Einführung in die Biostatistik

Vitaly Belik

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

Oct 06, 2020

Kontakt

Prof. Dr. Vitaly Belik

Fachbereich Veterinärmedizin

Institut für Veterinär-Epidemiologie und Biometrie

Juniorprofessor

Leitung Arbeitsgruppe Systemmodellierung

Adresse	Königsweg 67 Raum 104 14163 Berlin
Telefon	+49 30 838 61129
Fax	+49 30 838 4 61129
E-Mail	vitaly.belik@fu-berlin.de
Homepage	Working Group Modelling

Werdegang

2004	MSc Physics / Biochemical Physics, Moscow Lomonosov University
2000-2001	Physics, HU Berlin, unterstützt durch Siemens AG
2004-2013	MPI für Dynamik und Selbstorganisation, Göttingen
2008	Dr. rer. nat. in Theoretischer Physik, Georg-August-Universität Göttingen
2010-2012	Massachusetts Institute of Technology, Cambridge, MA, USA
2013-2015	Gastwissenschaftler am Helmholtz-Zentrum für Infektionsforschung, Braunschweig
2014-2016	TU Berlin
2016-	Professor (W1) FU Berlin, AG Systemmodellierung, Institut für Veterinär-Epidemiologie und Biometrie

Neue Medien – Persönlichkeitsrechte - Copyright

- Keine unautorisierten Foto-, Ton- und Video-Aufzeichnungen während der Vorlesung
- Kein unberechtigtes Weiterverbreiten oder “ins Netz stellen” von Inhalten aus der Veranstaltung
- Klare Kennzeichnung von Zitaten
- Quellenabgabe aller “fremder” Materialien

Was ist Statistik / Biostatistik (Biometrie)?

- Was sind Ihre Erwartungen?
- Go to http://menti.com

Was ist Statistik?

- Welche Daten soll man zur Beantwortung einer gegebenen Aufgabenstellung ermitteln?
- Wie viel Daten soll man ermitteln?
- Auf welche Art soll man das Untersuchungsmaterial auswählen?
- Wie soll mans eine Untersuchungsdaten ermitteln?
- Wie sollen die gewonnenen Daten geordnet werden?
- Wie sollen die Daten beschrieben und übersichtlich dargestellt werden?
- Wie wertet man die Daten aus?
- Welche Schlüsse lassen sich ziehen?
- Wie zuverlässig sind die getroffenen Aussagen?
- Welche weiterführenden Fragestellungen haben die Ergebnisse aufgeworfen?

Was ist Statistik? (1)

1. Aufgabenstellung. Nach präziser Formulierung der Fragestellung muss eine geeignete Wahl von Merkmalengetroffen,eine Mess-bzw. Beobachtungsmethode festgelegt und ein Versuchsplan aufgestellt werden.
2. Datengewinnung. Gewinnung des Untersuchungsmaterials (Ziehen der Stichprobe) und Ausführungder Messungen bzw. Beobachtungen an diesem Material.
3. Datenverarbeitung. Das gewonnene Datenmaterial muss graphisch und rechnerisch aufbereitet werden, dann sind Schlüsse von der Stichprobe auf die Grundgesamtheit zu ziehen; diese werden anschließend geprüft und interpretiert.

Was ist Statistik? (2)

Statistik

ist eine wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung von Methoden zur Datenerhebung, -beschreibung und -analyse sowie der Beurteilung der Ergebnisse ist. Dabei unterscheidet man:

Was ist Statistik? (2)

Statistik

ist eine wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung von Methoden zur Datenerhebung, -beschreibung und -analyse sowie der Beurteilung der Ergebnisse ist. Dabei unterscheidet man:

Deskriptive (beschreibende) Statistik:

Methoden zur Auswertung und übersichtlichen Darstellung und Zusammenfassung von Daten.

Was ist Statistik? (2)

Statistik

ist eine wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung von Methoden zur Datenerhebung, -beschreibung und -analyse sowie der Beurteilung der Ergebnisse ist. Dabei unterscheidet man:

Deskriptive (beschreibende) Statistik:

Methoden zur Auswertung und übersichtlichen Darstellung und Zusammenfassung von Daten.

Induktive (schliessende) Statistik:

Methoden zum Treffen von vernünftigen Entscheidungen im Falle von Unsicherheit bzw. Risiko. “Den Zufall in den Griff bekommen”. “Sicherheit über Unsicherheit gewinnen”.

Biostatistik oder Biometrie

Biostatistik (Biometrie)

angewandte Statistik zur Beschreibung, Modellierung und Beurteilung biologisch-naturwissenschaftlicher Phänomene.

Biostatistik oder Biometrie

Biostatistik (Biometrie)

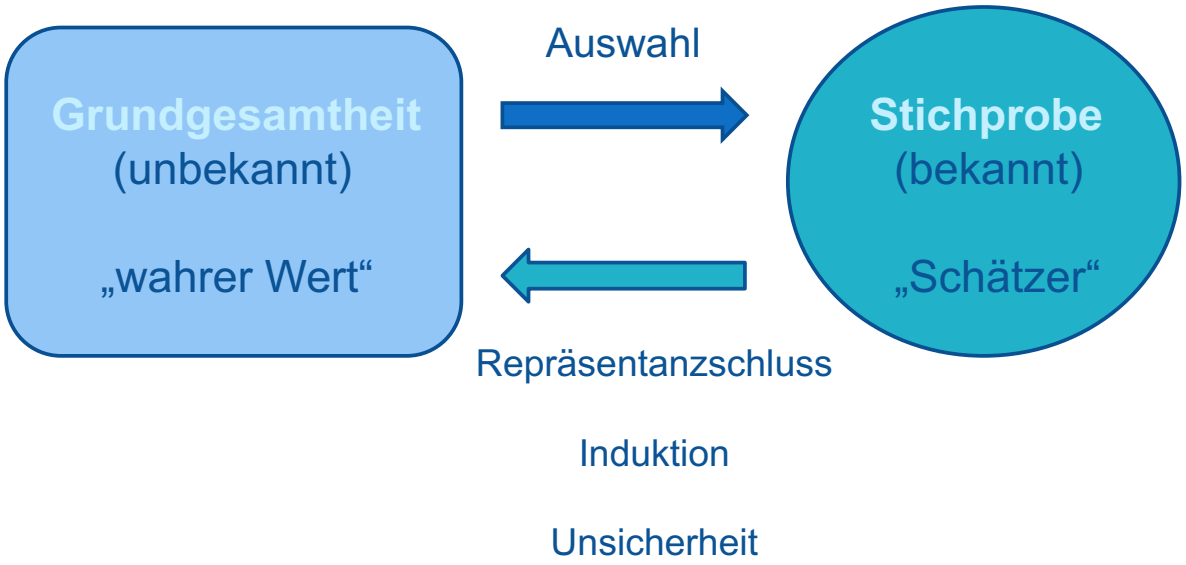
angewandte Statistik zur Beschreibung, Modellierung und Beurteilung biologisch-naturwissenschaftlicher Phänomene.

Beispiele

- Wie sicher ist das Ergebnis eines Diagnosetests zur Bestimmung einer Erkrankung?
- Wie viele Versuche müssen durchgeführt werden, um Verbesserung eines Produktes zu gewährleisten?

Deskriptive Statistik wird manchmal als *explorative* und Schliessende Statistik als *konfirmatorische Datenanalyse* bezeichnet.

Daten und Zufall



[Grafik: M. Doherr]

Statistische Schlüssweise

- Schätzen der unbekannten Parameter der Grundgesamtheit. “Finde eine Größe aus den Daten der Stichprobe, die”möglichst nah" an der unbekannten Wirklichkeit ist."
- Angabe von Konfidenzintervallen (Vertrauensbereichen). “Gebe basierend auf den Daten der Stichprobe ein Intervall an, das den wahren Wert (Populations-Parameter) mit einer gewissen Wahrscheinlichkeit überdeckt.”
- Entscheiden mittels eines statistischen Tests, ob anhand der Daten der Stichprobe eine Aussage über einen Parameter der Grundgesamtheit (bspw. Unterschied eines Mittelwertes zwischen Gruppen) wahr oder falsch ist.

Lernziele des Kurses

Ziel des Kurses is es Ihnen die wichtigsten statistischen Methoden zur Planung und Auswertung der Versuche und Daten aus wissenschaftlicher Studien zu vermitteln.

Sie sollen die Notwendigkeiten, Möglichkeiten und Grenzen grundlegender statistischer Analysen verstehen und selbst einfache statistische Berechnungen durchführen können.

Falls nötig, sollen Sie in der Lage sein bei einer statistischen Beratung, Ihr Anliegen sicher zu kommunizieren.

Kursübersicht

Nr	Date	Topic
1	06.10.20	Einführung. Arten von Daten.
2	13.10.20	Deskriptive Statistik
3	20.10.20	Schliessende Statistik. Parameterschätzung. Konfidenzeintervalle.
4	27.10.20	Hypothesentest, p-value
5	03.11.20	Zusammenhänge in Daten (kategoriell und kontinuierlich)
6	10.11.20	Grundlagen von Modellbildung. Lineare Regression. Modeldiagnostik
7	17.11.20	Generalized linear models. Mit einer unabhängigen und mehreren Variablen
8	24.11.20	Vergleich von zwei Mittelwerten ANOVA (einfache)
9	01.12.20	ANOVA 1 (mehrfache)
10	08.12.20	ANOVA 2 (mit Messwiederholungen)
11	15.12.20	Gemischte Modelle
12	05.01.21	Logistische Regression
13	12.01.21	Cluster- und Diskriminanzanalyse
14	19.01.21	Nichtparametrische Tests
15	26.01.21	Survival analysis
16	02.02.21	Elemente der Versuchsplanung
17	09.02.21	Konsultationen

Organisatorisches

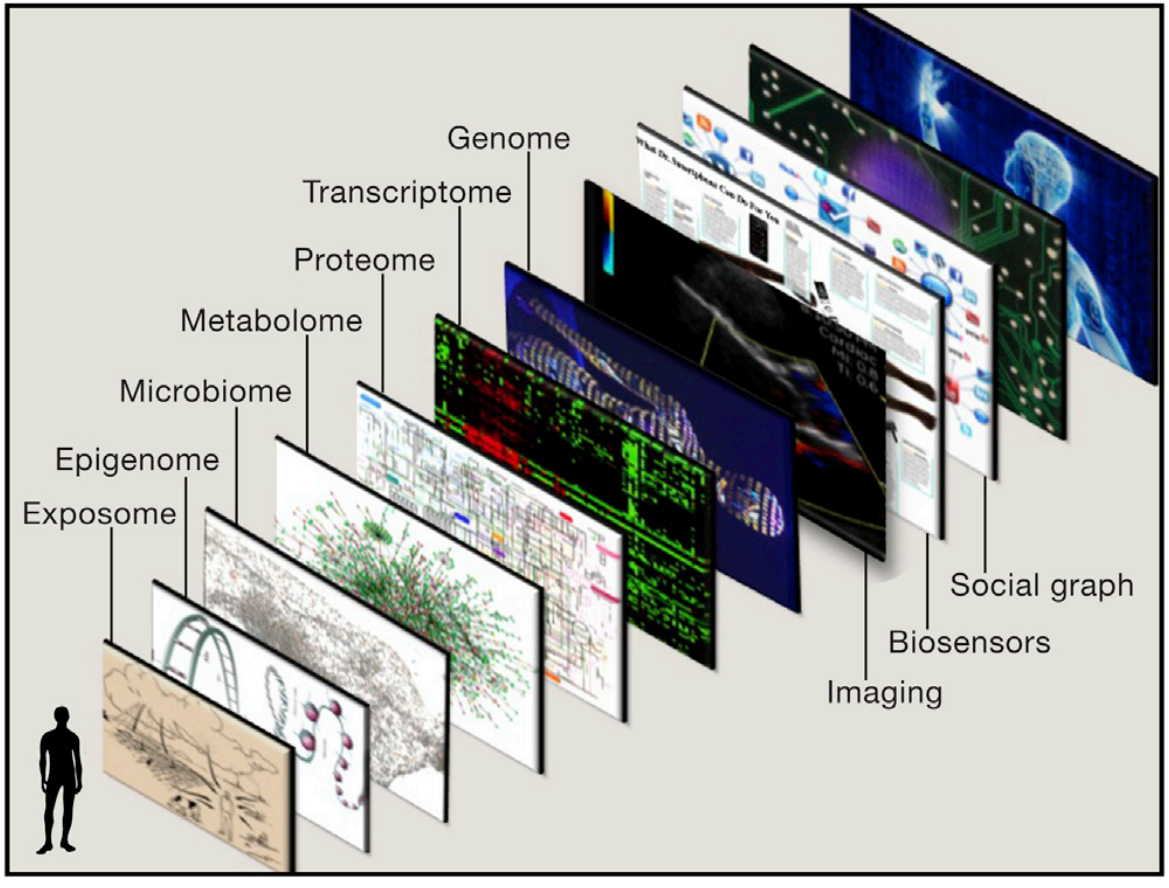
- Der Kurs besteht aus *Vorlesungen* und praktischen *Übungen*
- Für die Übungen wird Programmiersprache *R* (*RStudio*) benutzt
- Am Ende des Kurses ist eine *Klausur* vorgesehen
- Alternative ist es möglich ein *Projekt* zu bearbeiten (z.B. Analyse von Twitter)

Literatur

1. W. Köhler et al. *Biostatistik. Eine Einfürung für Biologen und Agrarwissenschaftler*
2. R. Kabacoff. *R in Action*
3. M. Crawley. *The R Book*
4. A. Field. *An Adventure in Statistics: The Reality Enigma*
5. A. Field *Discovering Statistics Using R*

Daten

In letzter Zeit mit der rasanten Entwicklung der ausgefallenen Sensoren (IoT), Rechner- und Speicherkapazitäten werden sehr viele Daten prozudiert.

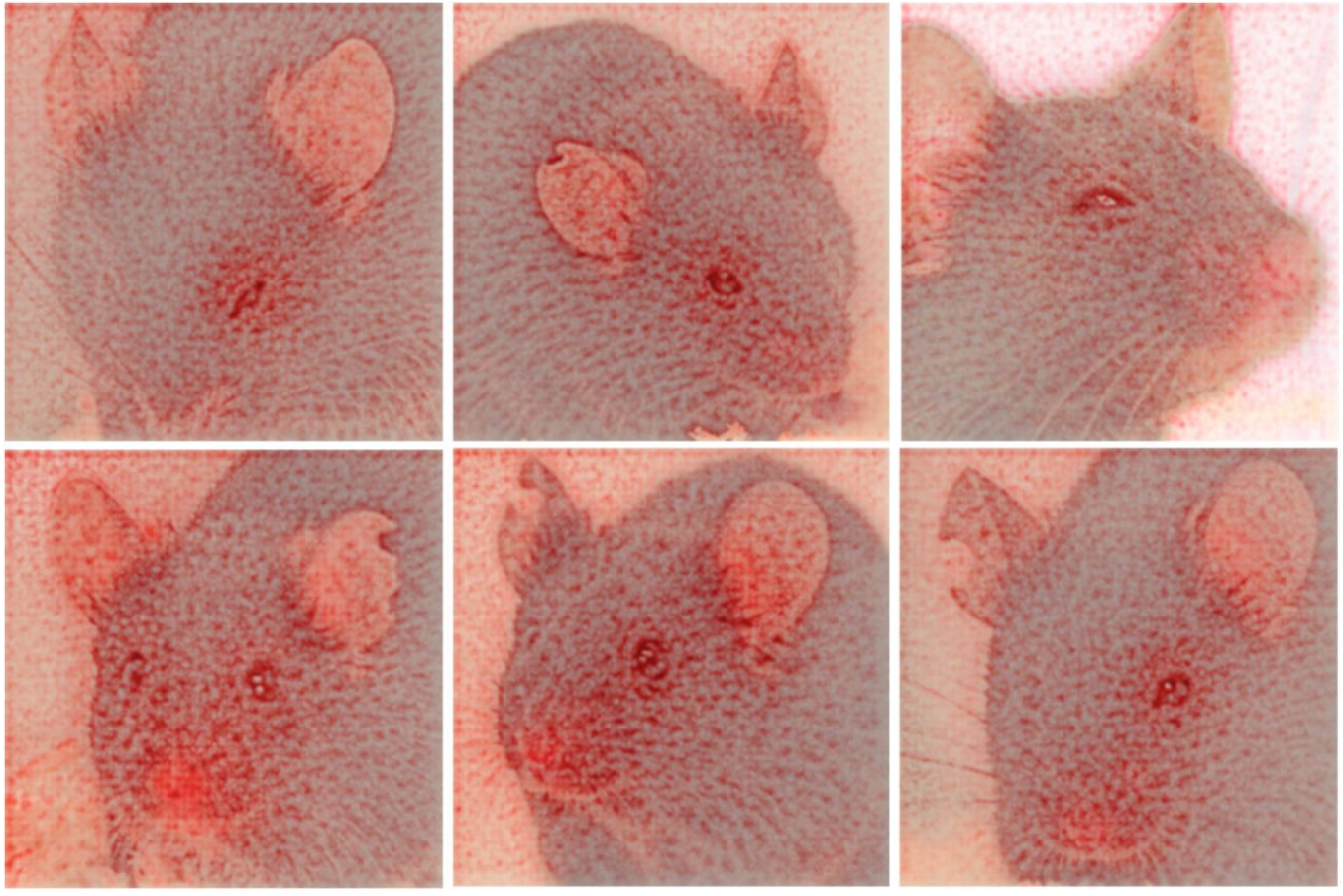


[Topol, 2014]

Daten sind heterogen

Bilder

Feature visualization von Mäuserbildern

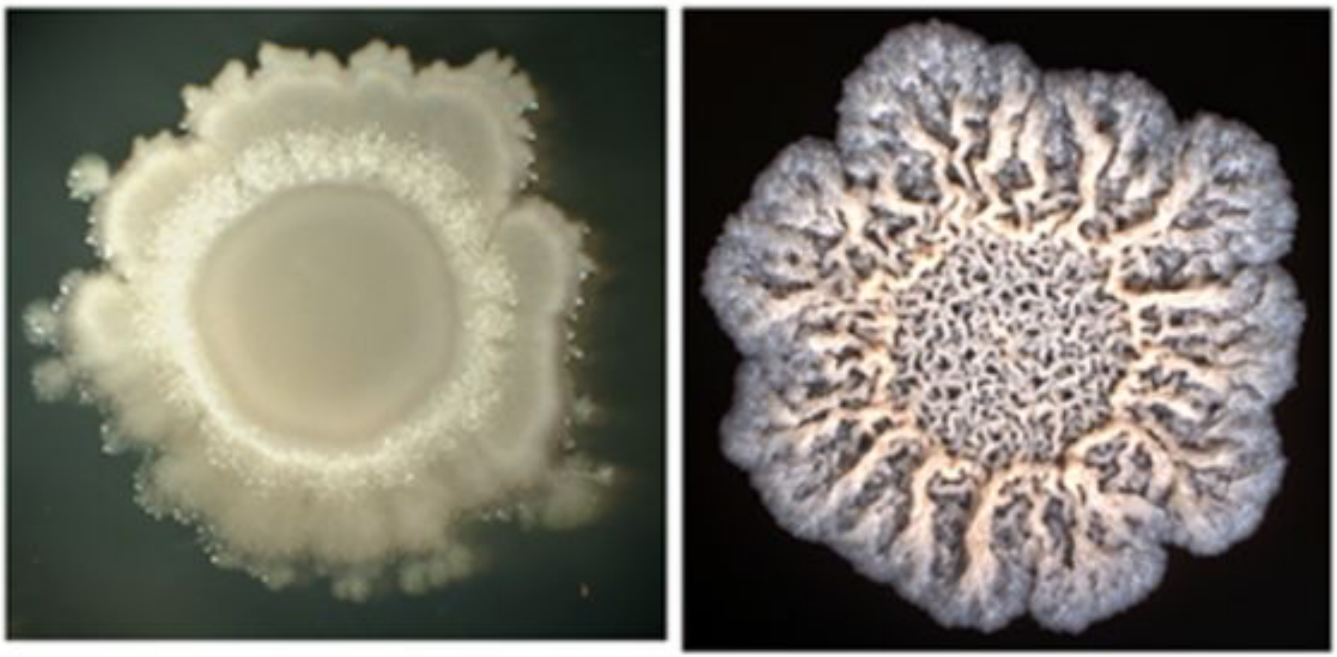


[https://doi.org/10.1371/journal.pone.0228059]

Daten sind heterogen (1)

Bilder

Biofilm von *B. subtilus*

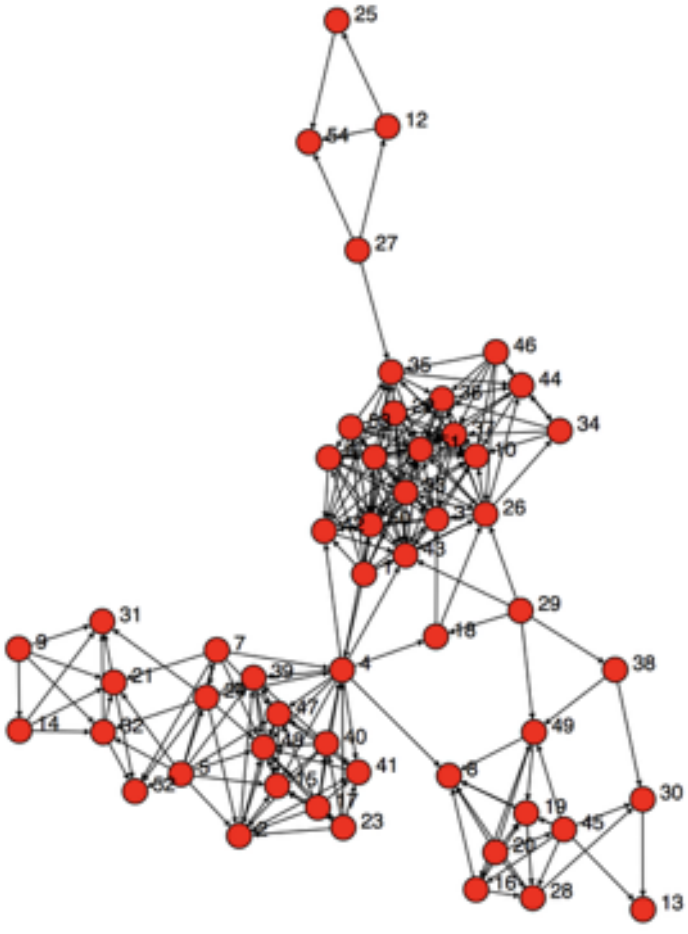


[https://doi.org/10.1128/JB.00028-13]

Daten sind heterogen (2)

Netzwerke

Kontaktnetzwerk von Tieren



[Daten: Thomas Selhorst]

Maschinelles Lernen

Große Mengen von heterogenen Daten (Big Data) verlangen nach entsprechenden Werkzeugen für die Datenanalyse. Dabei können, ausser klassischen statistischen Methoden, das *maschinelle Lernen* (z.B. *künstliche Neuronale Netze*) sehr hilfreich sein.

Maschinelles Lernen

Große Mengen von heterogenen Daten (Big Data) verlangen nach entsprechenden Werkzeugen für die Datenanalyse. Dabei können, ausser klassischen statistischen Methoden, das *maschinelle Lernen* (z.B. *künstliche Neuronale Netze*) sehr hilfreich sein.

Es stellt sich sogar die Frage, ob sich die Versuchsplanung und Datenanalyse nicht von einer Maschine erledigen lässt.

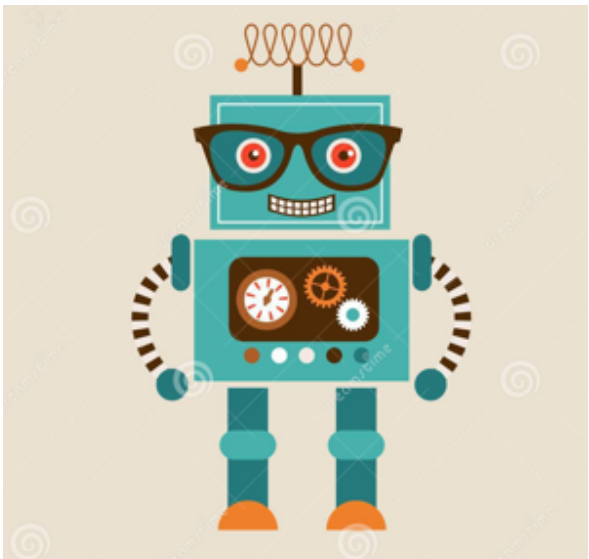
Roboscientist

SCIENCE VOL 324 3 APRIL 2009 **85**

The Automation of Science

Ross D. King,[1]* Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research, we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested treelike structure, 10 levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalization describes how a machine contributed to scientific knowledge.



Zurück zu eigentlichen Biostatistik!

Daten als Tabelle

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False
10	1	3	female	4.0	1	1	16.7000	S	Third	child	False	G	Southampton	yes	False
11	1	1	female	58.0	0	0	26.5500	S	First	woman	False	C	Southampton	yes	True
12	0	3	male	20.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
13	0	3	male	39.0	1	5	31.2750	S	Third	man	True	NaN	Southampton	no	False
14	0	3	female	14.0	0	0	7.8542	S	Third	child	False	NaN	Southampton	no	True
15	1	2	female	55.0	0	0	16.0000	S	Second	woman	False	NaN	Southampton	yes	True

Merkmale

Individuen oder Untersuchungsobjekte, die einer Erhebung / Untersuchung zu Grunde liegen, d.h. an / von denen Daten gesammelt werden, bezeichnet man als statistische Einheit, Merkmalsträger oder Untersuchungseinheiten.

Die Eigenschaften, die hinsichtlich des Untersuchungsziels an der statistischen Einheit untersucht werden, heißen *Merkmale*.

Studierendendaten

- Geschlecht, Körpergröße, Geburtsjahr
- Stadtnah oder ländlich aufgewachsen
- Wunsch, nach dem Studium in einem bestimmten Unternehmen zu arbeiten

Auswahl geeignete Merkmale

Objektivität

Die Ausprägung der zu ermittelneden Merkmales ist unabhängig von der Person des Auswerters eindeutig festszustellen.

Reliabilität

Das Merkmal gestattet reproduzierbare Mess- (bzw. Beobachtungs-) Ergebnisse, bei Wiederholung liegen also gleiche Resultate vor. Statt Reliabilität spricht man auch von “Zuverlässligkeit”.

Validität

Der Merkmal in seinen Ausprägungen spiegelt die für die Fragestellung wesentlichen Eigenschaften wider. Auch “Gültigkeit” oder “Aussagekraft” genannt.

Charakterisierung von Merkmalen

quantitative Merkmale:

Untersuchungseinheiten unterscheiden sich im absoluten (Zahlen-) Wert. - z.B. Alter, Gewicht, Temperatur, Anzahl Keime, Betriebsgröße, Schadstoffgehalt, ...

Charakterisierung von Merkmalen

quantitative Merkmale:

Untersuchungseinheiten unterscheiden sich im absoluten (Zahlen-) Wert. - z.B. Alter, Gewicht, Temperatur, Anzahl Keime, Betriebsgröße, Schadstoffgehalt, ...

qualitative Merkmale:

Untersuchungseinheiten unterscheiden sich in ihrer Ausprägung (Art) - z.B. Geschlecht, Name, Befund, Rasse, Therapie, Haltungsform, Region, ...

Skalenniveaus von Merkmalen

nominale (qualitative) Skala:

die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar.

- z.B. Name, Geschlecht, Rasse, Haltungsform, Therapieform, pathologische Klassifikation

Skalenniveaus von Merkmalen

nominale (qualitative) Skala:

die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar.

- z.B. Name, Geschlecht, Rasse, Haltungsform, Therapieform, pathologische Klassifikation

ordinale (qualitativ oder semiquantitative) Skala:

die Werte unterliegen einer Rangfolge, aber die Abstände zwischen den Werten der Skala lassen sich nicht interpretieren.

- z.B. Bewertung (Bonituren, Noten), Gesundheitszustand, Grad der Belastung mit Keimen (-, +, ++, +++)

Skalenniveaus von Merkmalen

nominale (qualitative) Skala:

die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar.

- z.B. Name, Geschlecht, Rasse, Haltungsform, Therapieform, pathologische Klassifikation

ordinale (qualitativ oder semiquantitative) Skala:

die Werte unterliegen einer Rangfolge, aber die Abstände zwischen den Werten der Skala lassen sich nicht interpretieren.

- z.B. Bewertung (Bonituren, Noten), Gesundheitszustand, Grad der Belastung mit Keimen (-, +, ++, +++)

metrische (quantitative) Skala:

die Werte unterliegen einer Rangfolge und die Abstände zwischen den Werten der Skala lassen sich interpretieren.

- z.B. Gewicht, Betriebsgröße, Keimzahlen, ...

Skalenniveaus von Merkmalen

nominale (qualitative) Skala:

die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar.

- z.B. Name, Geschlecht, Rasse, Haltungsform, Therapieform, pathologische Klassifikation

ordinale (qualitativ oder semiquantitative) Skala:

die Werte unterliegen einer Rangfolge, aber die Abstände zwischen den Werten der Skala lassen sich nicht interpretieren.

- z.B. Bewertung (Bonituren, Noten), Gesundheitszustand, Grad der Belastung mit Keimen (-, +, ++, +++)

metrische (quantitative) Skala:

die Werte unterliegen einer Rangfolge und die Abstände zwischen den Werten der Skala lassen sich interpretieren.

- z.B. Gewicht, Betriebsgröße, Keimzahlen, ...

Skalenniveaus von Merkmalen (1)

Es wird auch unterschieden zwischen

Intervallskala

Die Abstände zwischen Merkmalsausprägungen lassen sich vergleichen.
Die Skalla ist kontinuierlich.

- z.B. Temperatur in Grad Celsius

Verhältnisskala

Nicht nur die Differenz, sondern auch der Quotient aus zwei Messwerten darf verwendet werden.

- z.B. Temperatur in Kelvin, Länge in Zentimetern

Skalenniveaus von Merkmalen (2)

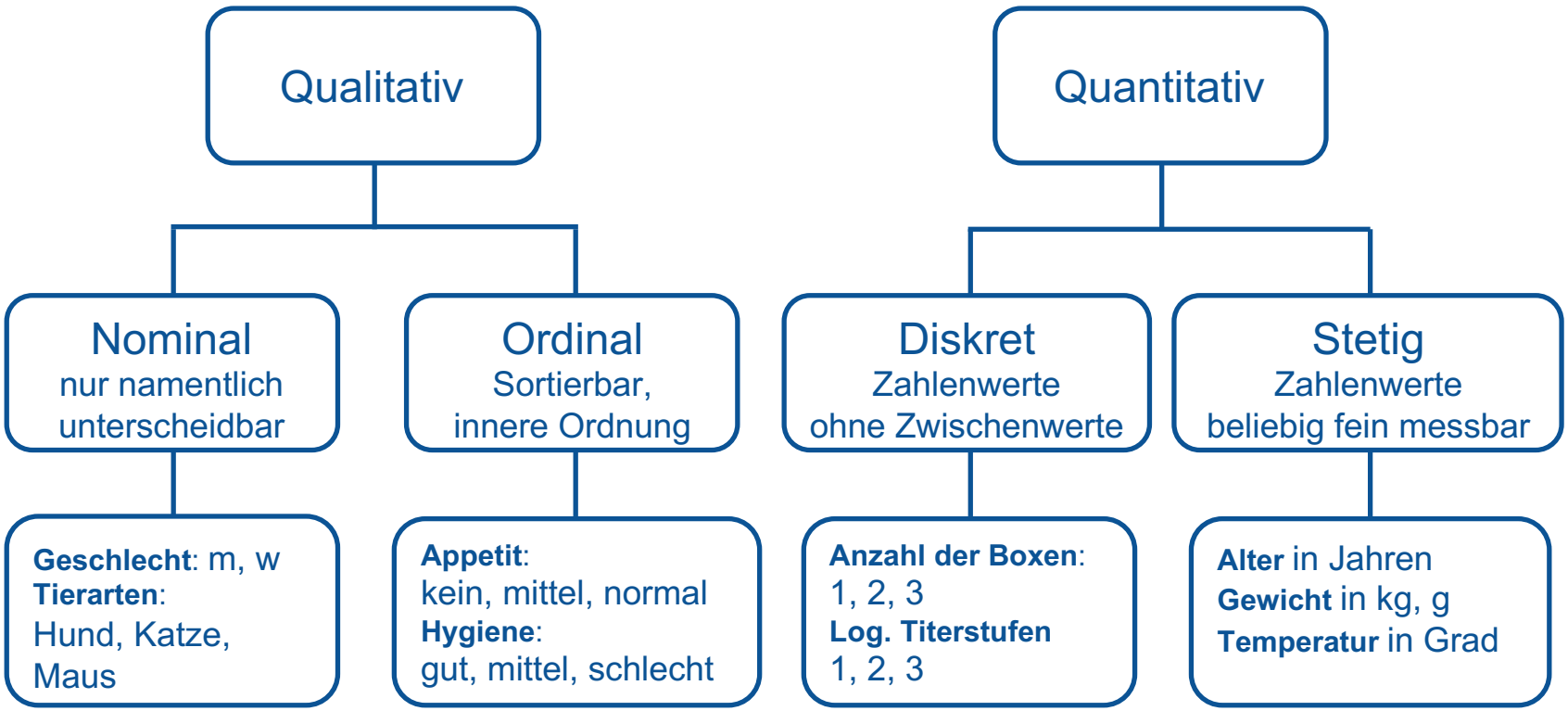
Die statistischen Auswertungsmöglichkeiten sind vom Skalenniveau abhängig, weil auf höherem Niveau mehr Information festgehalten und ausgewertet werden kann, als bei niedrigeren Skalierungen.

Skalenniveaus von Merkmalen (2)

Die statistischen Auswertungsmöglichkeiten sind vom Skalenniveau abhängig, weil auf höherem Niveau mehr Information festgehalten und ausgewertet werden kann, als bei niedrigeren Skalierungen.

Debei soll den Aufwand für den zusätlichen Informationgewinn berücksichtigt werden.

Skalenniveaus von Merkmalen (3)



Nicht jede Zahl ist eine Zahl.

Häufig werden Daten verschlüsselt, um die anschliessende Datenverarbeitung zu erleichtern

- Schulnoten: 1, 2, 3, 4, 5, 6 (ordinal)
- Testergebnis: 1, 0 (nominal)
- Kreiskennziffern: 3253, 3351 (nominal)
- Zuchtbuch-Nummern: 0511572 (nominal)

Einfache Datendarstellungen