Einführung in die Statistik und Datenanalyse (3)

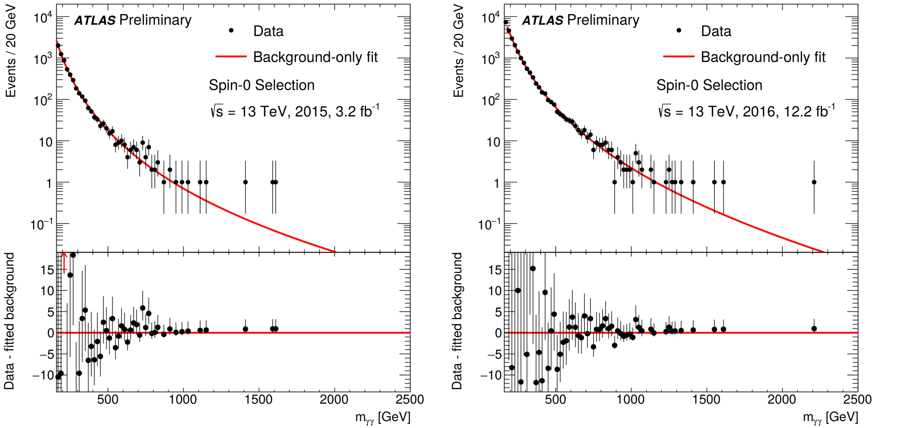
Martin Völkl
martin.voelkl@uni-tuebingen.de

Universität Tübingen
2018-03-20

Aufbau der Vorlesung

1. Einführung und systematische Unsicherheiten
2. Werkzeugkiste
3. **Wahrscheinlichkeit und Unsicherheit**
   - Das Induktionsproblem
   - Klassischer und axiomatischer Wahrscheinlichkeitsbegriff
   - Bayesianischer Wahrscheinlichkeitsbegriff
   - Frequentistischer Wahrscheinlichkeitsbegriff
4. Frequentistische Methoden
5. Bayesianische Methoden

Was bedeutet 3σ?



Search for scalar diphoton resonances with $15.4 fb^{-1}$ *of data collected at* $\sqrt{s} = 13$ TeV *in 2015 and 2016 with the ATLAS detector,* ATLAS Kollaboration

- 3.4σ Abweichung vom Untergrund
- Stellte sich als statistische Fluktuation heraus
- Wie sinnvoll Schlussfolgerungen aus Statistik ziehen?

Deduktiver Schluss

$$\frac{\text{Alle Tiere mit Federn haben Schnäbel} \quad \text{Dieser Rabe hat Federn}}{\text{Dieser Rabe hat einen Schnabel}}$$

- Typisch: Allgemeines Gesetz $\rightarrow$ Einzelfall
- Logisch korrekt

Induktiver Schluss

Dieser Rabe hat einen Schnabel
Dieser Rabe hat Federn

Alle Tiere mit Federn haben Schnäbel

- Logisch nicht korrekt
- Aus Einzelfall folgt nicht allgemeines Gesetz

Induktiver Schluss (2)

Gefiedertes Tier 1 hat einen Schnabel
Gefiedertes Tier 2 hat einen Schnabel
Gefiedertes Tier 3 hat einen Schnabel
$\vdots$

Alle Tiere mit Federn haben Schnäbel

- Auch aus vielen Einzelfällen folgt kein Gesetz
- Dies ist aber, was die Wissenschaft versucht
- Bekannt als *Induktionsproblem*

Induktiver Schluss (3)

Wenn A wahr ist, dann folgt B
B ist wahr

Also ist A wahr

- Keine korrekte Schlussfolgerung
- Können wir die Anforderungen ändern?

Induktiver Schluss (4)

Wenn A wahr ist, dann folgt B
B ist wahr

Dadurch wird A plausibler

- Besser, aber typisch ist etwas Anderes

Induktiver Schluss (5)

Wenn A wahr ist, dann wird B plausibler
B ist wahr

Dadurch wird A plausibler

- Beispiel:
  - A=”Die vier Lichtupunkte nahe des Jupiter sind Monde die um ihn kreisen”
  - B=”Die Lichtpunkte sind bei der nächsten Messung in einem geringen Winkel von Jupiter entfernt”

Selbstverständnis von Wissenschaftlern

- Intuitiv: Messungen geben Informationen über die Welt
- Eine Theorie, deren Vorhersagen genau überprüft wurden hat einen anderen Wert als eine ohne zugehörige Messungen
- Wenn Messergebnisse deutlich besser zu einer Theorie passen als zu einer anderen, ändert dies das Wissen über die beiden Theorien
- Zur Quantifizierung: Genaueres Verständnis von Wahrscheinlichkeit wichtig

Wahrscheinlichkeit in der Sprache

Wie wird der Begriff in den Beispielen verwendet?

- Das Projekt ist wahrscheinlich bis Ende der Woche fertig
- Die Wahrscheinlichkeit für einen 6er-Pasch mit zwei Würfeln ist $1/36$
- Wahrscheinlich war ein Meteorit für das Aussterben der Dinosaurier verantwortlich
- Die Regenwahrscheinlichkeit für morgen beträgt 75%
- Der wahre Wert befindet sich wahrscheinlich im Intervall $a \pm \sigma_a$
- Der Patient hat wahrscheinlich eine Grippe
- Du wirst wahrscheinlich diese Woche nicht im Lotto gewinnen
- Die Nationalmannschaft wird ihren Titel mit einer Wahrscheinlichkeit von 15% erfolgreich verteidigen

Alle Beispiele haben im Sprachgebrauch eine Bedeutung

Klassischer Wahrscheinlichkeitsbegriff

- Experiment mit mehreren (N) möglichen Ausgängen, die sich gegenseitig ausschließen
- Alle Ausgänge gleich Wahrscheinlich (Symmetrie zwischen Resultaten)
- Gruppe von N_A davon ausgewählt (z.B. "rot")
- Dann ist die Wahrscheinlichkeit:

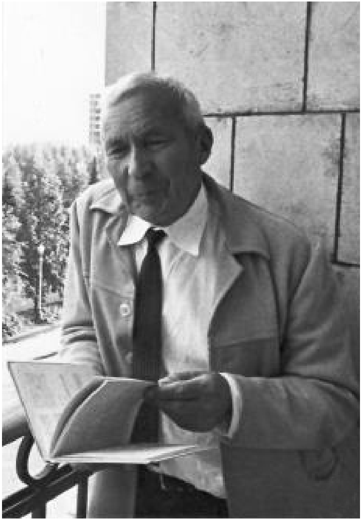
$$p(A) = \frac{N_A}{N}$$

- Häufig Def. im Schulunterricht
- Keine nützliche Definition für die Physik

Quelle: Wikipedia

Axiomatischer Wahrscheinlichkeitsbegriff (Kolmogorov)

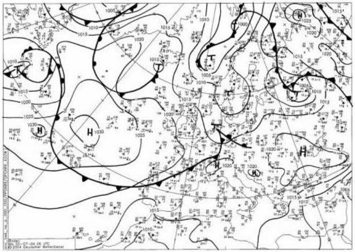
- Kolmogorov Axiome für Wahrscheinlichkeit $p(E)$ einer Ergebnisses E aus einer Menge möglicher Ergebnisse Ω:
  1. $P(E) \in \mathbb{R}$, $P(E) > 0 \; \forall E \in \Omega$
  2. $P(\Omega) = 1$
  3. $A \cap B = \emptyset \implies$ $P(A \cup B) = P(A) + P(B)$
- Erlaubt mathematisches Fundament
- Viele wichtige Formeln können hergeleitet werden
- Keine Verbindung zur echten Welt – für Physik unbrauchbar
- Aber: Die anderen Wahrscheinlichkeitsbegriffe erfüllen auch die Axiome



Quelle: Wikipedia

Bayesianischer Wahrscheinlichkeitsbegriff

- Wahrscheinlichkeit bezeichnet Wissensstand
- Beispiele:
  $p(\text{"Dunkle Materie sind WIMPS"})$,
  $p(\text{"Morgen wird es regnen"})$
- Sehr nahe an der umgangssprachlichen Nutzung des Begriffes
- Um in Zahlen auszudrücken: Rationale Wette mit momentanem Wissensstand
- Da $p(A) + p(\bar{A}) = 1$: Wenn nicht für A gewettet wird, dann für $\bar{A}$
- Wahrscheinlichkeiten ändern sich mit neuen Informationen
- Unter Symmetrie: $1/N$ einzig rationale Zuordnung (enthält klassischen Fall)



Quelle: Wikipedia

Hat das dann noch etwas mit Zufall zu tun? (Münzwurf)

- Gegebene Information A beeinflusst Wahrscheinlichkeitsverteilung $p(x)$
- Schreibe $p(x|A)$
- Beispiel: x="Es regnet morgen", A="Es regnet heute"
- Poisson: $p(x|\lambda)$
- $p(\mathrm{Waldbrand\ in\ 24h}|\mathrm{T})$

Quelle: Wikipedia

- Produktregel: $p(x, y) = p(x|y) \cdot p(y)$
- Die Wahrscheinlichkeit, dass Carlsen sein nächstes (Weiß-)Spiel mit "e4" eröffnet und gewinnt ist die Wahrscheinlichkeit, dass er es mit "e4" eröffnet mal der Wahrscheinlichkeit, dass er das Spiel gewinnt, falls er es mit "e4" eröffnete

Bayesianisches Wahrscheinlichkeitskalkül

- *Bayesian Inference*: Wahrscheinlichkeit als Erweiterung der Logik
- Wahre Aussage: $p = 1$, falsche Aussage $p = 0$
- Kalkül muss mindestens die formale Logik enthalten
- R.T. Cox: 3 "Axiome" (auch *Desiderata*)

1. Axiom

Die Plausibilität einer Aussage wird durch eine reelle Zahl repräsentiert

- $p(A)$ ist die Plausibilität von A (für den Moment noch nicht Wahrscheinlichkeit)
- $p(A|B)$ – Plausibilität von A, gegeben B
- $p(A, B|C)$, Plausibilität von "A UND B" gegeben C
- $p(A|C) > p(B|C)$ – A ist plausibler, gegeben C

2. Axiom

Die Plausibilität einer Aussage korrespondiert qualitativ mit dem gesunden Menschenverstand

- Wenn neues Wissen die Plausibilität von A erhöht, senkt dies die Plausibilität von $\bar{A}$, d.h.

$$p(A|C') > p(A|C) \iff p(\bar{A}|C') < p(\bar{A}|C)$$

3. Cox Axiom

3. Axiom

Die Berechnung der Plausibilität erfolgt konsistent

- Logisch äquivalente Aussagen sind gleich plausibel
- Wenn eine Schlussfolgerung auf unterschiedliche Weisen erfolgen kann, dann führen alle Wege zur gleichen Schlussfolgerung
- Das Kalkül erlaubt nicht, Informationen zu ignorieren; die Schlussfolgerung erfolt unideologisch

Cox' Theorem

Axiome:

1. Die Plausibilität einer Aussage wird durch eine reelle Zahl repräsentiert
2. Die Plausibilität einer Aussage korrespondiert qualitativ mit dem gesunden Menschenverstand
3. Die Berechnung der Plausibilität erfolgt konsistent

Es ergibt sich eine Klasse von äquivalenten Lösungen, eine davon hat die Eigenschaften:

1. Eine sichere Aussage wird repräsentiert durch $p(A) = 1$, eine sicher falsche Aussage durch $p(A) = 0$
2. $p(A) + p(\bar{A}) = 1$ (Summenregel)
3. $p(A, B) = p(A|B) \cdot p(B)$ (Produktregel)

p funktioniert also wie eine Wahrscheinlichkeit und kann so benannt werden

Der Satz von Bayes

- Wollen Wahrscheinlichkeiten möglicher Aussagen $A \in \{A_1, A_2 \ldots\}$ bestimmen
- Wissen B vorhanden, suchen $p(A|B)$
- Produktregel: $p(A, B) = p(A|B) \cdot p(B)$
- Aber auch: $p(A, B) = p(B|A) \cdot p(A)$
- Daraus folgt:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

- $p(B) = \sum_i p(A_i, B) = \sum_i p(B|A_i) \cdot p(A_i)$

$$\boxed{p(A|B) = \frac{p(B|A) \cdot p(A)}{\sum_i p(B|A_i) \cdot p(A_i)}}$$

Der **Satz von Bayes**

Der Satz von Bayes (2)

$$\boxed{p(A|B) = \frac{p(B|A) \cdot p(A)}{\sum_A p(B|A_i) \cdot p(A_i)}}$$

$p(A|B)$ *Posterior*-Wahrscheinlichkeitsverteilung: Wahrscheinlichkeit für A, wenn das Wissen B vorhanden ist

$p(A)$ *Prior*-Wahrscheinlichkeitsverteilung: Wahrscheinlichkeit für A, ohne das Wissen B

$p(B|A)$ *Likelihood*-Funktion

Nenner Normierung des Zählers

Anmerkung: Im Deutschen auch "a-priori-" und "a-posteriori-"Wahrscheinlichkeitsverteilung. Wegen Verwechselungsgefahr mit Kant, hier die englischen Begriffe

Bayesianische Schlusfolgerung – Diagnose

- Patient bei Vorsorgeuntersuchtung, Test für seltene Krankheit
- Test gibt in 99.95% der Fälle die richtige Antwort (Likelihood), also

$$p(\text{Test positiv}(+)|\text{Patient krank}(\text{k})) = 0.9995$$

$$p(\text{Test negativ}(-)|\text{Patient gesund}(\text{g})) = 0.9995$$

- Nur 1 Mensch in 100000, die zur Vorsorgeuntersuchung kommen hat die Krankheit, d.h. $p(\text{Patient krank}) = 10^{-5}$ (Prior)
- Der Test fällt positiv aus. Ist der Patient krank?
- Bayes:

$$p(k|+) = \frac{p(+|k) \cdot p(k)}{p(+|k) \cdot p(k) + p(+|g) \cdot p(g)} \approx 0.02$$

- Aber: Wahrscheinlichkeit um Faktor 2000 höher als vorher
- Wiederholte Testresultate wären wohl stark korreliert

Bayesianische Schlusfolgerung – Diagnose (2)

- $p = 0.02$, weiterer, unabhängiger Test ist auch positiv
- Test ist schwächer $p(+|k) = 0.99$, $p(-|g) = 0.99$
- Verwende alle Informationen! Neuer Prior ist Posterior des vorherigen Tests: $p(k) = 0.02$
- Jetzt folgt: $p(k|+) = \frac{0.99 \cdot 0.02}{0.99 \cdot 0.02 + 0.01 \cdot 0.98} \approx 0.67$
- Die Wahrscheinlichkeiten geben den momentanen Wissensstand an – und damit auch die Unsicherheit
- Laplace: *Wahrscheinlichkeit ist gesunder Menschenverstand ausgedrückt in Mathematik*

- Ein unbekannter aber fairer Würfel wirft die Zahlen 3, 6, 1, 5, 7, 4
- Wie viele Seiten ($3 \leq n \leq 20$) hat der Würfel?
- Wie Prior bestimmen?
- Konkret: Ein Kollege hat gewürfelt und fragt
- Prior: "Meiste Würfel $n = 6$, $n = 4, 10, 20$ kommen auch vor, vielleicht hat der Kollege auch einen besonders ungewöhnlichen gefunden"
- Ausgedrückt als $p(w_6) = 0.5$, $p(w_{20}) = p(w_{12}) = p(w_{10}) = p(w_4) = 0.1$, alle anderen gleiche Wahrscheinlichkeit

Quelle: Wikipedia

- Ein Wurf: k
- Suchen $p(w_n|k)$ (Posterior)
- Was ist $p(k|w_n)$? (Likelihood)

$$p(k|w_n) = \begin{cases} 1/n & k \leq n \\ 0 & \text{sonst} \end{cases}$$

- Für m Würfe:

$$p(k_1, k_2, \ldots k_m|w_n) = \begin{cases} 1/n^m & k_i \leq n \ \forall i \\ 0 & \text{sonst} \end{cases}$$

- Resultat: $p(w_6) = 0$, $p(w_7) = 0.258$, $p(w_8) = 0.116$, $p(w_{10}) = 0.395$, $p(w_{12}) = 0.13$

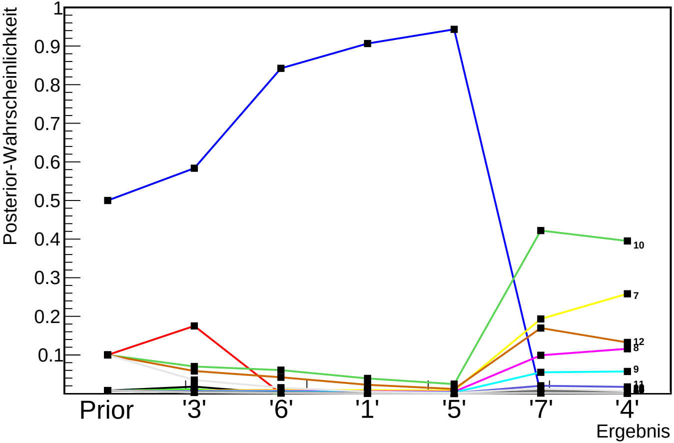
$$\boxed{p(A|B) = \frac{p(B|A) \cdot p(A)}{\sum_A p(B|A_i) \cdot p(A_i)}}$$

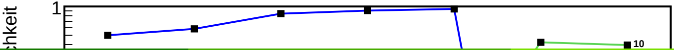
Quelle: Wikipedia

Bayesianische Schlussfolgerung – Würfel (3)

- Man kann alle Würfe auf einmal behandeln, oder einen nach dem anderen
- Im zweiten Fall wird der Posterior aus dem letzten Schritt der Prior zum nächsten
- Beides ergibt exakt das gleiche Resultat (Konsistenzaxiom)
- Die Bestimmung des Priors ist nicht einfach
- Tipp: Wie viel ist man bereit auf oder gegen z.B. w_{13} zu wetten?

Bayesianische Schlusfolgerung – Würfel (4)



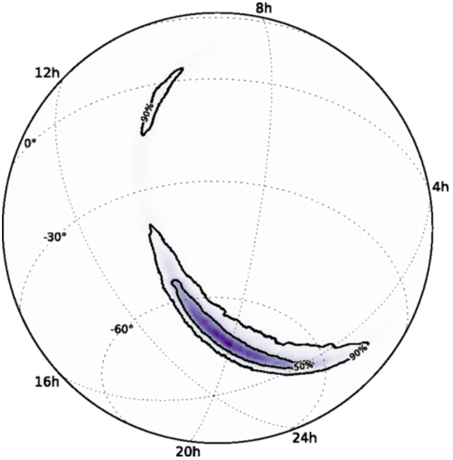


Unendlich Hypothesen und Parameterschätzung

- Würfel: 18 Hypothesen
- Verallgemeinerung zu unendlich Hypothesen
- Übergang z.B. von diskreten Punkten, die immer dichter werden
- Im Beispiel: $p(\phi, \theta|\mathrm{Signal})$
- In dem Fall Prior: Gleichverteilung im Raumwinkel
- Summe in Bayes' Formel wird Integral:

$$\boxed{p(\phi|d) = \frac{p(d|\phi) \cdot p(\phi)}{\int \mathrm{d}\phi' \mathrm{p}(\mathrm{d}|\phi') \cdot \mathrm{p}(\phi')}}$$

- Wichtig: Nenner hängt nicht von ϕ ab



Properties of the Binary Black Hole Merger GW150914, LIGO Kollaboration

Störparameter

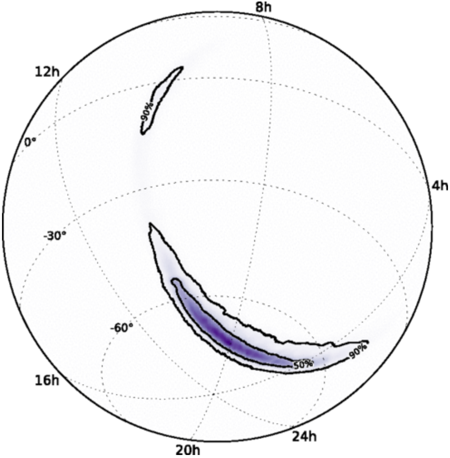
- Das Modell hat häufig unbekannte Parameter, die nicht interessant sind
- Z.B. wenn uns die Quelle der Gravitationswellen interessiert:

$$p(\theta, \phi, M_1, M_2 | \mathrm{Daten})$$

– auch die Massen sind im Modell

- Solche Parameter heißen "Störparameter"
- Um sie loszuwerden: Marginalisierung

$$p(x) = \int \mathrm{d}y \; p(x, y)$$



Properties of the Binary Black Hole Merger GW150914, LIGO Kollaboration

Beispiel – Zerfallskanäle

- $^{40}\mathrm{K}$ kann auf zwei Weisen zerfallen $^{40}\mathrm{K} \rightarrow \mathrm{e} + \mathrm{X}$ (A) oder $^{40}\mathrm{K} \rightarrow \gamma + \mathrm{X}'$ (B)
- Detektor misst Elektronen und Photonen mit gleicher Effizienz
- Nach einer bestimmten Zahl von Beobachtungen (jeweils A oder B): Was ist das Verzweigungsverhältnis (θ_A) für Zerfall A?
- Prior: Beide Zerfälle wurden beobachtet. Der Einfachheit halber: $p(\theta_A) = 1$ (zwischen 0 und 1)
- Wichtig: Dieser Prior schließt z.B. $\theta_A = 0$ aus. (Zerfall wurde ja beobachtet)

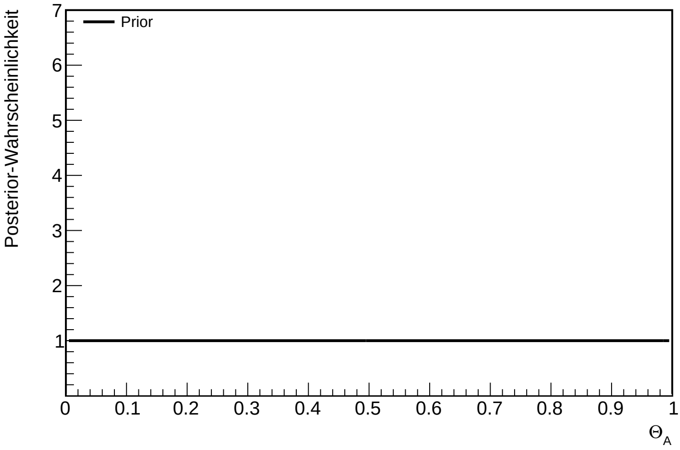
Beispiel – Zerfallskanäle (2)

$$\boxed{p(\theta_A|d) = \frac{p(d|\theta_A) \cdot p(\theta_A)}{\int \mathrm{d}\theta'_\mathrm{A}\ \mathrm{p}(\mathrm{d}|\theta'_\mathrm{A}) \cdot \mathrm{p}(\theta'_\mathrm{A})}}$$

- Prior: $p(\theta_A) = 1$
- Likelihood (Binomialverteilung)
  - Eine Messung (A oder B): $p(A|\theta_A) = \theta_A$, $p(B|\theta_A) = 1 - \theta_A$
  - N Messungen mit N_A mal Ergebnis A:

$$p(N_A|\theta_A, N) = \binom{N}{N_A} \theta_A^{N_A} (1 - \theta_A)^{N - N_A}$$

Beispiel – Zerfallskanäle (3)



Frequentistischer Wahrscheinlichkeitsbegriff

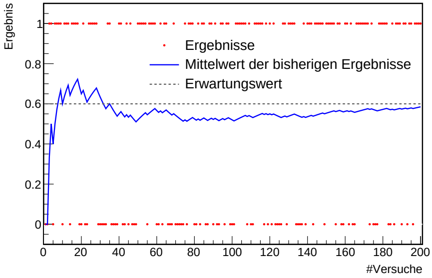
- $p(1) = \theta$, $p(0) = 1 - \theta$, bei N Messungen treten N_E Erfolge auf
- Bereits gesehen:

$$\lim_{n\to\infty} \frac{N_E}{N} = \theta$$

- Frequentistische Herangehensweise: Nutze dies als Definition für p:

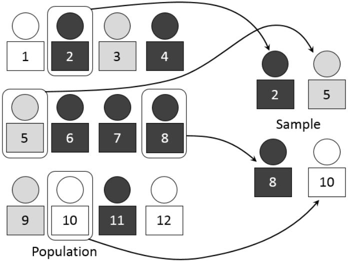
$$p(E) \equiv \lim_{n\to\infty} \frac{N_E}{N}$$

- D.h. nur Zufallsvariablen können eine Wahrscheinlichkeit besitzen



Grundgesamtheit und Stichprobe

- Endlicher Fall: Befrage eine Auswahl von Menschen, ziehe eine Auswahl von Bällen aus einem Behälter
- Grundgesamtheit: Die Menge, aus der Ausgewählt wird (Bevölkerung, Bälle, Tagesproduktion einer Maschiene)
- Stichprobe: Eine Untermenge, die Ausgewählt wird
- In Physik: Unendlicher Fall (Messung über Zeitraum, Münzwurf)
- Stichprobe sind die Messungen
- Grundgesamtheit komplizierter, unendlich mögliche Messungen



Quelle: Wikipedia

Frequentistische Unsicherheit

- Unsicherheit – wie variiert das Ergebnis bei Wiederholungen des Experiments
- Anders gesagt: Variation zwischen Stichproben der gleichen Grundgesamtheit
- Die genaue Definition der Grundgesamtheit kann große Unterschiede machen

Beispiel: Diagnose

Erinnerung: 1/100000 hat die Krankheit, Test in 99.95% der Fälle korrekt, Test positiv

Grundgesamtheit: Mögliche Tests des Patienten

- Wir nehmen an, alle Tests sind unkorreliert
- "Wenn ich den Test wiederhole, gibt er mir in 99.5% der Fälle das richtige Ergebnis"
- Anders gesagt: Wäre der Patient gesund, ist die Wahrscheinlichkeit für den positiven Test 0.0005, also gering

Grundgesamtheit: Mögliche Tests eines zufällig ausgewählten Patienten der Bevölkerung

- Hierbei wird von zwei Wahrscheinlichkeiten auf eine dritte geschlossen
- Es existiert eine relative Häufigkeit von "der Patient ist krank"
- Für relative Häufigkeiten gilt auch frequentistisch Bayes' Theorem
- $\implies p(k|+) = 0.02$

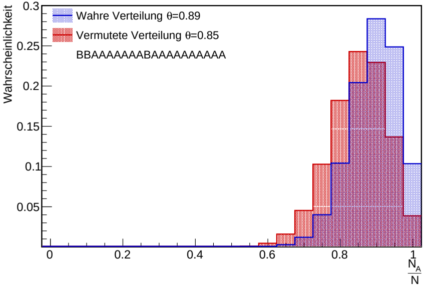
Beispiel: Diagnose (2)

- "Würde der Test wiederholt, ist das Ergebnis in 1999/2000 Fällen richtig."
- "Ein Patient mit positivem Test ist in 49/50 Fällen gesund."
- Beides korrekte Schlussfolgerungen
- **Wichtig:** Beides keine Aussagen über die Wahrscheinlichkeit dass *dieser bestimmte* Patient krank ist!
- Frequentistisch gibt es dafür keine Wahrscheinlichkeit – es ist wahr oder falsch
- Was bedeutet es, "dieses Experiment" zu wiederholen?

Frequentistische Parameterschätzung

- Kein Posterior – keine zentrale Größe, die berechnet wird
- Unsicherheit $\leftrightarrow$ Was geschieht bei Wiederholung des Experiments?
- Viele unterschiedliche Methoden
- zwei grobe Gruppen
  - Stelle Regel für Abschätzung des Wertes auf; wie ändert sich der geschätzte Wert bei Wiederholung des Experimentes?
  - Stelle Regel für Definition eines Intervalles auf; wie häufig enthält das Intervall den wahren Wert bei Wiederholung des Experimentes?

- Wahrer Wert θ_A bekannt – Unsicherheit gegeben durch $p(N_A|\theta_A, N)$
- Schätze θ_A ab durch N_A/N
- Verteilung gibt Unsicherheit
- $\rightarrow$ Die Unsicherheit hat eine Unsicherheit



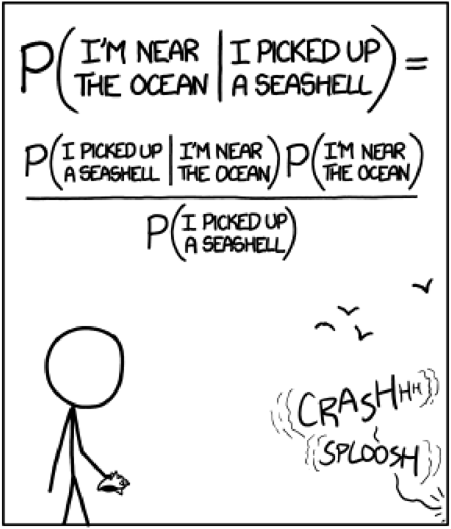
Frequentistische und Bayesianische Herangehensweise

	Frequentistisch	**Bayesianisch**
Wahrscheinlichkeits-begriff	Relative Häufigkeit	Grad der Überzeugung
Anwendung auf	Zufallsvariablen	Zufallsvariablen und Hypothesen
Datenanalyse	Die Daten sprechen für sich selbst	Daten ändern den Wissensstand
Unsicherheiten definiert durch	Verhalten bei Wiederholung des Experimentes	Posterior-Wahrscheinlichkeits-verteilung
Umgang mit Störparametern	Uneindeutig	Marginalisierung
Systematische Unsicherheiten	Separates Konzept	Wahrscheinlichkeits-verteilung
Nachteile	Schwierige Interpretation des Ergebnisses	Formulierung von Vorwissen als Prior

Frequentistische und Bayesianische Herangehensweise (2)

Bayesianer behandeln die Frage, die alle interessiert indem sie Annahmen machen, die keiner glaubt, während Frequentisten tadellose Logik benutzen um Probleme zu lösen, die niemanden interessieren.

Louis Lyons



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Quelle: xkcd.com