

Modellbasierte Korpusforschung und Bayessche Statistik

Workshop

Christoph Finkensiep¹, Martin Rohrmeier²

GMTH Jahreskongress, Freiburg 23.09.2023

¹Universiteit van Amsterdam, c.finkensiep@uva.nl

²École Polytechnique Fédérale de Lausanne, martin.rohrmeier@epfl.ch

Agenda

- Theorie
 - Modelle
 - Wahrscheinlichkeiten (basics)
 - Bayessche Inferenz
- Praxis
 - Probabilistic Programming
 - Ausführliches Beispiel

Vorwissen: Bruchrechnung, Dreisatz

Modelle, Wahrscheinlichkeiten, Inferenz

Das Problem

Ich habe einen Datensatz erhoben / annotiert / berechnet. Was nun?

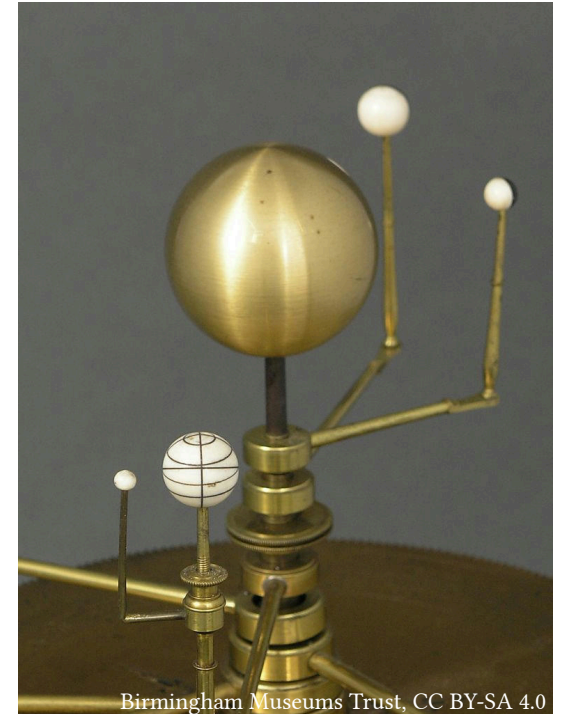
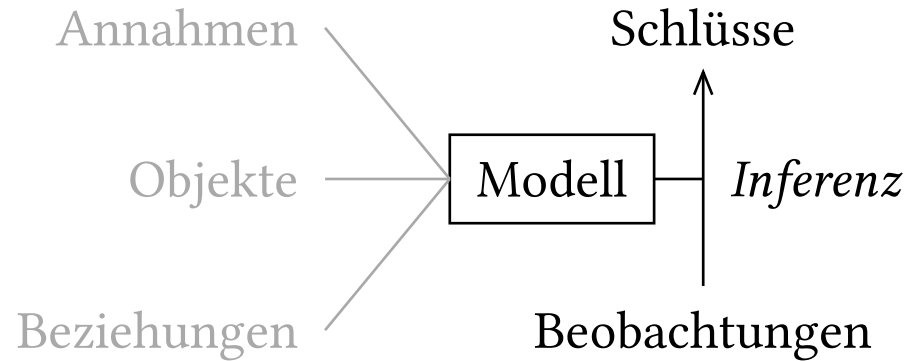
Statistik ist kompliziert...

- viele verschiedene Methoden
- Annahmen und Implikationen nicht offensichtlich
- Was mache ich in komplexen Fällen?

Der Ansatz

Modelle!

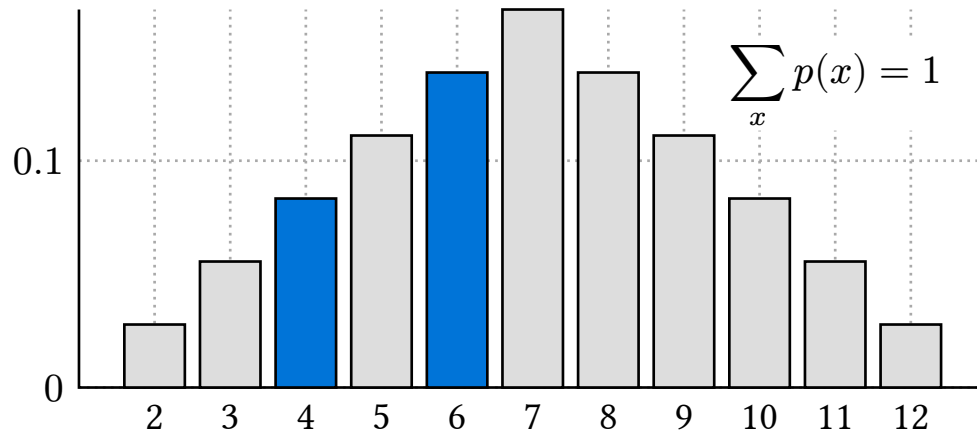
- beschreiben einen Ausschnitt der Welt (vereinfacht)
- relevante **Objekte** und **Beziehungen**
- explizite **Annahmen**
- erlauben **Simulation** und **Inferenz**



Basics: Verteilungen

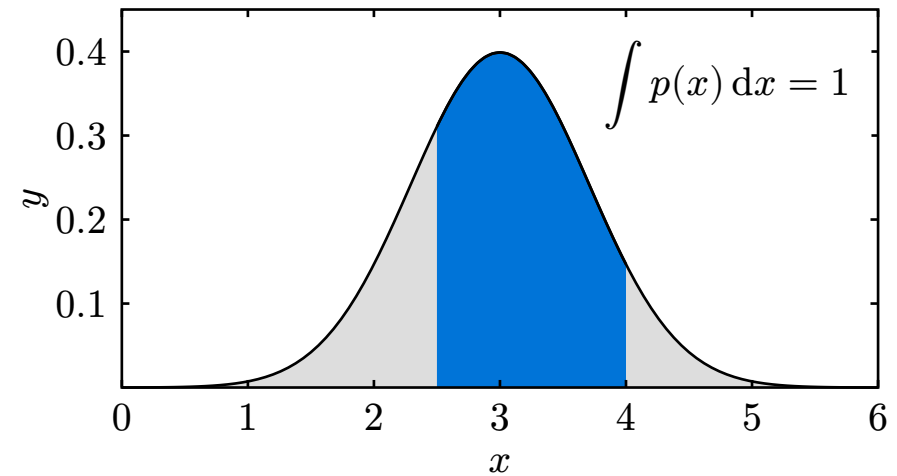
Zufallsvariable X :

diskret: „Massefunktion“ $p(x)$



$$P(X \in \{4, 6\}) = p(4) + p(6)$$

kontinuierlich: „Dichtefunktion“ $p(x)$



$$P(2.5 \leq X \leq 4) = \int_{2.5}^4 p(x) dx$$

Basics: Die Grundrechenarten

Gemeinsame Verteilung:

$$p(x, y)$$

	c	d	e	...
Dur	0.5	0.04	0.02	...
Moll	0.09	0.12	0.08	...

Randverteilung:

$$p(x) = \sum_y p(x, y)$$

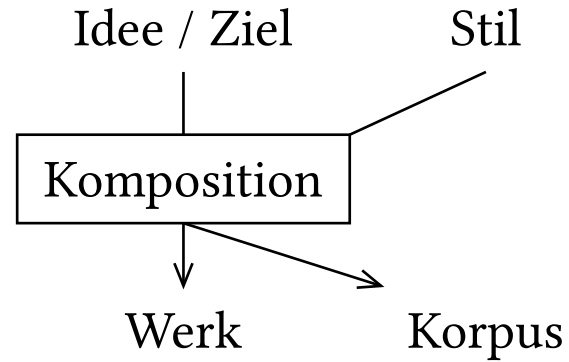
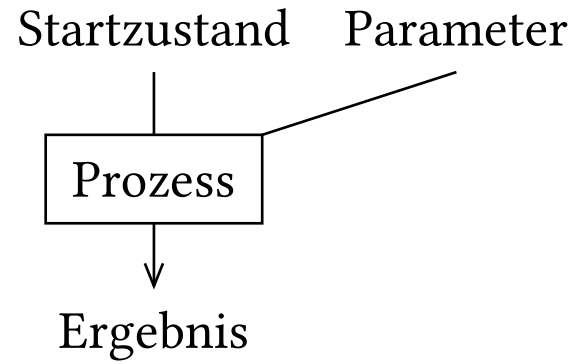
	0.7	c	d	e	...
Dur	0.7				
Moll	0.3	0.71	0.06	0.03	...

Bedingte Verteilung:

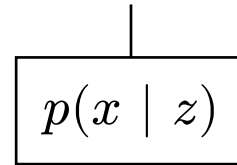
$$p(x|y) = \frac{p(x, y)}{p(y)}$$

	c	d	e	...	
Dur	0.71	0.06	0.03	...	= 1
Moll	0.3	0.4	0.27	...	= 1

Inferenz



latente Variablen (z / θ)



beobachtete Variablen (x)

Inferenz

$$p(z \mid x) = \frac{p(x, z)}{p(x)} = \frac{p(x \mid z) \cdot p(z)}{p(x)}$$

Der Satz von Bayes

x : beobachtete Variablen

z : latente Variablen

The diagram shows the equation for Bayes' theorem with arrows pointing from descriptive labels to the corresponding mathematical terms:

- posterior** points to $p(z \mid x)$
- joint** points to $p(x, z)$
- evidence** points to $p(x)$ in the denominator
- likelihood** points to $p(x \mid z)$
- prior** points to $p(z)$

$$p(z \mid x) = \frac{p(x, z)}{p(x)} = \frac{p(x \mid z) \cdot p(z)}{p(x)}$$

Beispiel: Dur oder Moll?

Beobachtet:

1 Dur
2 Dur
3 Moll
4 Dur
...

Dur: 112

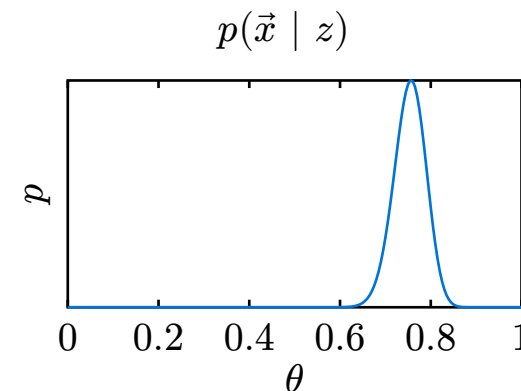
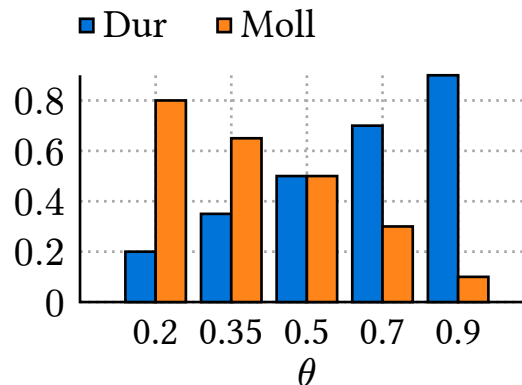
Moll: 36

Modell (Prozess):

- wähle Wahrscheinlichkeit θ
- für jedes Stück:
 - wirf Münze mit Wk. θ
 - Kopf: Dur
 - Zahl: Moll

Likelihood:

$$p(\vec{x} \mid \theta) = \prod_i p(x_i \mid \theta) = \prod_i \begin{cases} \theta & (x_i = \text{Dur}) \\ 1 - \theta & (x_i = \text{Moll}) \end{cases}$$



$$\max_{\theta} p(\vec{x} \mid \theta) = \frac{112}{112 + 36} = 0.757$$

Make it Bayessch

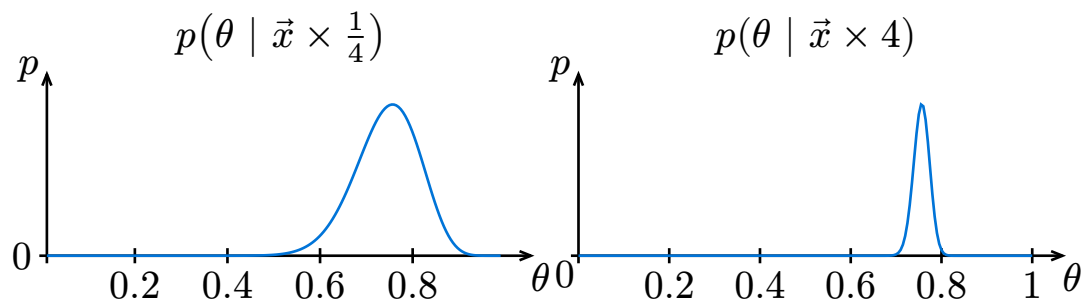
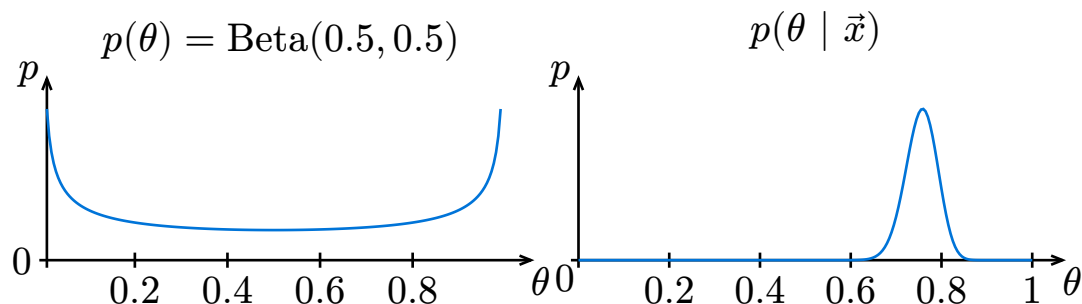
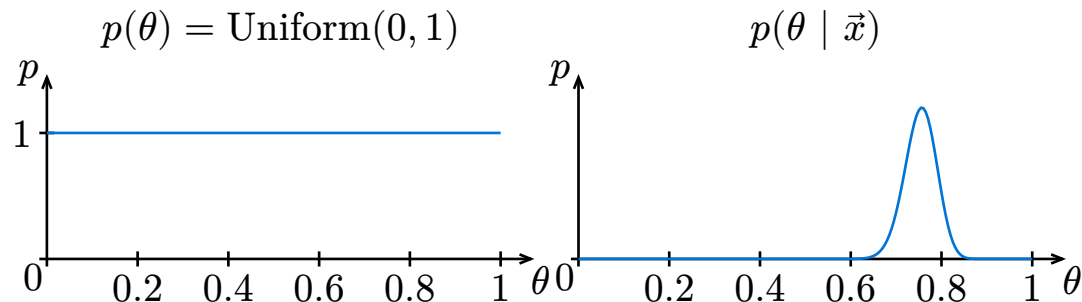
$$p(\theta \mid \vec{x}) = \frac{p(\vec{x} \mid \theta) \cdot p(\theta)}{p(\vec{x})}$$

Modell:

- wähle $\theta \sim \text{Uniform}(0, 1)$
(oder $\theta \sim \text{Beta}(0.5, 0.5)$)
- für jedes Stück i :
 - wähle $x_i \sim \text{Bernoulli}(\theta)$

Problem:

$$p(\vec{x}) = \int p(\vec{x}, \theta) d\theta \quad ???$$



Probabilistic Programming

Ein Modell als Programm

Tonart wählen:

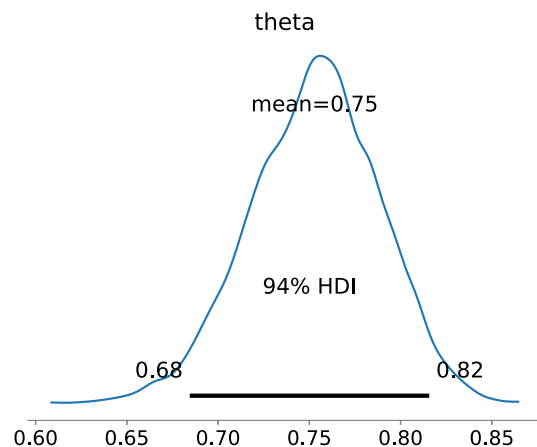
- wähle $\theta \sim \text{Uniform}(0, 1)$
- für jedes Stück i :
 - wähle $x_i \sim \text{Bernoulli}(\theta)$

```
def generate_keys(n):  
    theta = uniform(0, 1)  
    xs = []  
    for i in range(n):  
        maj = bernoulli(theta)  
        xs.append("d" if maj else "m")  
    return xs
```

Ein Modell als Programm

Tonart wählen:

- wähle $\theta \sim \text{Uniform}(0, 1)$
- für jedes Stück i :
 - wähle $x_i \sim \text{Bernoulli}(\theta)$



```
import pymc as pm
import arviz as az
```

```
keys = [0, 0, 1, 0, ...]
```

```
with pm.Model() as model:
    theta = pm.Uniform("theta", 0, 1)
    pm.Bernoulli("obs", p=theta, observed=keys)
```

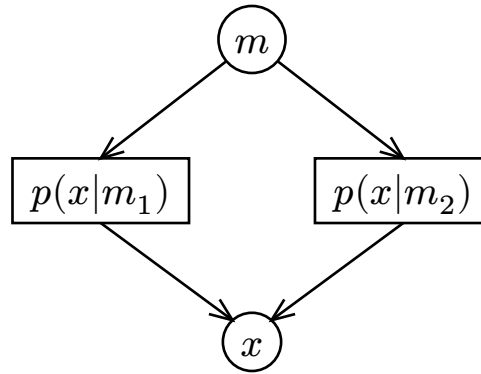
```
with model:
    idata = pm.sample(5000, chains=2)
```

```
az.plot_posterior(idata)
```

Ausführliches Beispiel: Notebook



Modellvergleich



Bayes Factor: $K = \frac{p(x \mid m_1)}{p(x \mid m_2)} = \frac{p(m_1 \mid x) \cdot p(m_2)}{p(m_2 \mid x) \cdot p(m_1)}$

Weiterführendes Material

- PyMC – viel Material für Einsteiger
- pyro – Beispiele und Tutorials für Variational Inference
- numpyro – komplexere Modelle, setzt etwas pyro voraus
- Bücher über Bayesian Statistics