefunktion her
- ... maximieren die Likelihoodfunktion
- schauen bei welchem Parameterwert das Maximum liegt. Voila!

Weiterführende ML Konzepte

- Ableitbarkeit von L am Maximum
- Fisher Information zur Verlässlichkeit des ML Schätzers (Konfidenzintervalle)
- Erwartungsuntreue von ML-Schätzern

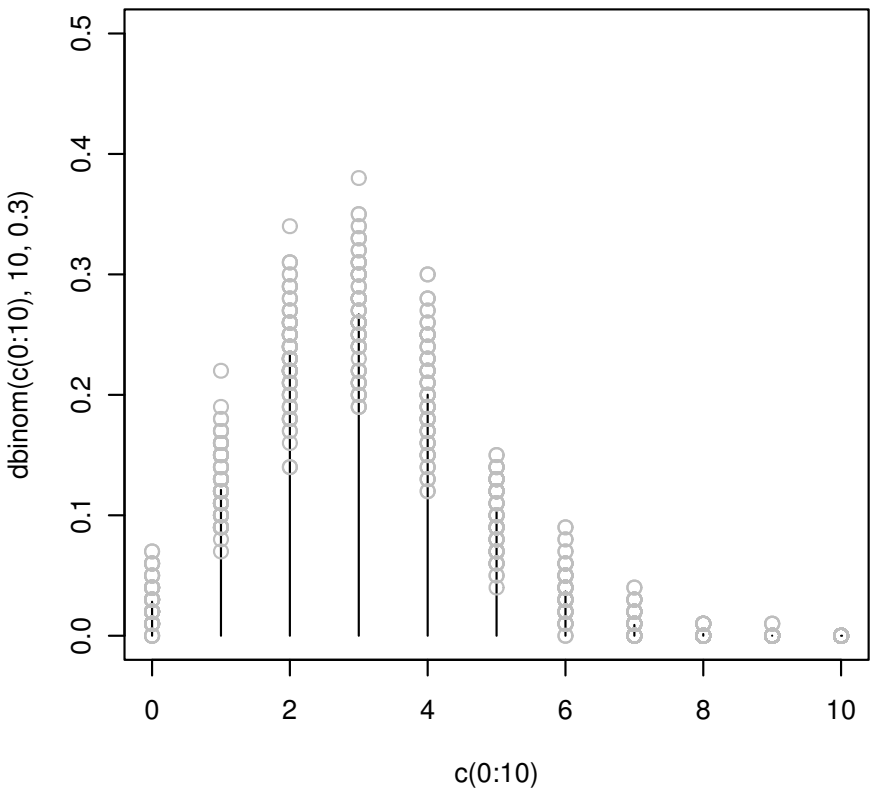
ML einiger bekannter Verteilungen

- Binomialverteilung: $\max L(p|D,n) = E(D)/n$
- Normalverteilung: $\max L(\mu|D) = E(D)$

Ist das immer so einfach? Nein! Sehen wir später!

Warum nicht Least Square Schätzung?

- Ist nur bei konstantem Fehlerterm erlaubt
- ist z.B. bei der Binomialverteilung nicht der Fall



Kann man mehrere Parameter gleichzeitig schätzen?

Klar! Bei zwei Parametern erhält man eine Likelihood Fläche.
Es könne auch mehr sein.

Was ist die Log Likelihoodfunktion?

- Wir suchen ja immer das Maximum der Likelihoodfunktion.
- Da der Logarithmus eine streng monoton steigende Transformation ist, gilt:
  $f(a) > f(b) \rightarrow \log f(a) > \log f(b)$
  $\max_x f(x) = \max_x \log f(x)$
- Durch Logarithmierung werden Produkte zu Summen
  $\log(L(\cdot|x_1)L(\cdot|x_2)) = \log L(\cdot|x_1) + \log L(\cdot|x_2)$
- Das vereinfacht oft die Berechnung

Beispiel: Zeckenpersönlichkeit

- eine zehntägige Kinderfreizeit im Wald ($n = 10$)
- jeden Abend Zeckencheck an allen $N = 20$ Kindern
- Wir zählen: Wieviele Kinder hatten an k Tagen Zecken

Zeckentage	0	1	2	3	4	5	6	7	8	9	10
Häufigkeit	5	4	6	3	1	1	0	0	0	0	0

Frage: Haben alle Kinder dieselbe W'keit p sich eine Zecke zu fangen?

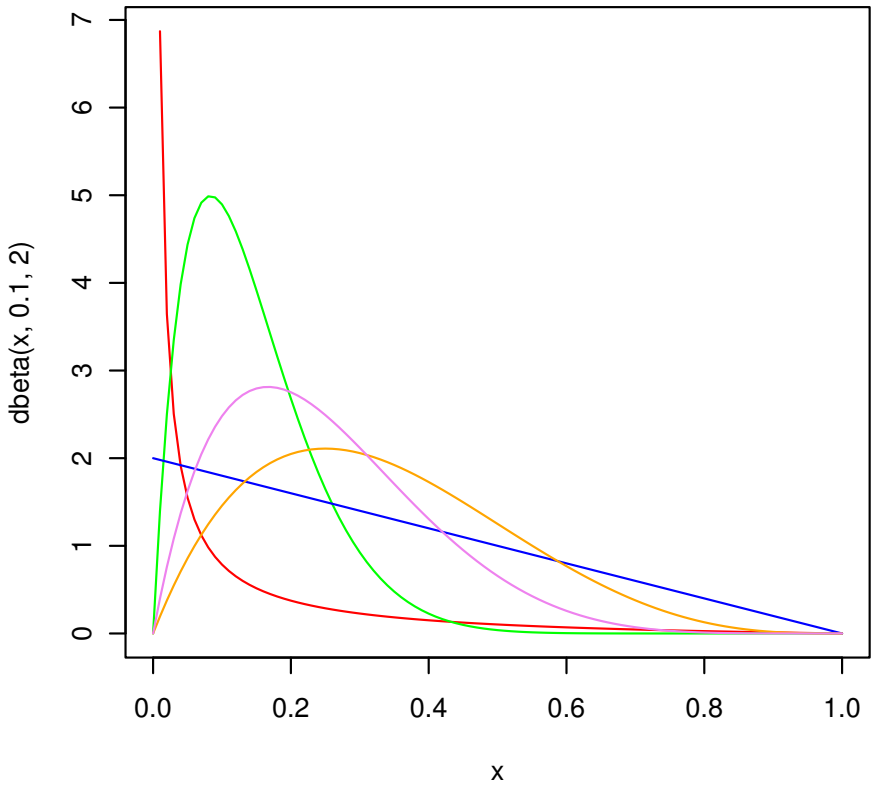
Oder: Mögen Zecken manche Kinder mehr als andere?

Zwei Modelle

1. Modell "homogenes p": Binomialverteilung $\text{Bin}(k|n = 10, p)$
2. Modell "heterogenes p"
   - p ist selbst eine Zufallsvariable
   - Gesucht: Eine Wahrscheinlichkeitsverteilung für p $[0;1] \rightarrow [0;1]$
   - Gefunden: Die Beta Verteilung

Die Beta Verteilung

- stetig in [0;1]
- zwei Parameter: a,b
- kann eine Vielzahl von Formen annehmen:
  - Gerade, konvex, konkav



Die Beta-Binomial Verteilung

- p ist Beta(a, b)verteilt
- Varianz der Binomialverteilung: $np(1-p)$
- Varianz der Betabinomialverteilung ist *größer*

ML Schätzer für die Betabinomiale Verteilung

- Wir nehmen die Dichtefunktion dbetabin$(k|n, a, b)$
- Die Likelihoodfunktion ist
  $L(a, b|D, n) = \prod_{i=1}^{N} \text{dbetabin}(d_i|n, a, b)$
- Brute Force: Wir lassen L über a und b laufen und suchen $\max L$

In R

```
lik.betabin<-function(n,K,a,b) prod(dbetabin.ab(K,n,a,b))
```

Numerische Optimierung

- Probleme: Brute Force dauert lange bei
  - unendlichen Bereichen für Parameter
  - bei vielen Parametern
- Lösung: Intelligente Suchalgorithmen
  - in R: optim()
  - Achtung: Es wird das *Minimum* gesucht, also $-\log L$

In R

```
fitbb.zecken<-optim( fn=nloglik.betabin, par=c(0,0), K=zecken,
n=10, lower=c(0.1), method="L-BFGS-B" )
```

Lösung des Zeckenbeispiels

Betabinomial			Binomial	
a	b	max *logL*	p	max *logL*
3.7	17.9	-33.26	0.17	-33.88

- Der max L der betabinomialen Schätzung ist größer
- Vorläufiger Schluss:
  - betabinomiales Modell erklärt besser
  - Es gibt die Zeckenpersönlichkeit

Occam's Razor

- Das Prinzip der Sparsamkeit: Die einfachere Theorie ist bei gleichem Erklärungswert zu bevorzugen
- Bei statistischen Modellen: Je weniger Parameter, desto sparsamer

Akaikes Information Criterion

- Bestrafung von Parameteranzahl k
- $AIC_c = -2 \log L_{max}(x_1 \dots x_k) + 2k + \frac{2k(k+1)}{n-k-1}$
- der kleinere AIC gewinnt
- Zusatzbemerkungen:
  - der AIC wirkt etwas beliebig, ist aber theoretisch fundiert (Informationstheorie)
  - oft sind Differenzen zwischen AICs sehr klein, das kann aber nicht interpretiert werden
  - bei linearen Modellen kann der AIC auch aus den Residuen berechnet werden (RSS)

AIC am Zeckenbeispiel

	Parameter	Stichprobe	$\log L_{max}$	AIC_c
Binomial	1	20	-33.26	69.998
Betabinomial	2	20	-33.88	71.235

- Der binomiale AIC ist kleiner
- Es gibt keine Zeckenpersönlichkeit

Ein zweiter Zeckendatensatz

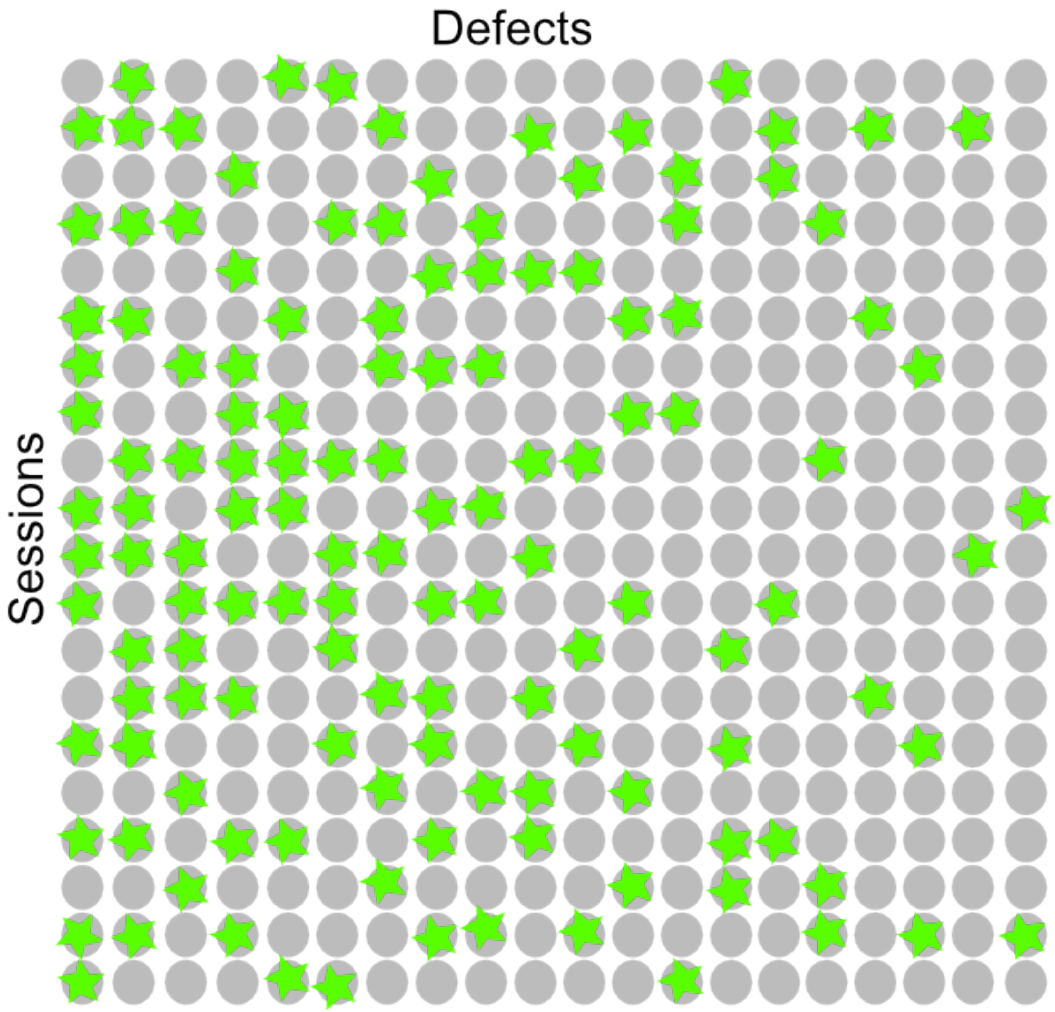
- $K \sim \text{Betabin}(10, 2, 4), N = 20$

Zeckentage	0	1	2	3	4	5	6	7	8	9	10
Häufigkeit	2	6	3	6	3	1	2	3	0	0	0

- Ergebnis: Es gibt eine Zeckenpersönlichkeit

	Parameter	Stichprobe	$\log L_{max}$	AIC_c
Binomial	$p = 0.34$	20	-43.92	90.061
Betabinomial	$a = 3.52; b = 6.85$	20	-41.60	87.911

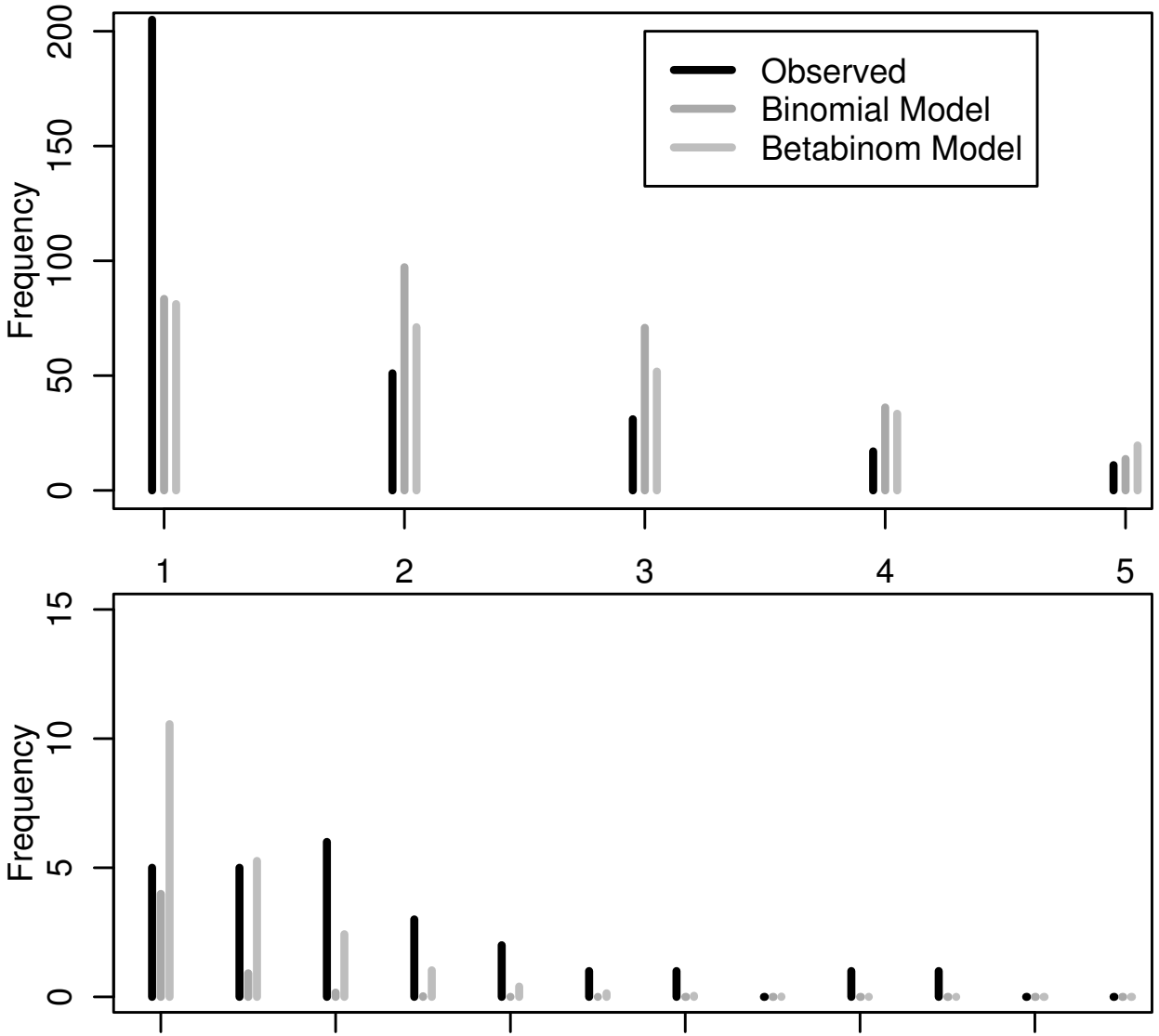
Entdeckung von Schwachstellen



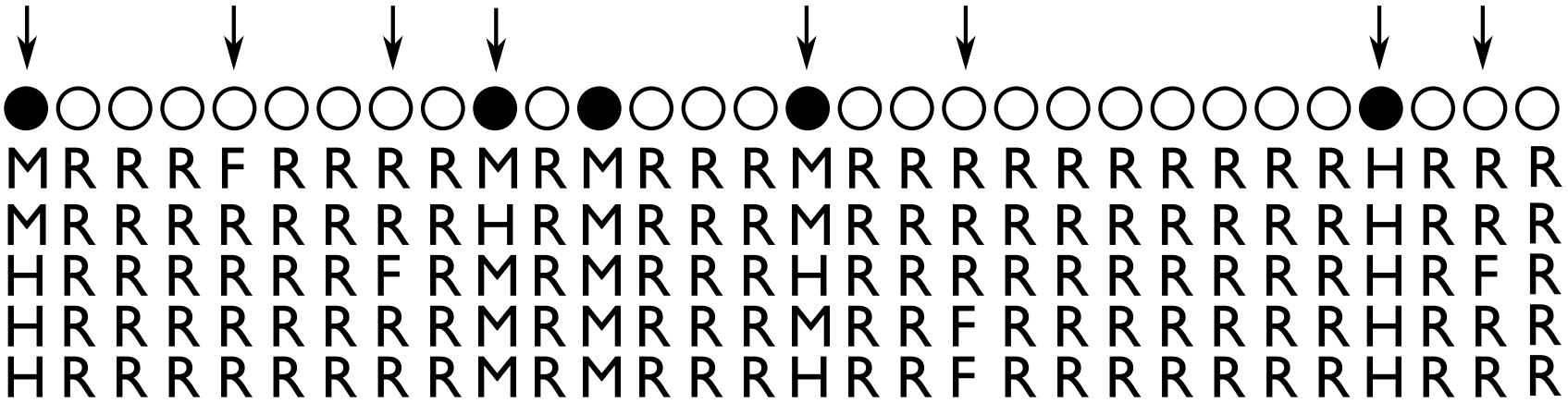
Heterogenität im Prozess

- Sind die Schwachstellen unterschiedlich schwierig zu finden?
- Überprüfung mit dem Modellvergleich Binomial vs. Betabinomial
- Ergebnis: Immer passt das betabinomiale Modell besser
- Schlussfolgerung: Es gibt Unterschiede in der Schwierigkeit

Eine Beobachtung im CUE-4 Datensatz

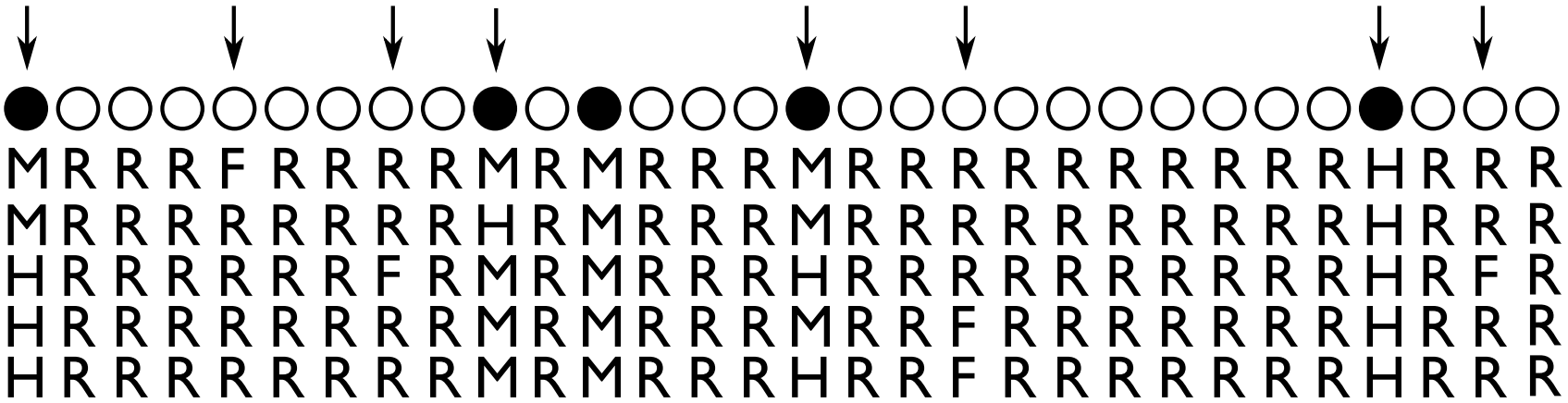


Das Konstrukt der Inspektionspunkte



- Der Entdeckungsprozess als “Abklappern” von Inspektionspunkten
- zwei Klassen von Inspektionspunkten:
  - Schwachstellen •
  - unverdächtige Punkte ○

Signalentdeckung



- korrekte Erkennung H
- Übersehen M
- falscher Alarm F
- korrekte Ablehnung R

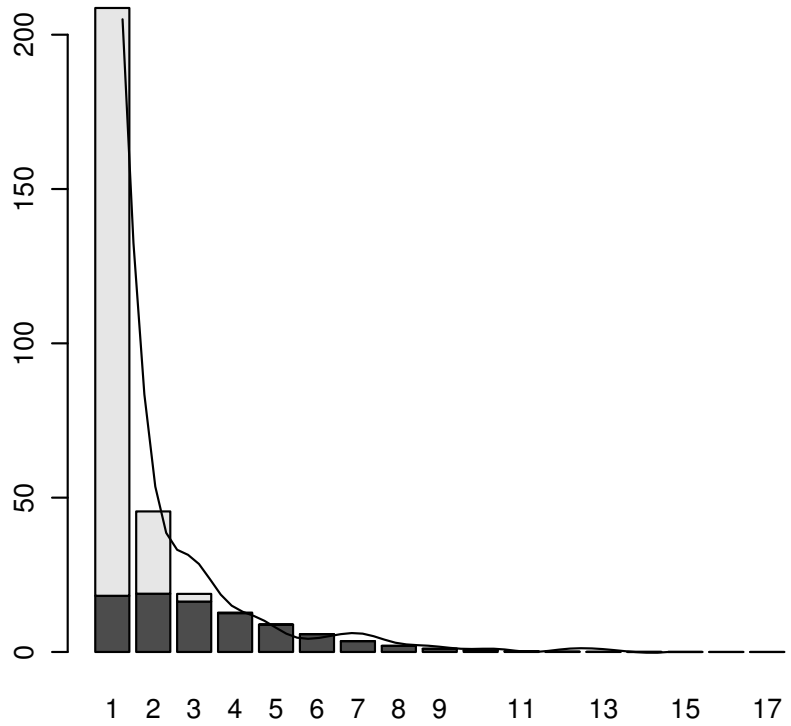
Beobachtet vs. unbeobachtet

X:=Times reported	**0**	**1**	**2**	**3**	**4**	**5**
Observed	-	4	3	1	0	1
True Defects	1	1	1	1	0	1
Unsuspicious Points	21	3	1	0	0	0

- beobachtet: Summe F+H
- unbeobachtet: nie berichtete Inspektionspunkte
- Problem: Verzerrung durch falsche Alarme

Lösungsansatz

- Betabinomiale Modellierung der True Defects
- Binomiale Modellierung der falschen Alarme (sehr kleines p)



Vorteile des Ansatzes

- bessere Schätzung der übersehenen Schwachstellen
  - im CUE-4 Datensatz:
    Betabin Modell: ca. 50
    Betabin+Binom: ca. 12
- Schätzung der Rate falscher Alarme

Noch zu tun

- ML Schätzer für zero-truncated Betabin+Binom
- Reliabilitätstest mit simulierten Daten
- Validierung an realen Datensätzen

Zusammenfassung

1. Grundidee der ML Schätzung
2. ML Schätzverfahren: grafisch, brute force, optim()
3. ML als Ausgangspunkt des Modellvergleichs (Alternative zum Hypothesentesten)
4. Aktuelles Forschungsproblem: betabinomiale Schätzung mit falschen Alarmen und unbeobachteten X=0