Einführung in die statistische Auswertung von Datensätzen

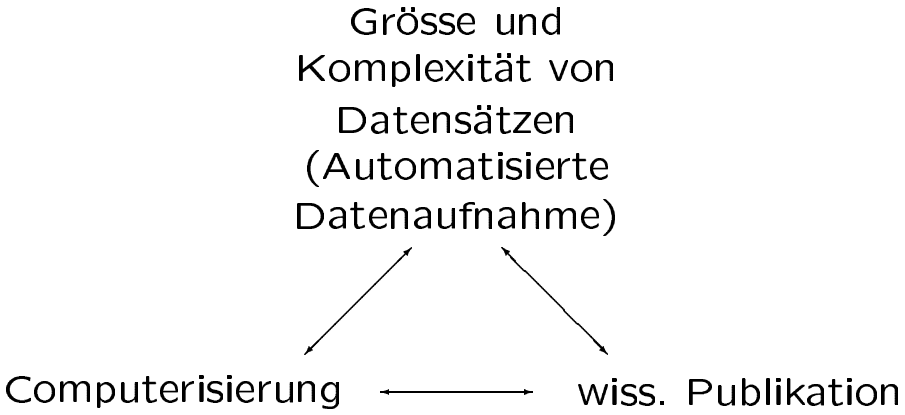
Lehrgang in wissenschaftlicher Ornithologie
für Studierende und fortgeschrittene Amateure

Lorenz Gygax
Statistician & proximate Biologist
brunner und hess software ag, Zürich

1. Juli 2000

Warum Statistik ?

- *saubere* Daten (Ausreisser finden)
- *Reduktion* von Daten
- eingängige und einfach zu interpretierende *graphische* Darstellungen
- quantitative *Beschreibung* von Daten
- quantitative Masse über *Unterschiede* und *Zusammenhänge*



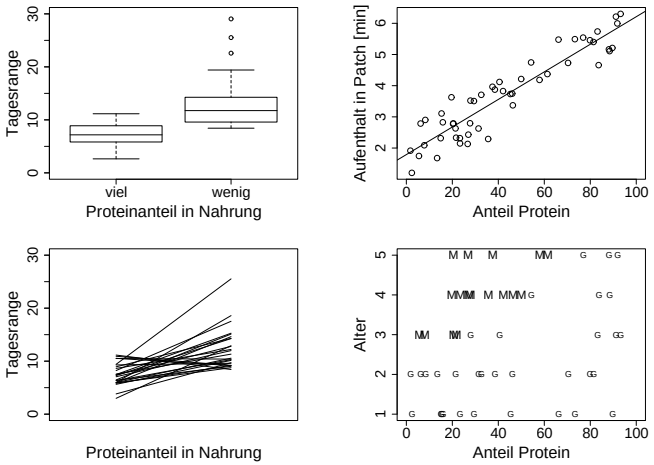
Inhalt

Vieles nur angetönt, Statistik eigenes Studium

- Graphisches
- Wann? Überhaupt / in einem Projekt
- Wie? Grundsätze statistischer Auswertungen
- Vorwissen
- Wahl der Methode, Überblick

Graphisches

- Darstellen von (meist) hochdimensionalen Daten in 2D
- Info auf einen Blick erkennbar, Lesbarkeit
- ideal: Replikate noch immer einzeln einsehbar
- Variabilität und Verteilung sichtbar
- Fantasie für den Spezialfall (Zeit !)
- Statistik "überflüssig"
- Flexibilität des Programmes



Idee von Statistik

Wie wahrscheinlich sind die Daten unter Annahme des Zufalls (p-Wert von H_0).

Widerspruchstest: H_0 kann abgelehnt, aber weder H_0 noch H_1 bewiesen werden.

Vorwissens $\rightarrow$ Problemwahrnehmung
"wir sehen, was wir glauben"

trotzdem: Wissen einsetzen $\rightarrow$ exakte Fragen

schliessende $\leftrightarrow$ explorative Statistik
eingeschränkte $\leftrightarrow$ offene Auswertung
Testen von Hypothesen $\leftrightarrow$ "Data-mining"
gute Kontrolle $\leftrightarrow$ viele Störeinflüsse
einfache Statistik $\leftrightarrow$ multivariate Verfahren
Experimente $\leftrightarrow$ Feldarbeit

Schätzungen
Voraussagen

Wann, wieviel

- Immer $\leftrightarrow$ Studie entsprechend designen
- d. h. vor Studienbeginn Überlegungen machen
- sich beraten lassen (Kontrolle, Baseline)
- multivariates Auswerten
  - Stärke des Einfluss wenn andere Variablen konstant gehalten
  - kein Verdecken von Einflüssen
- multiples Testen (α)

Replikate

- zufällig ausgewählte Beobachtungseinheiten
- $\rightarrow$ Stichprobengrösse ($N = 5 \cdot \#V$?)
  curse of dimensionality, Heterogenitäten
- statistische Unabhängigkeit (Individuen, Gebiete)
- 1 Datensatz pro Replikat (1 Zeile in Datenfile)
- Investition/Kosten
  # *Replikate* $\leftrightarrow$ Deatails für jedes Replikat

Fehlerarten

- *Fehler 1. Art* (α, p-Wert):
  P [H_0 fälschlich zu verwerfen]
- Multiples Testen
- *Fehler 2. Art* (β):
  P [H_0 fälschlich beizubehalten]
- *Macht* ($1 - \beta$):
  P [wahre Alternative anzunehmen]
- Power-Analysen

Skalen

- nominal/kategoriell, binär
- geordnet (ordinal), binär
- Anzahlen: > 0, Wurzeltransformation
- Beträge: > 0, kontinuierlich, Log-transf.
- intervall ($-\infty$ bis ∞)

Vorgehen

- Daten- und Variablenreduktion
- Wahl der Auswertungsmethode,
  was passt zu Daten?
- Wahl des spezifischen Modelles
  Einfachheit versus Detailliertheit
- Statistisches Testen
- Prüfen der Voraussetzungen/Annahmen
  (meist: Verteilung der Residuen)

Daten- und Variablenreduktion

- oft riesige Datenmengen (autom. Verfahren)
  $\rightarrow$ *Zufallsauswahl* (Muster bereits sichtbar)
- Problem: *Kollinearität*
  – Korrelierte Variablen (z. B. Wetter)
  – nicht unabhängig variierend
- Lösung: z. B. *PCA*
  – Interpretierbarkeit der Faktoren?
- Interpretation (Konstrukt), explorativ!

Wahl der Methoden

- Komplexität der Studie (KS)
- Skala der Zielvariable (ZV)
- Skala der erklärenden Variablen (EV)
- Verteilung der Fehler (VF)
- Schätzung der Fehler = Teilresultat

  Modelle rechnen $\leftrightarrow$ Modellwahl

KS: einfach, ZV: ordinal/intervall, EV: egal, VF: symmetrisch

Rangtests

	RANGTESTS (nicht-parametrisch, resampling)

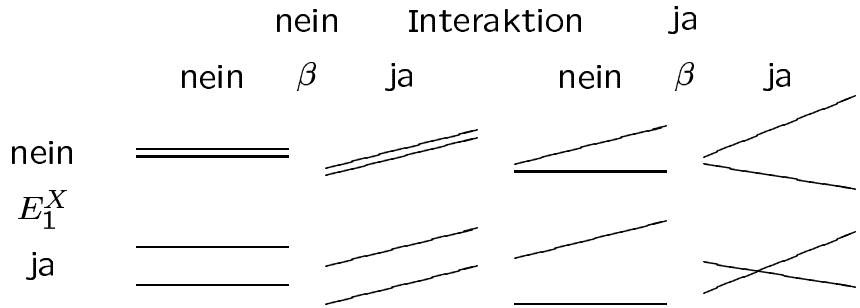
KS: (beliebig) komplex, ZV: > ordinal, EV: egal, VF: normalverteilt

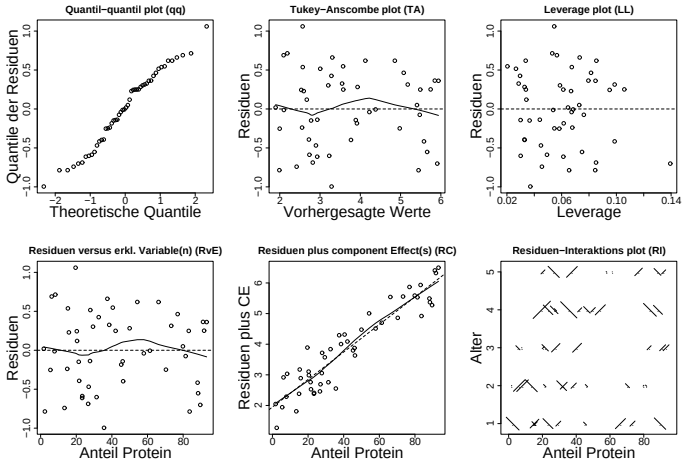
Lineare Modelle

- ANOVA:
  $Y = \mu + E^{X_1} + E^{X_2} + \ldots + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$
- Regression:
  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon_i,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$
- linear? X^2, $\log(X)$, $\sqrt{X}$, $\sin(X)$ möglich
- Residuenanalyse; Tests: Levene,
  Graphiken: qq, TA, LL, RvE, RC, RI plots

Interaktionen

- multiplikative Effekte (alle Skalen!):
  $Y = \alpha + E^{X_1} + \beta X_2 + \beta_I E^{X_1} X_2$





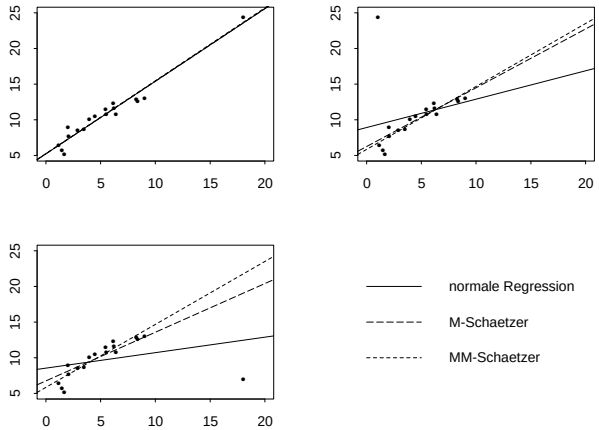
Wiederholte Messung ("gepaarte Daten")

- mehrere Messungen am selben Replikat
- Verlaufskurven
- Repeated Measure; Annahmen
- einige Kennzahlen schätzen pro Replikat
  - gute Charakteristika des Verlaufes
  - Maximum, Zeit bis . . . , Krümmung
- Auswertung pro Kennzahl

KS: komplex, ZV: > ordinal, EV: egal, VF: fast normalverteilt

Robuste Methoden

- Normalverteilung mit langen Schwänzen
- . . . oder kleiner Anzahl Ausreisser
- hat man eigentlich *immer*
- spezielle Methoden (eingesch. Implement.)
- z. B. Mittelwert versus Median
- z. B. Standardabweichung versus MAD
  $\text{mad} = C \cdot \text{med}(X_i - \text{med}(X))$



Generalisierte lineare Modelle

KS: komplex, ZV: ordinal, EV: egal, VF: multinominal

- kumulative Logits

KS: komplex, ZV: nominal/binär, EV: egal, VF: multin. normalvert.

- Diskriminanzanalyse

KS: komplex, ZV: nominal/binär, EV: egal, VF: multinominal

- Logistische Regression (siehe später)

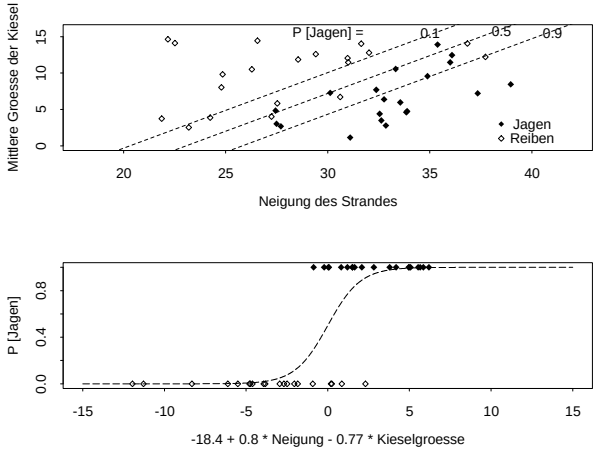
KS: komplex, ZV: Andere, EV: egal, VF: Andere

- log-lineare Modelle
- Poissonregression

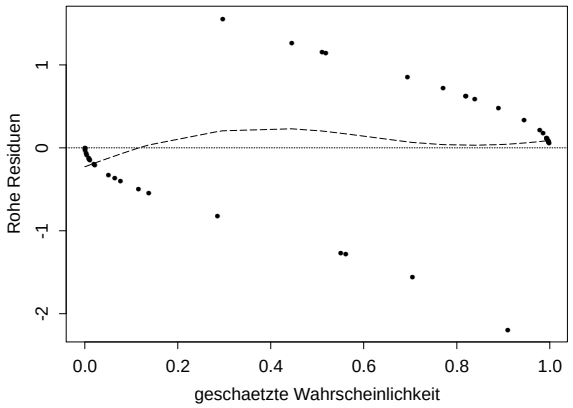
KS: komplex, ZV: binär, EV: egal, VF: binomial

Logistische Regression

- binäre Zielvariable
- Modell, das passt
- Interaktionen möglich
- Schätzwerte $\widehat{=}$ Wahrscheinlichkeiten


Mittlere Groesse der Kiesel
P [Jagen] =
0.1
0.5
0.9
Jagen
Reiben
Neigung des Strandes
P [Jagen]
-18.4 + 0.8 * Neigung - 0.77 * Kieselgroesse

25


Rohe Residuen
geschaetzte Wahrscheinlichkeit

26

Arbeiten mit %en

- Prozente nahe an 0, 1 nicht normalverteilt
- ev. arcsin $\sqrt{Y}$
- "Sum 1 constraint" $\leftrightarrow$ unabhängige Analyse
- ev. eine oder mehrere Anteile weglassen
- nur alle Ergebnisse zusammen interpretieren
- Compositional Analysis

Spezielles / Modernes

- nichtlineare Regression
- nichtparametrische Regression (z. B. lowess)
- resampling Verfahren: Jackknife / Bootstrap
  Fehlerverteilung aus Daten geschätzt
  so gut wie ursprüngliche Daten
- Zeitreihen (wenig Replikate, lange Reihen)
- neuronale Netze

Statistische Software

- an Anforderungen angepasst (Forscher und Problem)
- was ist vorhanden
- SPSS (www.spss.com)
- Glim, Genstat, SAS
- S-plus (www.mathsoft.com)
- R (www.ci.tuwien.ac.at/R)

Es gibt immer verschiedene Wege ein statistisches Problem anzugehen, verschiedene Methoden sollten bei starker Struktur zur gleichen Interpretation führen!

Statistik ist ein Hilfsmittel zur Entdeckung und quantitativen Beschreibung von Mustern und Zusammenhängen in hochdimensionalen Daten.

“Lehre jemanden den t–Test und er wird für einen Tag glücklich sein; lehre jemanden die Regression und er wird für eine Woche lang glücklich sein; lehre jemanden Statistik und er wird sein ganzes Leben lang Probleme haben.”

(unbekannter Statistiker)