Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

# Stichprobenziehung (Sampling) Grundlagen

## CAS practical machine learning

▸ Violeta Vogel, TI BFH

# Agenda

- Was ist eine Stichprobe?
- Representativität
- Grösse der Stichprobe
- Arten der Stichproben

# Was ist Stichprobe?

- Eine Stichprobe ist ein vollständiges verkleinertes Spiegelbild der Grundgesamtheit, die damit auch alle (wesentlichen) Eigenschaften der Grundgesamtheit korrekt wiedergibt

- Mit einer Probeziehung wird ein repräsentativer Teil einer Population ausgewählt, um Rückschlüsse auf die Gesamtheit zu ziehen

- Repräsentativ: Aus einer Stichprobe lassen sich zutreffende Rückschlüsse auf eine Grundgesamtheit ziehen
  - eine Stichprobe dann repräsentativ, wenn alle Beobachtungen der Grundgesamtheit die gleiche Chance besessen haben, Teil dieser Stichprobe zu werden

# Stichprobengrösse

- Wissenschaft n > 30
- Marktforschung n > 50
- Bevölkerungsrepräsentative Studien n > 1000
- Maschinelles Lernen: Als Faustregel gilt, dass Sie mindestens zehnmal so viele Datenpunkte/Beobachtungen benötigen, wie Ihr Datensatz Features enthält

# Art des Ziehens

- Ziehen *mit* Zurücklegen:  Ein und dasselbe Element kann *mehrfach* in die Stichprobe gelangen.
- Ziehen *ohne* Zurücklegen: Ein und dasselbe Element *nur einmal* in die Stichprobe gelangen.

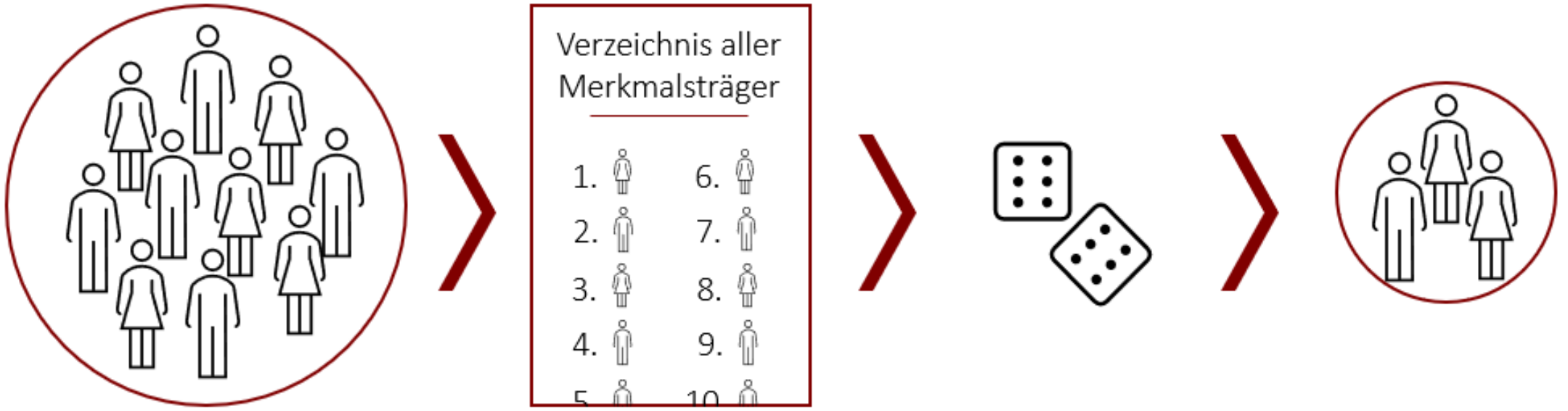
# Arten der Zufallsstichproben

- Einfache Stichprobe
- Geschichtete Stichprobe
- Klumpenstichprobe

# Einfache Stichprobe (Auch Zufallsstichprobe)

- Rein zufällig Features (bzw. Merkmalsträger) gezogen: Jede Beobachtung hat die gleiche Chance, ausgewählt zu werden (z.B. Generator der Zufallszahlen)
- Vorteil: Ist die Zufallsstichprobe groß genug, so werden automatisch auch alle Beobachtungen der Grundgesamtheit in einem ähnlichen Verhältnis in der Stichprobe auftauchen
- Vorrausetzung: Jede zur Grundgesamtheit gehörende Beobachtung dieselbe Auswahlwahrscheinlichkeit aufweist. D.h. die gleiche Chance hat, in die Stichprobe mit aufgenommen zu werden

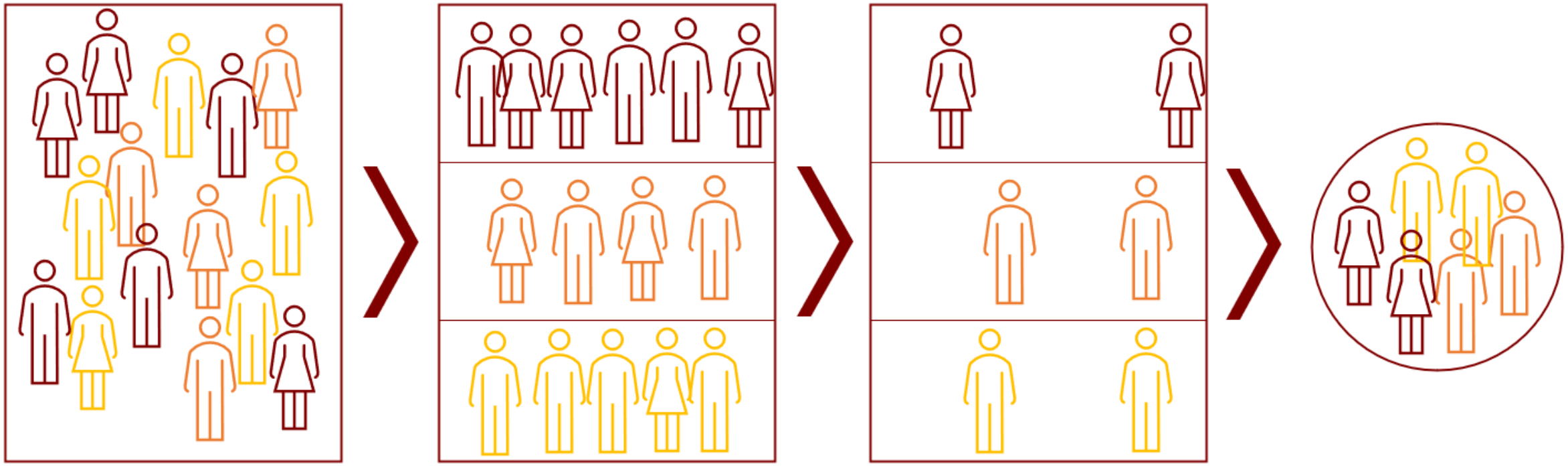
# Einfache Stichprobe



# Geschichtete Stichprobeziehung (Stratifizierung)

- Ziel der geschichteten Stichprobenziehung ist die Grundgesamtheit so genau wie möglich darzustellen.
- Anforderungen an die Schichten: Sie sollten sich gegenseitig ausschließen und gemeinsam erschöpfend sein
- Vorgehen:
  1. die relevanten Variablen identifizieren, die die Grundgesamtheit definieren. ( Alter, Geschlecht, Einkommen oder Standort)
  2. die Grundgesamtheit auf der Grundlage dieser Variablen in Untergruppen (Schichten) einteilen
  3. Zufallsstichprobe aus jeder Schicht auswählen:
     1. Proportional: entsprechend dem Anteil der Schicht an der Grundgesamtheit
     2. Disproportional: etwa Gleich grosse Stichproben aus jeder Schicht

# Stratifizierung



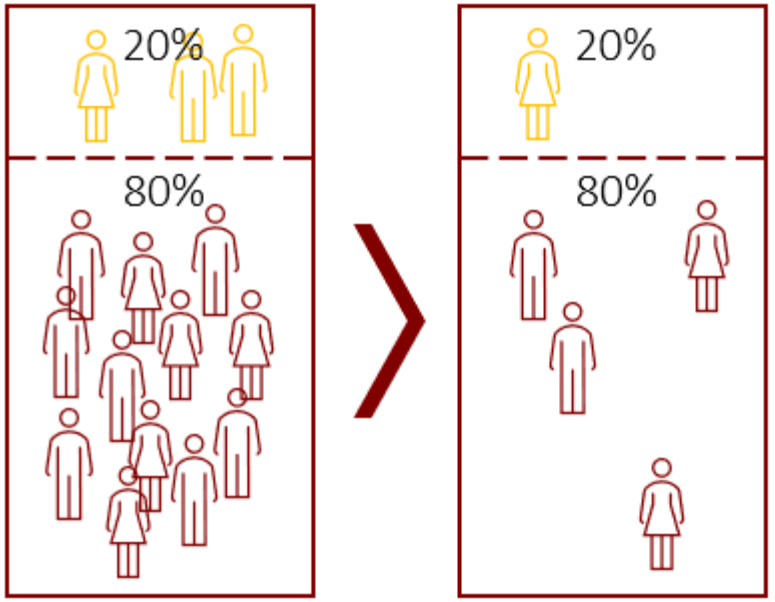
Grundgesamtheit

Einteilung in Schichten

Ziehung einer Zufallsstich-
probe aus jeder Schicht

Schichten-
stichprobe

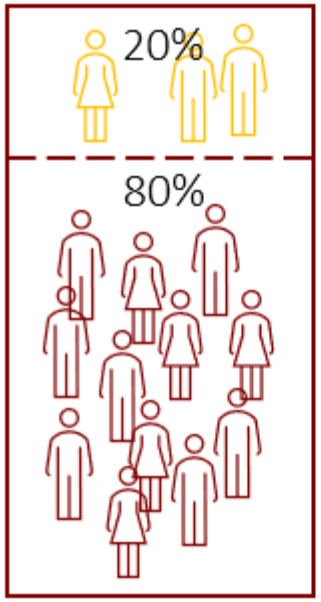
# Proportionale versus disproportionale Stratifizierung



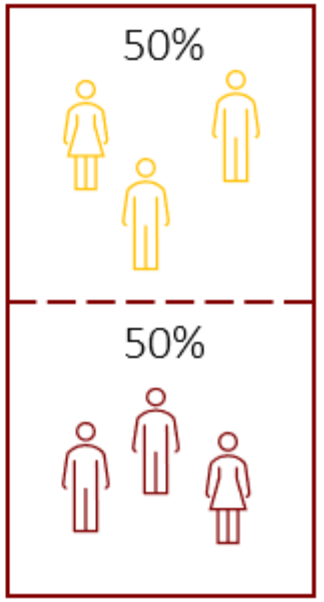

Grundgesamtheit




proportionale Stichprobe




Grundgesamtheit




disproportionale Stichprobe

# Proportionale Stratifizierung

- Die Schichten werden originalgetreu gewichtet
- Vorteile:
  - Repräsentativität: Die Grundgesamtheit ist in der Stichprobe gut repräsentiert
  - Präzise Schätzungen: Es können genauere Schätzungen für die Grundgesamtheit erzielt werden
- Nachteile:
  - Aufwand: Die Identifizierung und Abgrenzung der Schichten kann aufwendig sein
  - Komplexität: Es kann komplexer sein als bei einfachen Stichprobenverfahren

# Disproportionale Stratifizierung

- Die Schichten werden unterschiedlich stark gewichtet
- Verwendung, wenn:
  - bestimmte Schichten von besonderem Interesse sind oder
  - die Kosten der Datenerhebung in den verschiedenen Schichten variieren
- Ziel: Eine ausreichende Anzahl von Elementen aus jeder Schicht zu erhalten, um aussagekräftige Ergebnisse zu erzielen, insbesondere bei Schichten, die in der Grundgesamtheit nur einen kleinen Anteil ausmachen
- Gewichtung: Faktor, der die ursprüngliche Verteilung in der Grundgesamtheit berücksichtigt
-

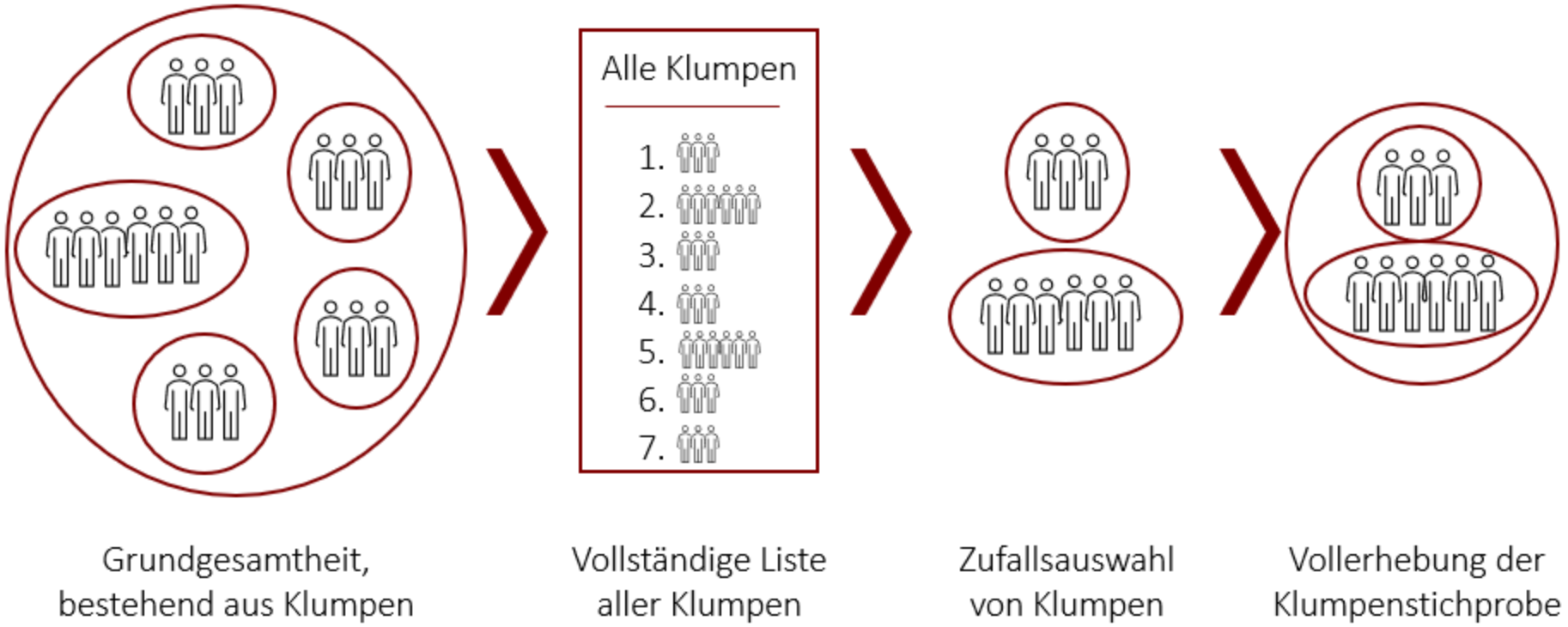
# Disproportionale Stratifizierung

- Vorteile:
  - Genauigkeit: Ermöglicht eine genauere Analyse von Schichten, die in der Grundgesamtheit unterrepräsentiert sind
  - Effizientere Datenerhebung: Kann kostengünstiger sein, wenn die Datenerhebung in einigen Schichten teurer ist
  - Berücksichtigung von Heterogenität: Erlaubt eine detailliertere Analyse der Unterschiede zwischen den Schichten
- Nachteile:
  - Verzerrte Ergebnisse**:** Führt zu verzerrten Ergebnissen, wenn die Ergebnisse nicht richtig gewichtet werden
  - Komplexe Analys: Erfordert zusätzliche Schritte bei der Datenanalyse -> Gewichtung

# Klumpenstichprobe

- Grundgesamtheit zunächst hinsichtlich eines Merkmals in natürliche Klumpen eingeteilt (Klassen in einer Schule)
- Klumpen untereinander homogen: Jeder Klumpen ist ein verkleinertes Abbild der Population und sich daher alle Klumpen stark ähneln
- Beispiele: Schulen, Wahlbezirke

# Klumpenstichprobe



# Klumpen versus Stratifizierung (Schichten)

| Klumpenstichprobe | Schichtenstichprobe |
|---|---|
| Jedes Element der Grundgesamtheit gehört zu genau einem Klumpen | Jedes Element der Grundgesamtheit gehört zu genau einer Schicht |
| In der Regel entsprechen die Klumpen „natürlichen“ Gruppierungen | In der Regel entsprechen die Schichten willkürlich gewählten Merkmalen |
| Es wird eine einfache Zufallsstichprobe aus der Menge der Klumpen gezogen | Alle Schichten werden berücksichtigt |
| Innerhalb eines ausgewählten Klumpens gelangen alle Elemente in die Stichprobe | Aus jeder Schicht wird jeweils eine Zufallsstichprobe gezogen |