Stichprobenverfahren

Einführung

Willi Mutschler
willi@mutschler.eu

Sommersemester 2017

Motivation

- 2017 ist Wahljahr (Bundestag und Landtag)
- Obwohl um 18 Uhr keine einzige Stimme ausgezählt ist, gibt es erste Prognosen, die erstaunlich genau sind
- Prognosen basieren üblicherweise auf Befragung von Wählern unmittelbar nach Abgabe ihrer Stimme (*Exit Polls*)
  - Wie schaffen wir es durch eine Befragung von 2000 Personen Aussagen über eine Bevölkerung von 80 Millionen Personen zu machen?
- Ziel der Veranstaltung ist es, Regeln zu finden für
  1. Strategie der Ziehung
  2. Auswertung der Antworten
- Anwendungen:
  - Marktforschung, sozial- und wirtschaftswissenschaftliche Erhebungen, medizinische Studien, Umweltforschung, ...

Gliederung

1. Erhebungsverfahren
2. Inklusionsindikator und Inklusionswahrscheinlichkeiten
3. Schätzfunktionen
4. Einfache Zufallsstichproben
5. Konfidenzintervalle
6. Schichtenverfahren
7. Klumpenverfahren
8. Gebundene Hochrechnung
9. Stichprobenregression
10. Modellbasierte Stichprobenverfahren
11. Capture-Recapture Auswahl, Ranked Set Sampling, Adaptive Sampling
12. ...

Zufallsgeneratoren (1)

- Voraussetzung:
  - Grundlegende Kenntnisse der Wahrscheinlichkeitsrechnung und der Mathematischen Statistik
- Die Wahrscheinlichkeitsrechnung ist aus der Beschäftigung von Mathematikern mit Glücksspielen in der zweiten Hälfte des 17. Jahrhunderts entstanden
- Für uns wichtig: Glücksspiele stellen spezifische Verwendungen von Zufallsgeneratoren dar
- Typische Beispiele: Werfen von Münzen und Ziehen von Kugeln aus Urnen
- Zufallsgeneratoren spielen somit in der Wahrscheinlichkeitsrechnung eine zentrale Rolle

Zufallsgeneratoren (2)

Charakterisierung von Zufallsgeneratoren

1. Ein Zufallsgenerator ist ein Verfahren, mit dem durch Aktivierung des Zufallsgenerators Sachverhalte erzeugt werden können
2. Die Beschreibung eines Zufallsgenerators besteht in der Beschreibung des Verfahrens zur Erzeugung von Sachverhalten (z.B. Urne mit Kugeln füllen, die sich nur durch die Beschriftung unterscheiden, mischen, blind ziehen, ...)
3. Zufallsgeneratoren können wiederholt (beliebig oft) angewendet werden
4. Mit dem Zufallsgenerator können *Sachverhalte unterschiedlichen Typs* (z.B. die Zahlen 1 bis 6 beim Würfeln) entstehen. Welcher Typ resultiert ist vor der Aktivierung unbestimmt.
5. Welcher Sachverhalt resultiert soll unabhängig von der vorherigen Verwendung des Zufallsgenerators sein (kein Gedächtnis)

Zufallsgeneratoren (3)

- Unsere *Definition von Wahrscheinlichkeit* als

$$\frac{\text{Zahl der günstigen Ereignisse}}{\text{Zahl der gleichmöglichen Ereignisse}}$$

  beruht auf der Vorstellung eines elementaren Zufallsgenerators
- *Idealer Würfel*, Urne mit Kugeln, die sich nur durch die Farbe, die Beschriftung unterscheiden

Datengenerierende Prozesse (1)

- Ursprünglich beziehen sich die Begriffsbildungen der Wahrscheinlichkeitsrechnung somit auf Zufallsgeneratoren
- In den Wirtschafts- und Sozialwissenschaften ist es allerdings üblich, die Begriffsbildungen der Wahrscheinlichkeitsrechnung auf soziale Prozesse anzuwenden
- D.h. es wird der spekulative Versuch unternommen, soziale Prozesse, durch die Sachverhalte unserer Erfahrungswelt entstehen, so zu deuten, als handle es sich um Realisierungen von Zufallsgeneratoren
- Beispiel: Renditen von Wertpapieren werden als Realisation eines Zufallsgenerators betrachtet, der normalverteilte, t-verteilte, Laplace verteilte,..., Zufallszahlen erzeugen kann

Datengenerierende Prozesse (2)

- Die stochastische Regressionsanalyse stellt ein wichtiges Instrument im Rahmen der *spekulativen Deutung sozialer Prozesse als Realisierungen von Zufallsgeneratoren* dar
- Ein Beispiel: die Einkommensfunktion

$$\mathit{Arbeitslohn}_i = \beta_0 + \beta_1 \cdot \mathit{Ausbildungsjahre}_i + u_i$$

- Überlegungen könnten dazu geführt haben, dass wir annehmen, es bestehe ein linearer Zusammenhang zwischen dem Arbeitslohn einer Person i und der Zahl der Ausbildungsjahre dieser Person i
- der durch eine Realisation u_i einer Zufallsvariable U_i additiv überlagert ist
- In diesem Beispiel wird also der Sachverhalt eines bestimmten Arbeitslohns einer Person i mit einer bestimmten Zahl an Ausbildungsjahren so interpretiert, als sei er durch einen Zufallsgenerator erzeugt worden

Datengenerierende Prozesse (3)

- Möglicherweise ist das Einkommen der Person *i* aber deshalb doppelt so hoch wie der durchschnittliche Lohn von Personen mit dieser Zahl an Ausbildungsjahren, weil *i* in dem Unternehmen seines Vaters angestellt ist
- Unter Ökonomen, Soziologen und Statistikern ist es üblich, derartige ausgedachte Zufallsgeneratoren als *datengenerierende Prozesse* zu bezeichnen
- Im Rahmen der Beschäftigung mit der Regressionsanalyse werden also Methoden betrachtet, mit denen man aus vorliegenden Daten einer bestimmten Anzahl von Personen Informationen über ausgedachte *datengenerierende Prozesse* gewinnen kann
- Solche ausgedachten *datengenerierenden Prozesse* werden auch als *Superpopulationsmodelle* bezeichnet

Gesamtheiten und Stichproben (1)

- Ein wichtiges Anwendungsfeld der Wahrscheinlichkeitsrechnung stellt die Stichprobentheorie dar
- Ausgangspunkt ist eine endliche Menge U von Einheiten
- An diesen Einheiten könnten die Ausprägungen des Merkmals Y gemessen werden
- Hätten wir für alle Einheiten von U das Merkmal Y gemessen, könnten wir mit den Methoden der deskriptiven Statistik den Informationsgehalt übersichtlich darstellen
- In der Stichprobentheorie beschäftigen wir uns mit dem Problem, dass uns die Ausprägungen des Merkmals Y nicht für alle Einheiten der Grundgesamtheit U; sondern für eine Teilmenge S vorliegen
- Die Teilmenge S mit $S \subset U$ wird als Stichprobe bezeichnet
- Wir interessieren uns für Aussagen über die Verteilung von Y in U; haben aber lediglich Angaben über Y in S vorliegen

Gesamtheiten und Stichproben (2)

- In der Stichprobentheorie untersuchen wir nun, was wir über die Verteilung von Y in U auf der Basis einer Stichprobe S sagen können
- Unmittelbar ersichtlich ist: über sich nicht in der Stichprobe befindende Einheiten kann auf Basis der Stichprobe nichts gesagt werden
- Aber auf der Basis von Stichproben, die durch ein bestimmtes Auswahlverfahren gewonnen wurden, können wir Hypothesen über die Verteilung von Y in U bilden oder die Plausibilität von Hypothesen über Y in U einschätzen
- Wahlbeispiel:
  - von allen (U) abgegebenen Stimmen wurde eine Stichprobe S gezogen und ausgezählt. Auf Basis des Auszählung der Stichprobe können Hypothesen über Y in U (Partei ... erreicht mehr als ... Prozent) eingeschätzt werden
- Grundlegende Voraussetzung der Anwendung der Wahrscheinlichkeitsrechnung im Rahmen der Stichprobentheorie ist die Verwendung eines Zufallsgenerators zur Auswahl der Einheiten aus U, die in die Stichprobe S gelangen

Gesamtheiten und Stichproben (3)

Zu beachten ist der folgende grundlegende Unterschied:

- Im Rahmen der Beschäftigung mit *datengenerierenden Prozessen* werden Sachverhalte (z.B. das Einkommen von Personen) als durch Zufallsgeneratoren erzeugt gedacht. Gewonnene Kenntnisse über ausgedachte *datengenerierende Prozesse* sollen helfen, über soziale Prozesse nachzudenken.
- In der Stichprobentheorie geht es nicht um das *Spekulieren* über das Zustandekommen sozialer Sachverhalte, sondern diese werden als gegeben vorausgesetzt. Zufallsgeneratoren dienen nur der Auswahl von Einheiten aus einer Grundgesamtheit U in eine Stichprobe S: Aufgrund der Stichprobe sollen dann Hypothesen über die Grundgesamtheit einschätzbar gemacht werden.
- Wir wollen im Folgenden unter einer Stichprobe *eine mit Hilfe eines Zufallsgenerators zufällig ausgewählte Teilmenge aus einer endlichen Grundgesamtheit* verstehen

Grundproblem der Stichprobentheorie (1)

Stichproben werden aus Grundgesamtheiten gezogen, um über Charakteristika der Grundgesamtheit etwas zu erfahren

- Im Folgenden wollen wir uns nur mit Zufallsstichproben beschäftigen:
  - D.h. wir wählen aus den N Elementen der Grundgesamtheit mit Hilfe eines Zufallsgenerators n Elemente aus
  - Diese n ausgewählten Elemente bilden eine Stichprobe S

Analog können wir auch die Stichproben betrachten:

- Aus der Menge aller möglichen Stichproben $\mathcal{S}$ wählen wir eine Stichprobe S aus
- Wenn wir eine bestimmte Stichprobe S gezogen haben, können wir die Werte der interessierenden Variable Y bei diesen n Einheiten der Stichprobe messen
- Leider: über die $N - n$ Einheiten, die nicht in der konkreten Stichprobe S sind, können wir auf Basis der Kenntnis der erhobenen n Einheiten nichts sagen

Grundproblem der Stichprobentheorie (2)

- Haben wir z.B. eine Grundgesamtheit mit 5 Frauen und 5 Männern und ziehen eine einfache Zufallsstichprobe vom Umfang $n = 4$; können wir nur den Anteil der Frauen in der Stichprobe ermitteln
- Über das Geschlecht der 6 nicht in die Stichprobe gelangten Personen können wir nichts sagen
- Wir können natürlich eine Stichprobe ziehen, die nur Männer oder nur Frauen enthält. Entsprechend würden wir dann zu ziemlich schlechten Vermutungen über den Anteil der Frauen in der Grundgesamtheit gelangen
- Warum dann die Beschäftigung mit Stichprobentheorie?

Grundproblem der Stichprobentheorie (3)

- Wir versuchen nicht direkt, auf Basis der konkreten vorliegenden Stichprobe etwas über die nicht erfassten Einheiten zu sagen
- Sondern wir gehen gedanklich von der Grundgesamtheit aus und betrachten alle möglichen Stichproben $\mathcal{S}$
- Einfaches Auszählen aller möglichen Stichproben ergibt dann die Wahrscheinlichkeitsverteilung (z.B. des Anteils von Frauen) einer Stichprobe $n = 4$
- Dieses Verfahren nennt man den *direkten Schluss*
- Auf diese Art können wir verschiedene Auswahl- und Schätzverfahren beurteilen
- Wir würden lieber solche Auswahl- (z.B. freie Zufallsauswahl) und Schätzverfahren (z.B. Mittelwert) wählen, die mit hoher Wahrscheinlichkeit zu Schätzwerten führen, die nahe bei dem interessierenden Grundgesamtheitsparameter liegen

Grundproblem der Stichprobentheorie (4)

- Tatsächlich kennen wir natürlich die Grundgesamtheit nicht
- D.h. wir können nur allgemeine Vorzüge und Nachteile bestimmter Verfahren beurteilen
- Auf Basis einer einzigen vorliegenden Stichprobe können wir nur versuchen, Hypothesen über die Grundgesamtheit einschätzbar zu machen
- Eine Stichprobe mit 4 Frauen ($p = 1$) würde uns (fälschlicherweise) der (in diesem Fall wahren) Hypothese $p = 0.5$ wenig Vertrauen entgegenbringen lassen
- Denn unter der Hypothese haben Stichproben mit $p = 1$ eine geringe Wahrscheinlichkeit gezogen zu werden

Populationswerte

Ziel der Stichprobentheorie sind Aussagen über Populationswerte

- Populationswerte sind Maßzahlen (Anteile, Mittelwerte, etc.) der endlichen Grundgesamtheit
- Stichproben sollen ausgewählte Teilmengen nur dann heißen, wenn die Auswahl zufällig war $\rightarrow$ der Prozess der Ziehung ist genau definiert (Zufallsgenerator!)
- Zufällig ist, welche Grundgesamtheitselemente in die Stichprobe gelangen
- Damit sind auch die für die Stichprobe berechneten Maßzahlen zufällig
- Schätzer sind Maßzahlen, die auf Basis der Stichprobe berechnet werden, die Maßzahlen der Grundgesamtheit aber möglichst gut *treffen* sollen
- Gewünschte Eigenschaften sind
  - Erwartungstreue
  - Geringe Varianz
  - Möglicherweise trade-off zwischen Erwartungstreue und Varianz

Design einer Zufallsstichprobe

Beispiel für Stichprobendesigns:

- Population von 5 Merkmalsträgern (A,B,C,D,E)
- Ziel ist es, 2 Einheiten in Form einer Stichprobe zu ziehen. Wir haben folgende Möglichkeiten:

$$s_1 = (A, B), \quad s_2 = (A, C), \quad s_3 = (A, D), \quad s_4 = (A, E), \quad s_5 = (B, C)$$
$$s_6 = (B, D), \quad s_7 = (B, E), \quad s_8 = (C, E), \quad s_9 = (C, E), \quad s_{10} = (D, E)$$

- Zuordnung von Wahrscheinlichkeiten:
  1. Alle Stichproben haben die gleiche Wahrscheinlichkeit von 1/10
     $\rightarrow$ **einfache Zufallsstichprobe**
  2. In Stichprobe soll ein Konsonant und Vokal vorkommen, also nur s_1, s_2, s_3, s_7, s_9 und s_{10} zulässig, diese bekommen jeweils die Wahrscheinlichkeit 1/6
     $\rightarrow$ **geschichtete Stichprobe**
  3. Element A soll besonders wichtig sein: alle Stichproben, die A enthalten, erhalten eine größere Wahrscheinlichkeit, zum Beispiel:
     $P(S_1) = P(s_2) = P(s_3) = P(s_4) = \frac{2}{14}$ und
     $P(s_5) = P(s_6) = P(s_7) = P(s_8) = P(s_9) = P(s_{10}) = \frac{1}{14}$
     $\rightarrow$ **probabilities proportional to size**

Materialien zur Vorlesung

- Infos und Materialien: `https://mutschler.eu/teaching`
- Passwort: dortmund
- Termine:
  - Vorlesung: Donnerstags, 12.30-14.00
  - Übung: Donnerstags, 14.15-15.45
  - Vorlesung und Übung werden nicht streng getrennt, wir haben also zwei Termine
  - Bitte bringen Sie ihren Laptop mit vorinstallierten R mit
- Achtung: Vorlesung fällt aus am 25.05. (Feiertag), 15.06. (Feiertag), 06.07, 13.07 und 20.07.
- Prüfung: Klausur, Termin nach Absprache

Literatur

- Behr (2015): Theory of Sample Survey with R
- Cochran (1972): Stichprobenverfahren
- Kauermann und Küchenhoff (2011): Stichproben
- Särndal, Swensson und Wretman (1992): Model Assisted Survey Sampling
- Thompson (1997): Theory of Survey Samples
- Thompson (2002): Sampling