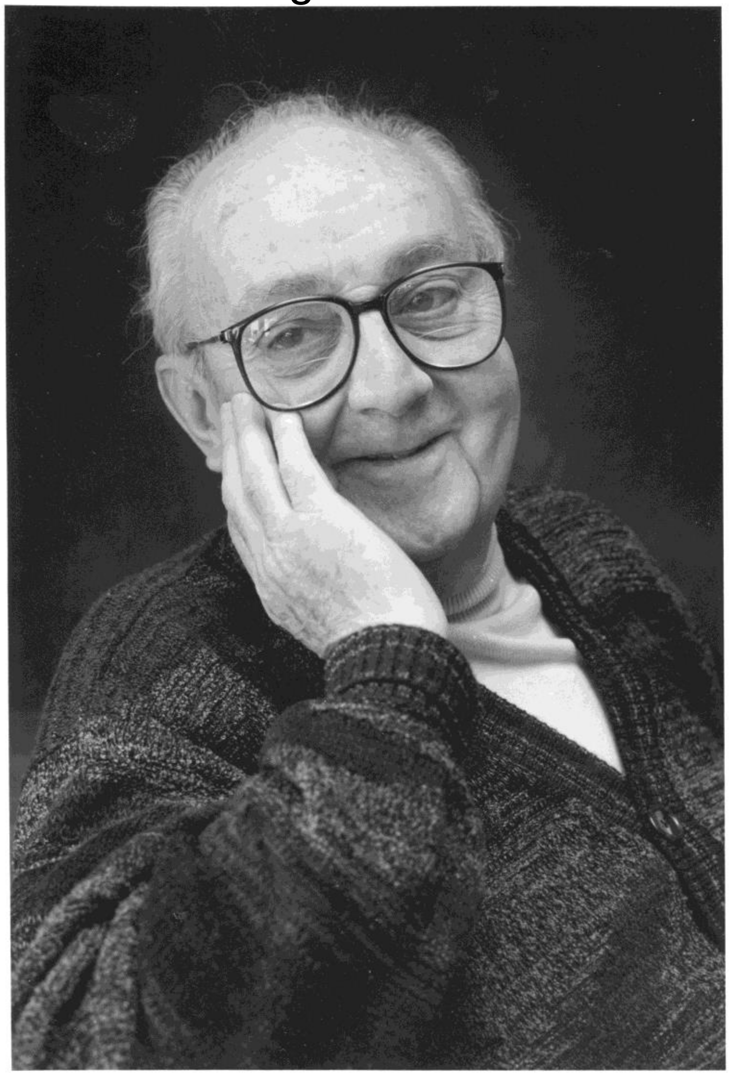
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Diskrete Wa.verteilungen: Eine Zooführung

Statistik (Biol./Pharm./HST) – FS 2014

Warum Wa.verteilungen?

George E.P. Box



“Essentially,
all models are
wrong,
but some are
useful.“

“Standard” Wa.verteilungen

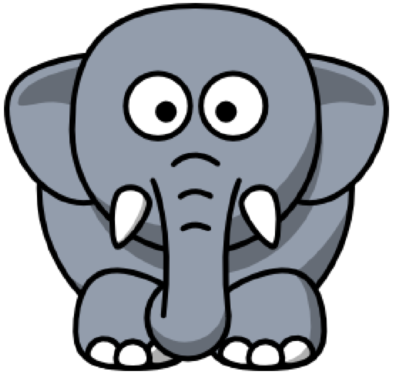
Details dieser Verteilungen in Büchern oder Software festgehalten

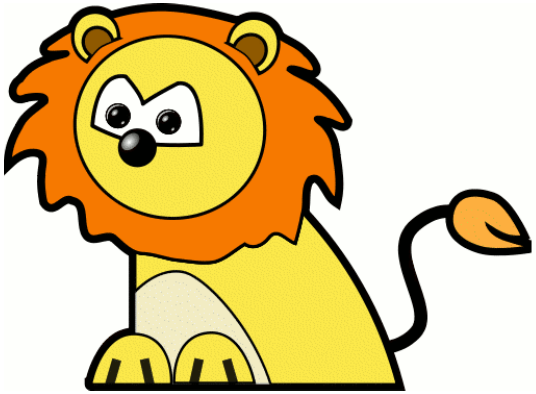
Viele typische Probleme einfach lösbar

Verteilungs-Zoo: Diskrete Wa.verteilungen

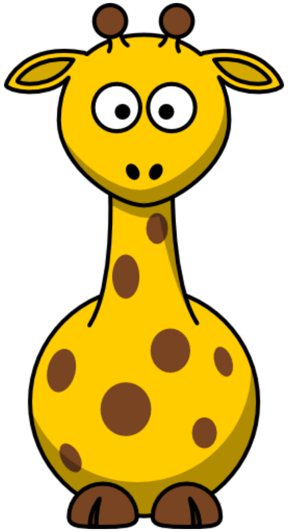
Uniforme Verteilung

Binomialverteilung

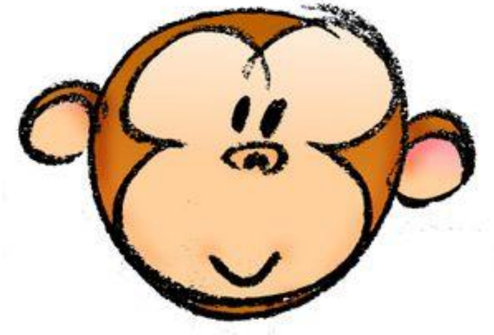




Poisson Verteilung



Hypergeometrische Verteilung

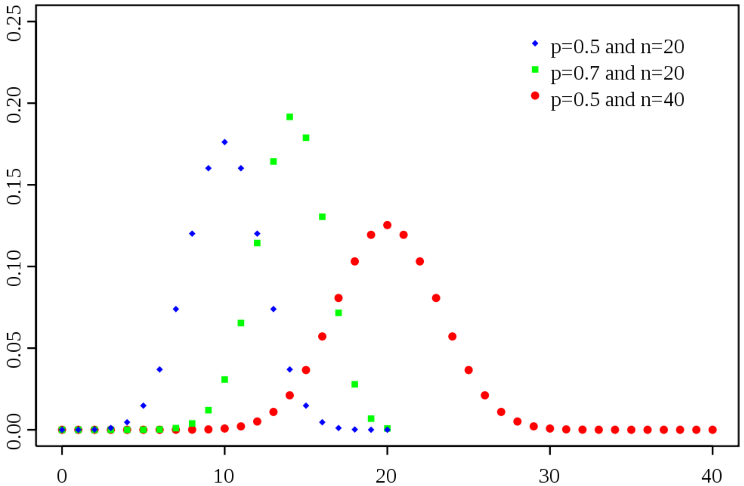


. . .

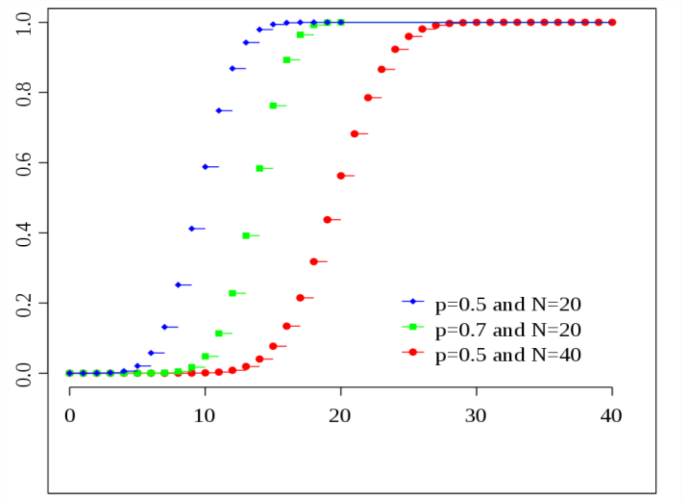
Binomialverteilung

- **Situation:** Ziehe n Lose an Losbude; gleiche Gewinnwa. für alle Lose; Lose unabhängig
- ZV X: Anzahl Gewinne unter n Losen
- $X \sim Bin(n, \pi)$
  "X ist binomial verteilt mit Parametern n und π"
- $P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x},\ x \in \{0,1,\ldots,n\}$
- $E(X) = n \cdot \pi,\ Var(X) = n \cdot \pi \cdot (1-\pi)$

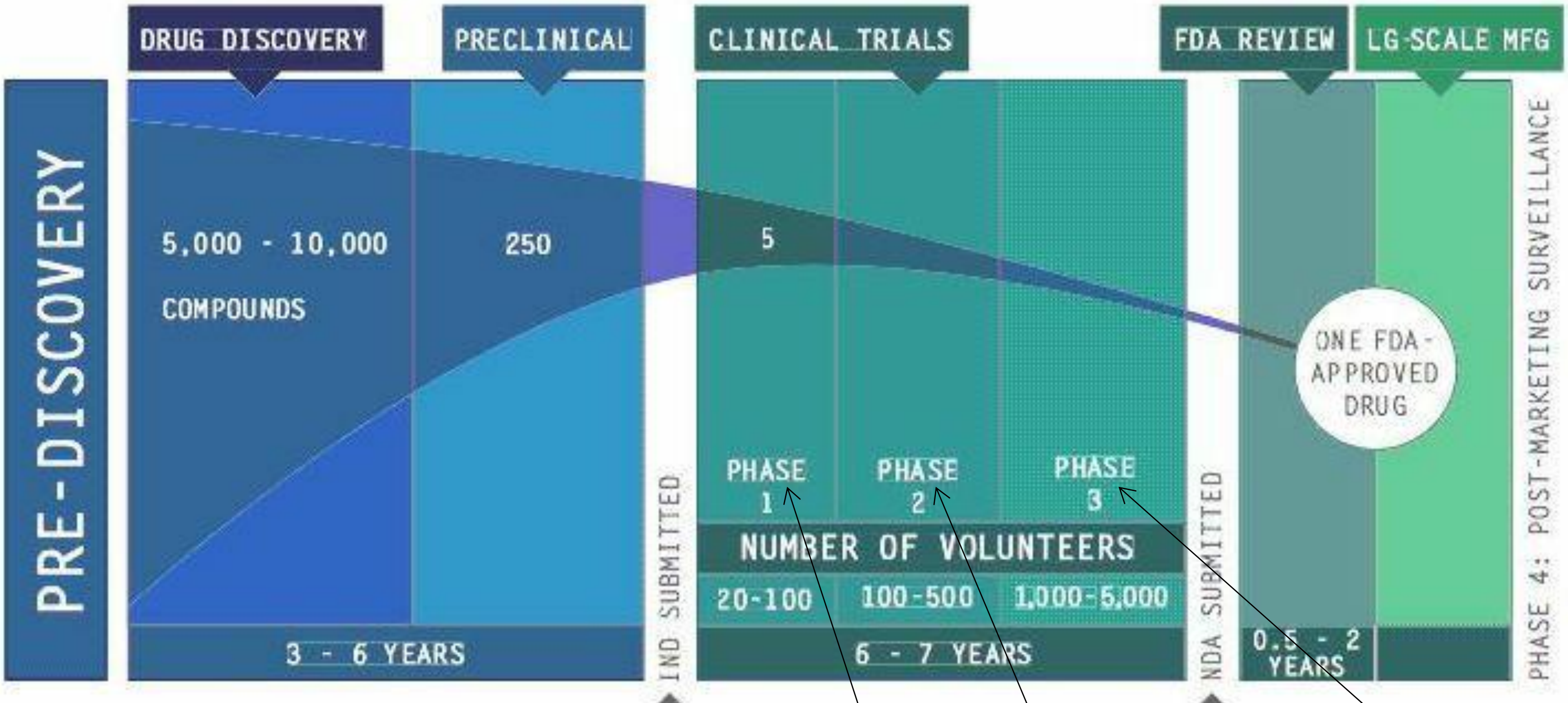
"probability mass function" (pmf)



"cumulative mass function" (cmf)



Beispiel: Klinische Studien



- **Lose**: Alle denkbaren Patienten
- **n gezogene Lose**: n Patienten in Studie
- **Gewinn**: Patient wird gesund
- π: Anteil aller denkbaren Patienten, bei denen Medikament wirkt

Giftig?

Grob: Wirksam?

Genau: Wirksam? Nebenwirkungen?

Bsp: Phase 2

- Hersteller behauptet: Neues Medikament wirkt in 80% der Fällen
- In einer Phase 2 Studie mit 100 Patienten werden aber nur 67 gesund
- Ist das plausibel, wenn die Heilungswa. 80% ist?
- X: Anzahl geheilter Patienten
- Falls Hersteller recht hat:

$$X \sim Bin(n = 100, \pi = 0.8)$$

Wie testen wir die Behauptung “$\pi = 0.8$”?

Versuch 1: $P(X = 67) = 0.0008$

Sind Sie überzeugt ?

Bsp: Phase 2 - Problem

- $X \sim Bin(n = 100, \pi = 0.8)$
- Angenommen, wir haben genau $n \cdot \pi = 80$ Genesungen gesehen; wir sollten dem Hersteller also unbedingt glauben

	n=100	n=1000	n=10'000	n=100'000
$P(X = n\pi)$	0.10	0.03	0.01	0.003

⇨ $P(X = 67)$ ist keine gute Kennzahl, weil die Wa. für jede beliebige Zahl klein wird, wenn man nur genug Beobachtungen hat!

	n=100	n=1000	n=10'000	n=100'000
$P(X \leq n\pi)$	0.54	0.51	0.504	0.501

⇨ $P(X \leq 67)$ ist eine gute Kennzahl; sie ist, unabhängig von der Stichprobengrösse, leichter zu interpretieren.

p-Wert ("Wa. für Beob. oder etwas noch extremeres"; später mehr dazu…)

Bsp: Phase 2

- Hersteller behauptet: Neues Medikament wirkt in 80% der Fällen
- In einer Phase 2 Studie mit 100 Patienten werden aber nur 67 gesund
- Ist das plausibel, wenn die Heilungswa. 80% ist?
- X: Anzahl geheilter Patienten
- Falls Hersteller recht hat:

$$X \sim Bin(n = 100, \pi = 0.8)$$

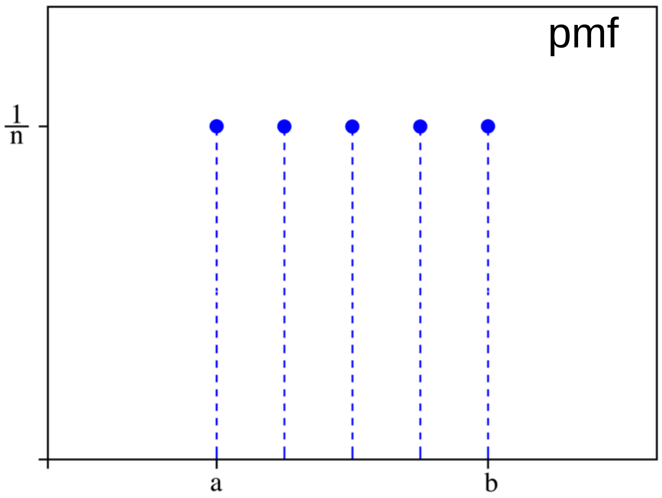
Wie testen wir diese Behauptung?

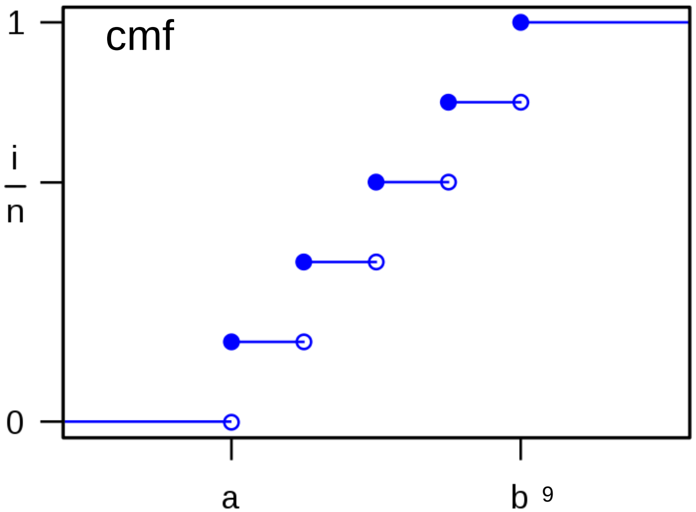
Versuch 2: $P(X \leq 67) = 0.001$

→ Beobachtung und Hypothese passen nicht zusammen; vermutlich wirkt das Medikament schlechter als 80%.

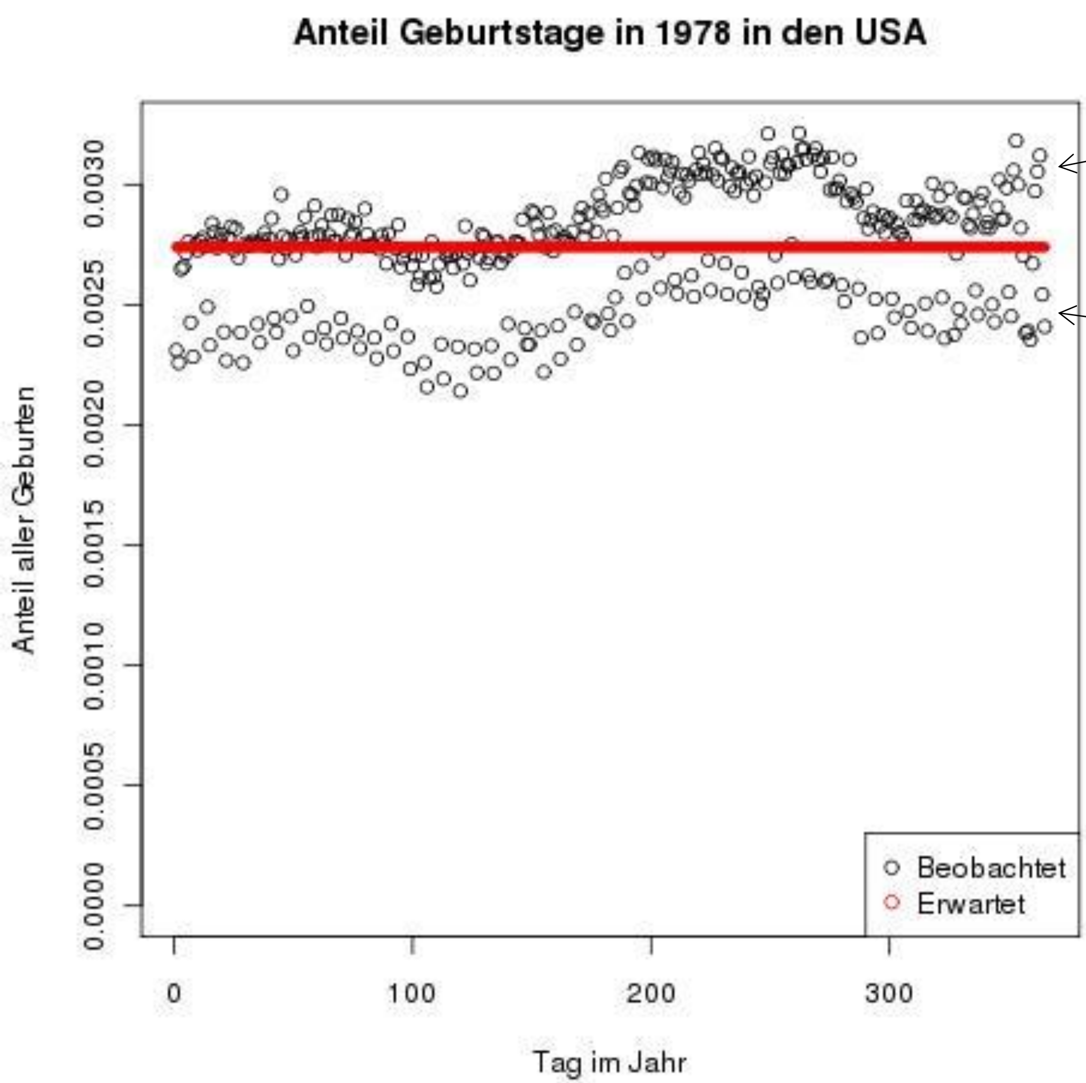
Uniforme Verteilung

- Situation: Ziehe eine Zahle aus {1,2,…,n}; alle Zahlen sind gleich wahrscheinlich
- ZV X: Gezogene Zahl
- $X \sim Unif(n)$
  “X ist uniform verteilt auf den Zahlen 1 bis n”
- $P(X = x) = \frac{1}{n}, x \in \{1,2,\ldots,n\}$
- $E(X) = \frac{n+1}{2}, Var(X) = \frac{(n+1)(n-1)}{12}$





Bsp: Sind Geburtstage uniform verteilt?



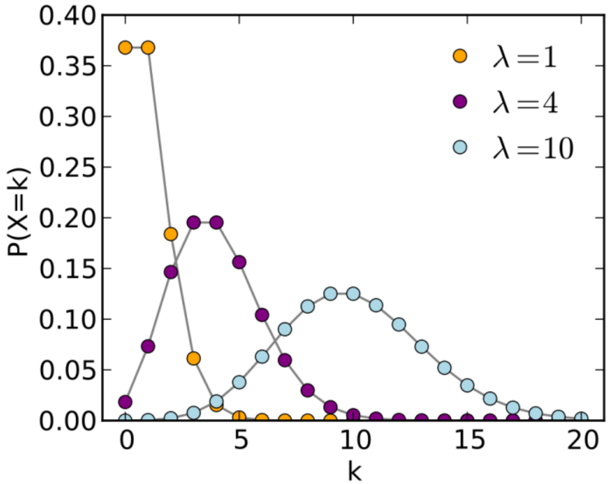
Wochentag

Wochenende

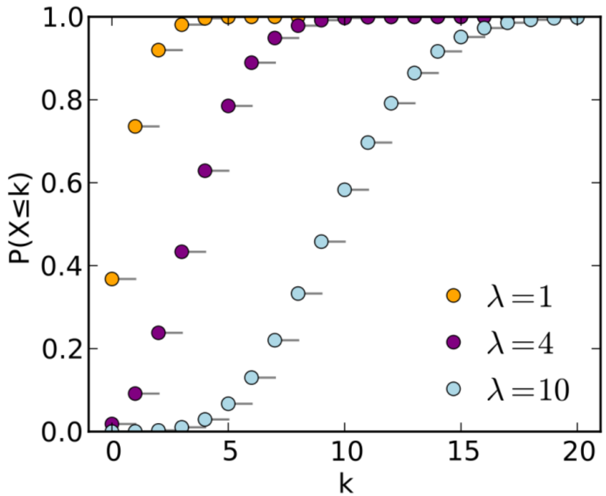
Ja, in grober Näherung schon

Poisson Verteilung

- Situation: Seltene Ereignisse werden in einem vorgegebenen Zeitraum gezählt
- ZV X: Anzahl beobachteter Ereignisse
- $X \sim Pois(\lambda)$
  ‘X ist poisson verteilt mit Paramter’
- $P(X = x) = \frac{\lambda^x}{x!} \exp(-\lambda)\ , x \in \{0,1,\dots,\infty\}$
- $E(X) = \lambda, Var(X) = \lambda$

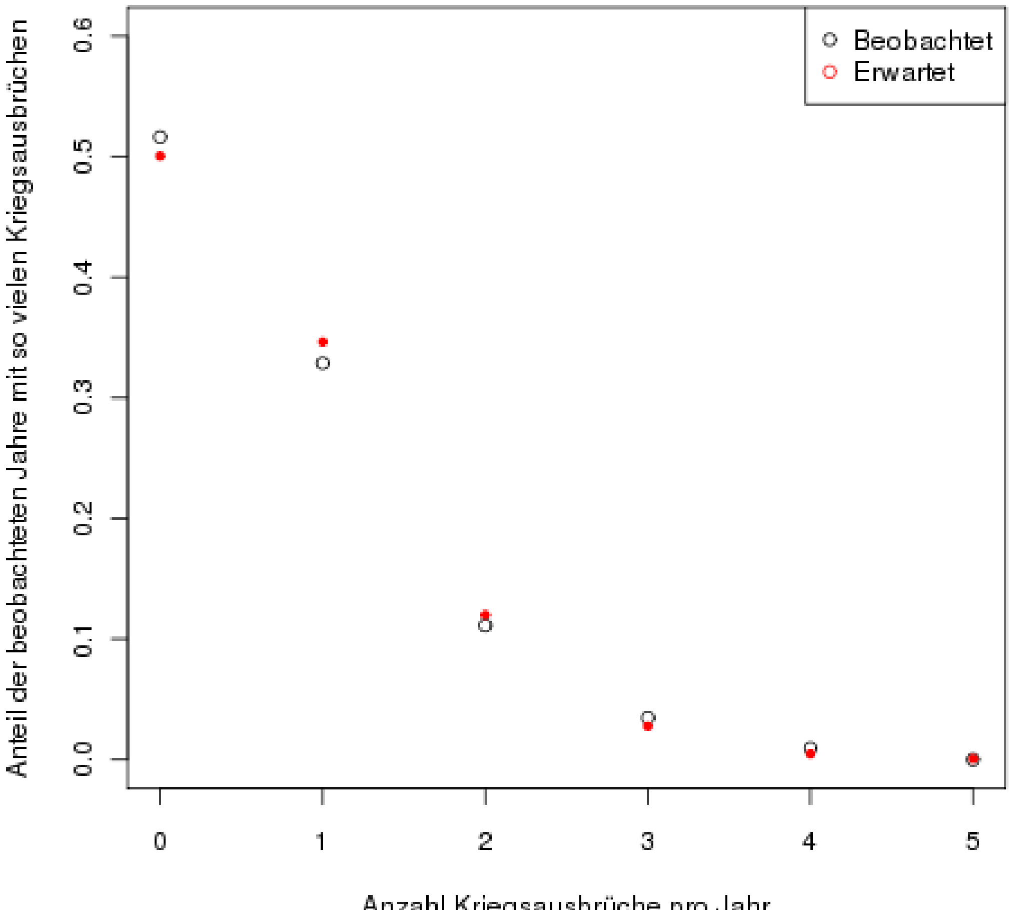


pmf



cmf

Bsp: Ist die Anzahl Kriege pro Jahr poisson verteilt?
(1500-1930, weltweit)



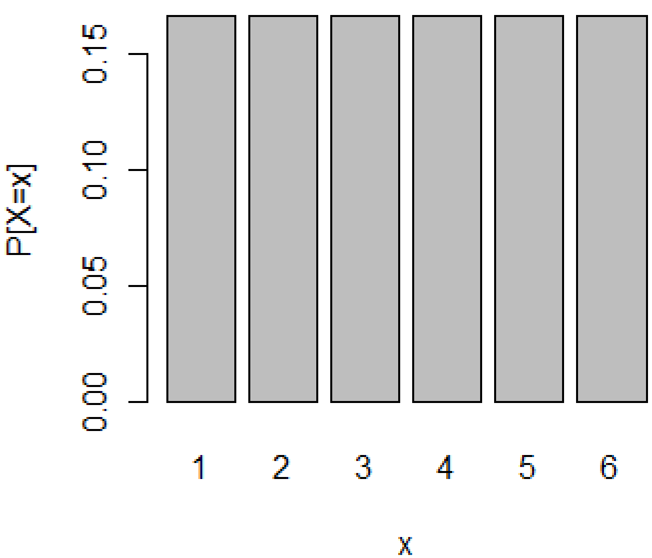
Besonderheit der Poissonverteilung

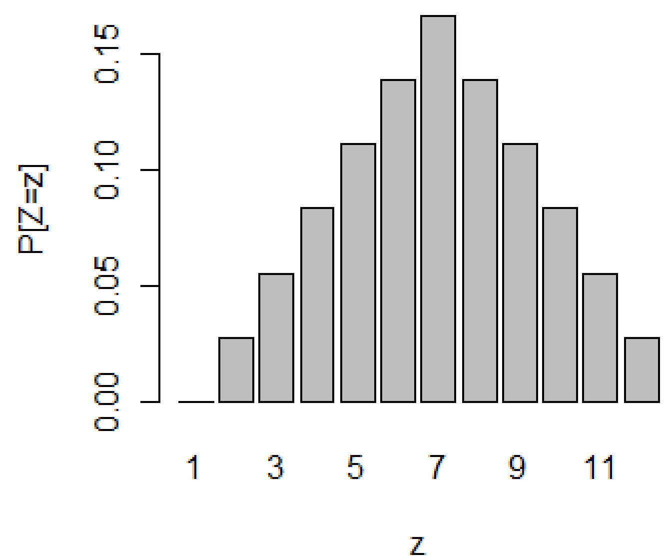
- Angenommen:
  - $X \sim Pois(\lambda_1), Y \sim Pois(\lambda_2)$
  - X, Y sind unabhängig
- Bilde neue Zufallsvariable: $Z = X + Y$
- Dann gilt: $Z \sim Pois(\lambda_1 + \lambda_2)$

- Das gilt normalerweise nicht!

Normalerweise: Summe von zwei Verteilungen gibt neue Verteilung

- Bsp: $X \sim Unif(\{1,2,3,4,5,6\}), Y \sim Unif(\{1,2,3,4,5,6\})$
  X, Y sind unabhängig
- $Z = X + Y$ ist nicht uniform verteilt (Augensumme 2 ist selten, Augensumme 7 ist häufig)



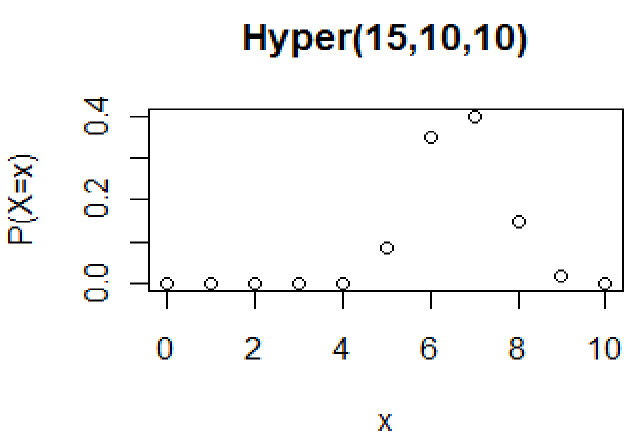


Hypergeometrische Verteilung

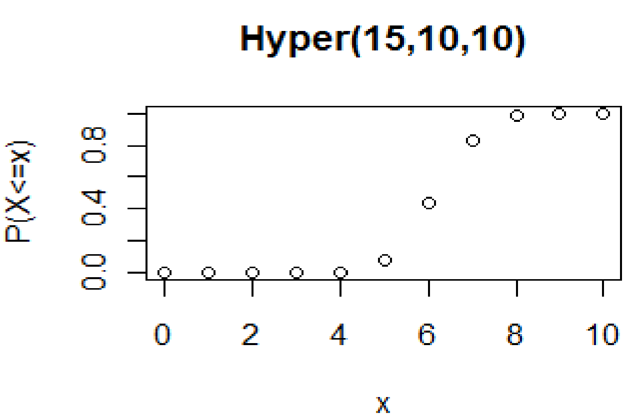
- Situation: Urne mit N Kugeln; m sind markiert; ziehen n Kugeln ohne Zurücklegen; wie viele markierte Kugeln?
- ZV X: Anzahl markierter gezogener Kugeln
- $X \sim Hyper(N, n, m)$

“X ist hypergeometrisch verteilt mit Paramtern N, n und m”

- $P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$ ‘günstig’ ‘möglich’ $x \in \{0, 1, \ldots, \min(n, m)\}$
- $E(X) = \frac{(n \cdot m)}{N}$, $Var(X)$ kompliziert; siehe z.B. Wikipedia



pmf



cmf

Bsp: Phase 3 Studie – Wirksamer als Placebo?

Doppel-blinde, randomisierte Studie

Markierte Bälle (m)

Gezogene und markierte Bälle

	Medikament	Placebo	Total
Geheilt	15	9	24
Nicht geheilt	10	11	21
Total	25	20	45

Gezogene Bälle (n)

Bälle in Urne (N)

Falls Medikament keine Wirkung hat: Es gibt 24 Personen, bei denen unabhängig von der Gruppenzuteilung fest steht, dass sie gesund werden

Urnenmodell → Hypergeometrische Verteilung

Bsp: Phase 3 Studie – Wirksamer als Placebo?

	Medikament	Placebo	Total
Geheilt	15	9	24
Nicht geheilt	10	11	21
Total	25	20	45

ZV X: Anzahl geheilter Patienten in Medikamenten-Gruppe
Falls Medikament keine Wirkung hat:

$$X \sim Hyper(N = 45, m = 24, n = 25)$$

Ist es dann plausibel in der Medikamenten-Gruppe 15 oder mehr geheilte Patienten zu beobachten?

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.76 = \boxed{0.24}$$ p-Wert

Falls das Medikament nicht wirkt, ist es durchaus plausibel 15 oder mehr geheilte Patienten in der Medikamentengruppe zu beobachten

Momentenmethode, Bsp 1:

- 100 zufällig ausgewählte Patienten bekommen neues Medikament
- 54 werden gesund
- Wie gross ist wohl die Wirkwahrscheinlichkeit in der gesamten Bevölkerung?

- X: Anzahl Patienten, die gesund wurden
  $X \sim Bin(n = 100, \pi = ?)$
  Beobachtung: $x = 54$

- Momentenmethode um π zu schätzen:

$$E(X) = n \cdot \pi; \;\; E(X) \approx x = 54 \to x \approx n \cdot \pi \to \pi \approx \frac{x}{n} = 0.54$$

"Erstes Moment"

Momentenmethode, Bsp 2: Capture-Recapture

- Wie gross ist eine Population, von der wir sonst *gar nichts* weiter wissen?
- Bsp: Ameisen in Ameisenhaufen; Fische in See
- Lincoln-Peterson Methode:
  - Fange m zufällige Tiere, markiere, lasse wieder laufen
  - Fange n zufällige Tiere
  - ZV X: Anzahl markierter Tiere im zweiten Fang
- $X \sim Hyper(N, n, m)$, wobei N die Grösse der Pop. ist; x markierte Tiere im zweiten Fang
- Idee: «Erwartungswert ≈ Beobachtung»
  - $E(X) = \frac{n \cdot m}{N} \approx x \rightarrow N \approx \frac{n \cdot m}{x}$
- Ungenau, aber OK für richtige Grössenordnung

"Erstes Moment"

Maximum-Likelihood Methode 1/3

Bsp: n=600 Personen erhalten neues Medikament;
x = 30 haben als Nebenwirkung Kopfschmerzen

Wie gross ist der Anteil Personen mit diesen Nebenwirkungen in der Gesamtbevölkerung (>600) ?

Binomialverteilung:

- X: Anzahl Personen mit Kopfschmerzen
- $X \sim Bin(n = 600, \pi)$
- $P(X = 30) = \binom{600}{30}\pi^{30}(1-\pi)^{570}$

Maximum-Likelihood Estimate (MLE) $\hat{\pi}$ für π, ist der Wert, der $P(X = 30)$ maximiert.

Maximum-Likelihood Methode 2/3: Computer

Berechne $P(X = 30)$ für verschiedene Werte von π mit dem Computer:

π	…	0.03	0.04	0.05	0.06	0.07	…
$P(X = 30)$		0.002	0.036	0.075	0.042	0.010	

Maximum
$\hat{\pi} \approx 0.05$

Maximum-Likelihood Methode 3/3: Analytisch

- $P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} =: f(\pi)$ "likelihood"
- Analysis: Finde π, sodass $f(\pi)$ maximal ist
  (s. Skript S.25)

- Ergebnis: $\hat{\pi} = \frac{x}{n} = \frac{30}{600} = 0.05$

Wir erwarten, dass bei etwa 5% der Gesamtbevölkerung Kopfschmerzen als Nebenwirkung auftritt.