

Lösungen einiger Beispiele zur Vorlesung Statistik 2:

Kap. 24: Stichprobentheorie

Zu Kap. 24.3. Auswahlverfahren I: Zufallsstichprobe

Im Rahmen einer Mobilitätsuntersuchung in Dresden sollen aus den 210000 Haushalten eine Zufallsstichprobe vom Umfang 1000 befragt werden. Als Zufallszahlen stehen zur Verfügung:

- (i) Eine genügend lange Folge von $(0, 9)$ -gleichverteilten ganzzahligen Zufallszahlen. Der Beginn könnte z.B. so aussehen: 23407623580134857...
- (ii) Stetige, $(0, 1)$ gleichverteilte "Pseudo"-Zufallszahlen aus der `random`-Funktion des Computers.

Wie geht man jeweils konkret vor?

Lösung Zunächst nummeriert man die Haushalte von 1 bis 210000, z.B. in alphabetischer oder in irgendeiner anderen beliebigen Reihenfolge. (man könnte auch von 0 bis 209999 nummerieren)

- (i) Einfachste Methode: Man teilt die Zufallszahlen in Sechser-Gruppen (da die höchste Nummer sechsstellig ist)
 - Falls die entsprechende Zahl zwischen 1 und 210000 liegt und das erste Mal vorkommt, nimmt man sie als Stichprobe
 - Ansonsten verwirft man die Zahl
 - Dieses Spielchen treibt man, bis man die 1000 Haushalte "beieinander" hat.

Besser, da mit weniger "Verbrauch" an Zufallszahlen verbunden, ist folgendes Schema:

- Gehe in der Zufallszahlen-Reihe bis zur nächsten Zahl, die zwischen 0 und 2 liegt,
- Bilde eine Sechser-Gruppe, beginnend mit dieser Zahl und verfare wie oben

Die ersten beiden Nummern, die mit dieser Methode gezogen werden, sind 076235 \Rightarrow 76235 und 013485 \Rightarrow 13485.

- (ii) Man multipliziert die Realisierungen x_i des Zufallsgenerators $X \sim G(0, 1)$ mit 210000 und rundet auf die nächste ganze Zahl auf, die gleich der ausgewählten Haushalts-Nummer entspricht

Zu Kap. 24.3. II: Systematische Stichprobe

Bei Fertigung von Motorventilen wird aus der laufenden Produktion jedes 50-te Teil auf Fertigungstoleranzen kontrolliert. Die Maschine hat jedoch nach jedem 100-ten Ventil eine automatische Nachjustierung. Erhält man eine repräsentative Stichprobe?

Antwort: Nein. Prüft man z.B. je das 1. und 51. Teil nach der Nachjustierung, erhält man systematisch bessere Ergebnisse als wenn man das je 50. und 100. Teil nach der Nachjustierung prüft. Korrekt wäre es, bei gleichem Auswahlsatz z.B. nach jedem 5000 ten Teil 100 Teile hintereinander auszuwählen.

Per Stichprobe (Auswahlsatz 1%) soll die Altersstruktur von Motorradbesitzern ermittelt werden. Da die meisten gar kein Motorrad besitzen, wäre eine Stichprobe unter allen erwachsenen Personen (Einwohnermeldeamt) uneffektiv. Effektiver ist es da schon, aus der Motorrad-Zulassungsstatistik jedes Hunderste Kennzeichen auszuwählen und das Alter des entsprechenden Besitzers zu erfassen.

- Was ist hier "faul?"
- Wie könnte man es beheben?

Antwort: Es gibt durchaus Motorradfans, die zwei oder mehrere zugelassene Motorräder besitzen! Diese Besitzer werden bei obigen Verfahren systematisch bevorzugt. Wäre die Zahl der Motorräder unabhängig vom Alter, so bewirkt dies nur eine minimale Vergrößerung des Stichprobenfehlers. Wegen des nötigen Kapitalbedarfs sind die "Mehrfachbesitzer" unter den Motorradliebhabern jedoch eher gesetzteren Alters und die Untersuchung erhalte einen systematischen Fehler in Richtung zu hohen Alters!

Abhilfe: Besitzer von n Motorrädern werden mit den Faktor $1/n$ gewichtet.

Zu Kap. 24.3. Auswahlverfahren IV: Quotenverfahren (geschichtete Stichprobe)

Wie kann man konkret eine Stichprobe vom Umfang $n = 1000$ ziehen, bei der folgende Quoten erfüllt sind:

- dass 54% der Grundgesamtheit (Einwohner) Frauen sind,
- und dass 20% der Frauen und Männer zwischen 0 und 20 Jahre, 40% der Frauen und 50% der Männer zwischen 20 und 50 Jahre alt sind.

Lösung:

Allgemein kann man nach folgendem Schema vorgehen:

Schritt 1: Aufspalten des Stichprobenumfangs nach Quoten:

$$n_i = nq_i$$

Schritt 2: Ermittlung von Zufallsstichproben vom Umfang n_i in den Teilmengen i

Schritt 3: Die Quoten-Stichprobe ist die Vereinigung dieser Zufallsstichproben.

Hier hat man insgesamt 6 Quoten:

Teilmenge	q_i	n_i
♀, 0 – 19 Jahre	0.108	108
♂, 0 – 19 Jahre	0.092	92
♀, 20 – 49 Jahre	0.216	216
♂, 20 – 49 Jahre	0.230	230
♀, ≥ 50 Jahre	0.216	216
♂, ≥ 50 Jahre	0.138	138

Aus jeder dieser sechs Grundgesamtheiten wählt man mit einem der bisherigen Verfahren n_i Elemente aus.

Quotenverfahren: Meinungsumfrage zu Radwegebau

Die Kommunalpolitiker einer großen Stadt wollen mittels einer Stichprobe vom Umfang 1000 die Meinung der Bürger bezüglich der Frage: "Halten Sie Sanierungen und Neubauten von Radwegen für dringend notwendig" wissen. Zur Planung der Stichprobe steht eine Meinungsumfrage vom vergangenen Jahr zur Verfügung. Außerdem ist bei der neben dieser Frage u. a. auch nach dem Autobesitz gefragt wurde. Das Ergebnis lautete wie folgt (Zahlen sind fiktiv!):

Autobesitz	mindestens eines	kein Auto
Quote q_i	40%	60%
Für mehr Radwege	20%	93.6%

- (a) (\Rightarrow Deskriptive Statistik). Wie groß war der Ja-Anteil in dieser letzten Umfrage?

Antwort: Ja-Anteil $\theta_0 = \sum_{i=1}^2 q_i \theta_i = 0.4 * 0.2 + 0.6 * 0.936 = \underline{\underline{64.15\%}}$

- (b) Wie lautet die Verteilungsfunktion des Ja-Anteils und wie groß ist der erwartete $1-\sigma$ Fehler, wenn man die neue Stichprobe mit dem Zufallsverfahren durchführt? Nehmen Sie dabei an, dass sich der wahre Anteil θ_0 der Radwege-Befürworter nicht wesentlich geändert hat und gleich dem Anteil der letzten Stichprobe angenommen werden kann. Benutzen Sie die Binomialverteilung für die einzelnen, unabhängigen Ja-Nein-Entscheidungen, den Zentralen Grenzwertsatz und die Additionsregel für die Varianzen unabhängiger Zufallsgrößen.

Lösung: Die Zahl X der Ja-Antworten ist durch

$$X = \sum_{i=1}^{1000} X_i \sim B(1000, \theta_0)$$

gegeben, wobei jedes $X_i \sim B(1, \theta_0)$ -binomialverteilt ist ($X_i = 1$ entspricht einer Ja-Entscheidung, $X_i = 0$ einer Nein-Entscheidung). Wenn sich die Ja-Anteile nicht wesentlich ändern (das kann man natürlich nur *a posteriori* überprüfen), dann ist (vgl. (a)) $\theta_0 = 0.6415$. Aufgrund der Zufallsauswahl und der Bedingung $n \ll N$ sind die X_i unabhängig. Da außerdem $n > 30$, ist der Zentrale Grenzwertsatz erfüllt und es gilt mit sehr guter Genauigkeit:

$$X \sim N(\mu, \sigma^2)$$


mit

$$\begin{aligned}\mu &= \sum_{i=1}^{1000} \mu_i = \sum_{i=1}^{1000} \theta_0 = 641.5, \\ \sigma^2 &= \sum_{i=1}^{1000} \sigma_i^2 = \sum_{i=1}^{1000} \theta_0(1 - \theta_0) = 230\end{aligned}$$

Damit ist der 1σ -Fehler gleich $\sqrt{\sigma^2} = 15.2$ Ja-Stimmen ($\approx 1.5\%$).

(c) *Wie könnte man bei der Quotenauswahl konkret vorgehen?*

Antwort: Auswahl von $0.4 \cdot 1000 = 400$ Autobesitzern und von $0.6 \cdot 1000 = 600$ Personen, die kein Auto besitzen, z.B. indem man erst mit Zufalls- oder Klumpenwahl beliebige Personen auswählt, bis die jeweiligen Quoten voll sind.

(d)  *Wie lautet die Verteilungsfunktion des Ja-Anteils und wie groß ist der erwartete 1σ Fehler, wenn man die neue Stichprobe mit der Quotenstichprobe unter Zuhilfenahme des Quotenmerkmals "Autobesitz" durchführt? Diskutieren Sie den Unterschied! Gehen Sie davon aus, dass sich die Quoten selbst (d.h. der Anteil der Autobesitzer) nur unwesentlich geändert haben.*

Lösung: Beim Quotenverfahren setzt sich die Zahl der Ja-Stimmen aus zwei *unabhängigen* Quotenanteilen zusammen:

$$X = X_{\text{Auto}} + X_{\text{keinAuto}},$$

die jede für sich den Zentralen Grenzwertsatz erfüllen:

$$\begin{aligned}X_{\text{Auto}} &\sim N(\mu_{\text{Auto}}, \sigma_{\text{Auto}}^2) = N(400 \cdot 0.2, 400 \cdot 0.2 \cdot 0.8) = N(80, 64), \\ X_{\text{keinAuto}} &\sim N(\mu_{\text{keinAuto}}, \sigma_{\text{keinAuto}}^2) \\ &= N(600 \cdot 0.936, 600 \cdot 0.936 \cdot (1 - 0.936)) = N(561.5, 36)\end{aligned}$$

Nun ist eine Summe von unabhängigen normalverteilten Zufallsvariablen wieder normalverteilt:

$$X \sim N(\mu, \sigma^2)$$

mit

$$\begin{aligned}\mu &= 80 + 561.5 = \underline{\underline{641.5}}, \\ \sigma^2 &= 64 + 36 = \underline{\underline{100}}\end{aligned}$$

Damit ist der 1σ -Fehler $\sqrt{\sigma^2}$ gegeben durch $\sqrt{100} = 10$ Ja-Stimmen (d.h. 1%) statt 15.2 Ja-Stimmen (1.52%) beim Zufallsverfahren!

Das Quotenverfahren setzt allerdings Konstanz des Quotenmerkmals voraus, hier also, dass sich der Anteil der Autobesitzer weniger stark ändert als die Meinung über neue Radwege. Das muss hier überprüft werden! Man sollte sich also doch lieber auf weitgehend konstante Quoten wie das Geschlecht beschränken.

Die Anteile θ_i und θ können sich hingegen durchaus deutlich von der Vor-Stichprobe unterscheiden. Ggf. muss man *a posteriori* eine neue Fehlerabschätzung vornehmen.

Weiteres Beispiel einer geschichteten Stichprobe: Haushalte

Hier ist das Quotenmerkmal die Haushaltsgröße (Zahl k der Personen im Haushalt) und das zu untersuchende Merkmal X z.B. die Zahl der Kfz pro Haushalt.

zu 24.3 V: Zufallsauswahl mit Korrekturfaktoren

Aufgabe mit statistischer Einheit = "Haushalt", Merkmal X_i =Zahl der Fahrzeuge im Haushalt, Quotenmerkmal Q =Haushaltsgröße (Ausprägungen q_k =Zahl der Personen im Haushalt) Verzerrung speziell bei Ziehungsgrundlage=Personenregister: Dann sind große Familiengrößen systematisch überrepräsentiert!