Werkzeuge der empirischen Forschung

Wolfgang Kössler

Institut für Informatik, Humboldt-Universität zu Berlin
SS2008

18. April 2008

Übersicht

1. Einleitung
2. Dateneingabe und Transformation
   - Allgemeine Syntax
   - Eingabe über die Tastatur
   - Transformationen
   - Eingabe durch externes File
   - Wichtige Varianten der INPUT-Anweisung
   - Ein- u. Ausgabe von SAS-Files
   - Zusammenfügen von Files
   - Output-Anweisung
   - DO-Schleifen im DATA-Step
1. Wahrscheinlichkeitsrechnung
   - Grundgesamtheit, Population
   - Wahrscheinlichkeit
   - Zufallsvariablen

Wahrscheinlichkeitsrechnung
Übersicht

3.1 Grundbegriffe

3.2 Wahrscheinlichkeit

3.3 Zufallsvariable

3.4 Diskrete Zufallsvariablen
Binomial-, Poisson-, geometrisch

3.5 Stetige Zufallsvariablen
Exponential-, Gleich-, Normal

3.6 Normalverteilung

3.7 Erwartungswert

3.8 Varianz

3.9 Besondere Eigenschaften der NV

Grundbegriffe

Grundgesamtheit, Population

Eine Grundgesamtheit (oder Population)

ist eine Menge von Objekten, die gewissen Kriterien genügen. Die einzelnen Objekte heißen <u>Individuen</u>.

- Menge aller Haushalte
- Menge aller Studenten
- Menge aller Studenten der HUB
- Menge aller Einwohner von GB
- Menge aller Heroin-Abhängigen
- Menge aller Bewohner Tibets
- Menge aller verschiedenen Computer
- Menge aller Schweizer Franken

Grundbegriffe

Zufällige Stichprobe

Die gesamte Population zu erfassen und zu untersuchen ist meist zu aufwendig, deshalb beschränkt man sich auf zufällige Stichproben.

Zufällige Stichprobe

Eine zufällige Stichprobe ist eine zufällige Teilmenge der Grundgesamtheit, bei der jedes Element mit ‘der gleichen Wahrscheinlichkeit’ ausgewählt wird.

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Nominale Merkmale

Die Ausprägungen sind lediglich Bezeichnungen für Zustände oder Sachverhalte.
Sie können auch durch Zahlen kodiert sein!

Bsp: Familienstand, Nationalität, Beruf

Dichotome Merkmale

Hat das (nominale) Merkmal nur 2 Ausprägungen, so heißt es auch binär oder dichotom.

gut - schlecht
männlich - weiblich
wahr - falsch

Klassifikation von Merkmalen

Ordinale und metrische Merkmale

Ordinale Merkmale (Rangskala)

Die Menge der Merkmalsausprägungen besitzt eine Rangordnung!

Rangzahlen einer Rangliste (z.B. beim Sport)
Härtegrade
Schulzensuren

Metrische Merkmale (kardinale/quantitative M.)

Werte können auf der Zahlengeraden aufgetragen werden (metrische Skala)

Meßwerte, Längen, Größen, Gewichte, Alter

Klassifikation von Merkmalen

Metrische Merkmale

Metrische Merkmale werden unterschieden nach:

Diskrete Merkmale

nehmen höchstens abzählbar viele Werte an.

Alter, Länge einer Warteschlange

Stetige Merkmale

können Werte in jedem Punkt eines Intervalls annehmen, z.B. $x \in [a, b]$, $x \in (-\infty, \infty)$.

Metrische Merkmale sind immer auch ordinal.

Grundbegriffe

Stichprobenraum

Der Stichprobenraum Ω eines zufälligen Experiments

ist die Menge aller möglichen Versuchsausgänge
Die Elemente ω des Stichprobenraums Ω heißen Elementarereignisse.

- Münzwurf $\Omega = \{Z, B\}$
- Würfel $\Omega = \{1, ..., 6\}$
- Qualitätskontrolle $\Omega = \{$gut, schlecht$\}$
- Lebensdauer einer Glühlampe $\Omega = [0, \infty)$
- 100m - Zeit $\Omega =$[9.81, 20)
- Blutdruck, Herzfrequenz
- Länge einer Warteschlange $\Omega = \{0, 1, 2, ...\}$
- Anzahl der radioaktiven Teilchen beim Zerfall
- Wasserstand eines Flusses $\Omega = [0, \ldots)$

Grundbegriffe

Ein Ereignis ist eine Teilmenge $A, A \subseteq \Omega$

Lebensdauer ≤ 10 min.
Augensumme gerade.
Warteschlange hat Länge von ≤ 10 Personen.

Realisierungen sind die Ergebnisse des Experiments

(die realisierten Elemente von Ω)

Verknüpfungen von Ereignissen werden durch entsprechende Mengenverknüpfungen beschrieben

$A \cup B$	A oder B tritt ein
$A \cap B$	A und B tritt ein
$\overline{A} = \Omega \setminus A$	A tritt nicht ein.

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Forderung (damit die Verknüpfungen auch immer ausgeführt werden können):
Die Ereignisse liegen in einem Ereignisfeld (σ-Algebra) $\mathfrak{E}$.

Ereignisfeld

Das Mengensystem $\mathfrak{E} \subseteq \mathfrak{P}(\Omega)$ heißt <u>Ereignisfeld</u>, falls gilt:

1. $\Omega \in \mathfrak{E}$
2. $A \in \mathfrak{E} \Longrightarrow \overline{A} \in \mathfrak{E}$
3. $A_i \in \mathfrak{E}, i = 1, 2, \ldots \Longrightarrow \bigcup_{i=1}^{\infty} \in \mathfrak{E}$.

Wahrscheinlichkeit

Das Axiomsystem von Kolmogorov

Sei $\mathfrak{E}$ ein Ereignisfeld. Die Abbildung

$$P : \mathfrak{E} \longrightarrow \mathbb{R}$$

heißt Wahrscheinlichkeit, falls sie folgende Eigenschaften hat:

1. Für alle $A \in \mathfrak{E}$ gilt: $0 \leq P(A) \leq 1$.
2. $P(\Omega) = 1$.
3. $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i),$$
   falls $\quad A_i \cap A_j = \emptyset \quad \forall i, i \neq j$

Wahrscheinlichkeit

Eigenschaften (1)

$P(\overline{A}) = 1 - P(A)$.

Beweis:

$$\begin{aligned} 1 &= P(\Omega) && \text{Axiom 2} \\ &= P(A \cup \overline{A}) \\ &= P(A) + P(\overline{A}) && \text{Axiom 3} \end{aligned}$$

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
- Syntax
- Tastatur
- Transformationen
- Externes File
- Input-Anweisung
- SAS-Files
- Zusammenfügen
- Output-Anweisung
- DO-Schleifen

Wkt.rechnung
- Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Beweis:

$$\begin{aligned}
P(A \cup B) &= P((A \cap B) \cup (A \cap \overline{B}) \cup (B \cap \overline{A})) \\
&= \underbrace{P(A \cap B) + P(A \cap \overline{B})} \\
&\quad + P(B \cap \overline{A}) \quad \text{Axiom 3} \\
&= P(A) + \underbrace{P(B \cap \overline{A}) + P(A \cap B)} \\
&\quad - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

Zufallsvariablen

Eine (meßbare) Abbildung heißt Zufallsvariable.

$$\begin{aligned} X: \quad \Omega &\longrightarrow \mathbb{R} \\ \omega &\longrightarrow r \end{aligned}$$

Diskrete Zufallsvariable

Die Zufallsvariable X heißt diskret, wenn X nur endlich viele oder abzählbar unendlich viele Werte x_i annehmen kann. Jeder dieser Werte kann mit einer gewissen Wkt. $p_i = P(X = x_i)$ auftreten. ($p_i > 0$)

- geografische Lage (N,O,S,W)
- Länge einer Warteschlange
- Anzahl der erreichten Punkte in der Klausur.

Stetige Zufallsvariable

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Stetige Zufallsvariable

Die Zufallsvariable X heißt stetig, falls X beliebige Werte in einem Intervall (a, b), $[a, b]$, $(a, b]$, $(a, b]$, $(-\infty, a)$, (b, ∞), $(-\infty, a]$, $[b, \infty)$, $(-\infty, \infty)$ annehmen kann.

- Wassergehalt von Butter
- Messgrößen (z.B. bei der Banknote)
- Lebensdauer von Kühlschränken

Verteilungsfunktion

Diskrete Zufallsvariable

$$F_X(x) := P(X \leq x) = \sum_{i:i \leq x} p_i = \sum_{i=0}^{x} p_i$$

heißt Verteilungsfunktion der diskreten zufälligen Variable X

Stetige Zufallsvariable

Die Zufallsvariable X wird mit Hilfe der sogen. Dichtefunktion f beschrieben,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Diskrete Zufallsvariablen

Bezeichnung

$$X \in \{x_1, x_2, x_3, \ldots\}$$

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots p_n & \cdots \end{pmatrix}$$

$$p_i = P(X = x_i) > 0, \quad i = 1, 2, 3, \ldots$$

$$\sum_{i=1}^{\infty} p_i = 1$$

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Zweimaliges Werfen einer Münze

$\Omega = \{ZZ, ZB, BZ, BB\}$, $\quad X :=$ Anzahl von Blatt

$$X : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Erfolge bei n Versuchen

X: Anzahl der “Erfolge” bei n Versuchen, wobei jeder der n Versuche eine Erfolgswahrscheinlichkeit p hat.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Binomialwkt.}$$

$$F_X(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \quad \text{Vf.}$$

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
- Syntax
- Tastatur
- Transformationen
- Externes File
- Input-Anweisung
- SAS-Files
- Zusamenfügen
- Output-Anweisung
- DO-Schleifen

Wkt.rechnung
- Population
- Wahrscheinlichkeit
- Zufallsvariablen
- Diskrete Zufallsvariablen
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert
- Varianz

Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

Einleitung
Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen
Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Diskrete Zufallsvariablen

Übungsaufgabe

Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

X: Anzahl der Sechsen.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - \sum_{i=0}^{3} P(X = i)$$

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

X: Anzahl der Sechsen.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - \sum_{i=0}^{3} P(X = i)$$

$$= 1 - \left(\frac{5}{6}\right)^{20} - 20\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{19} - \frac{20 \cdot 19}{2}\left(\frac{1}{6}\right)^{2}\left(\frac{5}{6}\right)^{18} - \frac{20 \cdot 19 \cdot 18}{6}\left(\frac{1}{6}\right)^{3}\left(\frac{5}{6}\right)^{17}$$

= 1 - CDF('Binomial',3,$1/6$,20)
= SDF('Binomial',3,$1/6$,20)
$\approx$ 0.43.

Diskrete Zufallsvariablen

Poisson (1)

X: Anzahl der Anrufe pro Zeiteinheit

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \end{pmatrix}$$

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \qquad \lambda > 0$$

$$\sum_{i=0}^{\infty} p_i = \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}}_{e^{\lambda}} e^{-\lambda} = 1.$$

Bez.: $X \sim Poi(\lambda)$, wobei λ ein noch unbestimmter Parameter ist. Er kann als mittlere Rate aufgefaßt werden.

Diskrete Zufallsvariablen
Poisson (2), Motivation

Sei $\{N_t\}_{t \in T}$ eine Menge von Zufallsvariablen (ein stochastischer Prozeß) mit den Eigenschaften:

V1: Zuwächse sind unabhängig, dh. die Zufallsvar. $N_{t+h} - N_t$ und $N_t - N_{t-h}$ sind unabhängig

V2: es ist egal wo wir das Zeitintervall betrachten, dh N_{t+h} und N_t haben dieselbe Verteilung

V3: Wkt., daß mindestens ein Ereignis in der Zeit h eintritt, z.B. ein Kunde ankommt.
$p(h) = a \cdot h + o(h), \quad a > 0, h \to 0$

V4: Wkt. für $k \geq 2$ Ereignisse in der Zeit h: $o(h)$

Diskrete Zufallsvariablen

Poisson (3)

Frage: Wkt. bis zum Zeitpunkt t genau i Ereignisse?
(eingetroffene Kunden, zerfallene Teilchen)

$$P_k(t) := P(N_t = k), \qquad P_k(t) = 0 \quad \text{für} \quad k < 0$$

$$P_k(t) = \frac{a^k t^k}{k!} e^{-at}, \qquad k \geq 0$$

Poisson-Verteilung mit Parameter $\lambda = at$.

Diskrete Zufallsvariablen
Poisson (4)

Binomial und Poisson

Seien $X_n \sim Bi(n,p)$ $\qquad Y \sim Poi(\lambda)$

Für $n \cdot p = \lambda$ gilt: $\qquad P(X_n = k) \longrightarrow_{n\to\infty} P(Y = k)$.

Beweis:

$$\begin{aligned}
P(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\
&= \frac{1}{k!} \underbrace{\frac{n(n-1)\cdots(n-k+1)}{(n-\lambda)^k}}_{\to 1} \lambda^k \underbrace{(1-\frac{\lambda}{n})^n}_{\to e^{-\lambda}}
\end{aligned}$$

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Münzwurf solange bis B(Blatt) kommt

Ω = {B, ZB, ZZB, ...}

X := Anzahl der Würfe bis zum ersten Blatt.

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ (1/2) & (1/2)^2 & (1/2)^3 & (1/2)^4 & \cdots & (1/2)^n & \cdots \end{pmatrix}$$

$$\sum_{i=1}^{\infty} p_i = \sum_{i=1}^{\infty} (1/2)^i = \frac{1}{1-\frac{1}{2}} - 1 = 1$$

geometrische Reihe

geometrische Verteilung mit p=1/2, $p_i = (1/2)^i$.

allgemeiner: $p_i = p^{i-1}(1-p)$.

Diskrete Zufallsvariablen

Hypergeometrische Verteilung (1)

Qualitätskontrolle

Warenlieferung mit N Stücken, davon genau n schlecht. Frage: Wkt., daß in einer Stichprobe vom Umfang m höchstens k Stück schlecht sind?

Diskrete Zufallsvariablen

Hypergeometrische Verteilung (1)

Qualitätskontrolle

Warenlieferung mit N Stücken, davon genau n schlecht. Frage: Wkt., daß in einer Stichprobe vom Umfang m höchstens k Stück schlecht sind?

X: Anzahl der schlechten Stücke in der Stichprobe.

$$P(X = x) = \frac{\binom{n}{x} \cdot \binom{N-n}{m-x}}{\binom{N}{m}}$$

$\binom{N}{n}$: $\#$ möglichen Stichproben.

$\binom{n}{x}$: $\#$ Möglichkeiten, aus n schlechten Stücken in der Population x schlechte Stücke zu ziehen.

$\binom{N-n}{m-x}$: $\#$ Möglichkeiten, aus $N - n$ guten Stücken in der Population $m - x$ gute Stücke zu ziehen.

Offenbar:

$0 \leq x \leq \min(n, m)$

$m - x \leq N - n.$

Eine Zufallsvariable mit der Verteilungsfunktion

$$F(k|H_{N,n,m}) = \sum_{x=0}^{k} \frac{\binom{n}{x} \cdot \binom{N-n}{m-x}}{\binom{N}{m}}$$

heißt hypergeometrisch verteilt.

Bemerkung: Für $N \to \infty$, $n \to \infty$, $\frac{n}{N} \to p$ gilt:

$$f(x|H_{N,n,m}) \to \binom{m}{x} p^x (1-p)^{m-x} = f(x|Bi(m,p))$$

SAS-Anweisungen

CDF('Binomial',m,p,n)	PDF('Binomial',m,p,n)
CDF('Poisson',m,λ)	PDF('Poisson',m,λ)
CDF('Geometric',m,p)	PDF('Geometric',i,p)
CDF('Hyper',K,N,n,m)	PDF('Hyper',k,N,n,m)

```
Descr_Binomial_neu.sas
Descr_Poisson.sas
Descr_Geometr.sas
Descr_Hypergeom.sas
```

In den Wahrscheinlichkeiten können Parameter auftreten, die in der Regel unbekannt sind.

Die Parameter sind anhand der Beobachtungen (der Daten) zu bestimmen/zu schätzen!

$\longrightarrow$ Aufgabe der Statistik

Werkzeuge der empirischen Forschung

W. Kössler

Einleitung

Datenbehandlung
Syntax
Tastatur
Transformationen
Externes File
Input-Anweisung
SAS-Files
Zusamenfügen
Output-Anweisung
DO-Schleifen

Wkt.rechnung
Population
Wahrscheinlichkeit
Zufallsvariablen
Diskrete Zufallsvariablen
Stetige Zufallsvariablen
Normalverteilung
Erwartungswert
Varianz

Sei X stetig auf (a,b), wobei a, b unendlich sein können, $a \leq x_0 < x_1 \leq b$
$P(X = x_0) = 0, \quad P(x_0 < X < x_1) > 0.$

Die Funktion f heißt Dichtefunktion (von X) falls:

1. $f(x) \geq 0, \quad a < x < b.$
2. $\int_a^b f(x)\, dx = 1.$

Die stetige Zufallsvariable X wird also durch seine Dichtefunktion beschrieben.

$$P(c < X < d) = \int_c^d f(x)\, dx.$$

Die Dichtefunktion hängt i.A. von unbekannten Parametern ab, die geschätzt werden müssen.