

Schätzung des Anteils kariesfreier Kinder mittels Clusterstichproben aus einer regionalen Grundgesamtheit von Kindergärten (mit Beispielrechnung kiga50agg.txt)

Für die Präzision der Schätzung erwarten wir etwa $\pm 3\%$ bei einem Stichprobenumfang von 50 Kindergärten (siehe "Durchführung"). Zur Berechnung mit dem Programm **WinPepi** werden die Originaldaten aus kiga50v170.txt in einer für Kindergärten aggregierten Datei kiga50agg.txt zusammengefasst. Zur Nomenklatur siehe "Durchführung".

Eine einfache Möglichkeit zur Schätzung des Anteils kariesfreier Kinder aus zahnärztlichen Daten, welche in Kindergärten erhoben wurden (Clusterstichproben), bietet das Programm **WinPepi, Modul Describe: J2 - cluster sample: different-sized clusters**. Hierfür müssen die Daten als Textdatei (*.txt) oder in einer Tabellenkalkulation (z.B. Excel, siehe Abb.) folgendermaßen angeordnet sein: Jede Zeile repräsentiert einen Kindergarten und jede Spalte eine Variable. In Kindergarten Nr. 2 wurden beispielsweise 59 Kinder untersucht, davon waren 38 kariesfrei ($\sum ng$). Solche aggregierten Daten werden häufig auch von kommerziellen Programmen (z.B. Oktoware) in Textform bereitgestellt.

kiganr	anzahl	$\sum ng$
2	59	38
7	61	30
9	52	35
12	50	35
13	58	29
19	22	15
20	24	16
21	57	36
25	29	18
29	47	29

Estimates prevalence (or any other proportion) from observations in a cluster or stratified sample or in pooled samples. The clusters may be clusters of subjects or of observations made made on each subject. The presence of the attribute under study is called a "hit". A pooled sample contains material from a number of individuals, which is then tested for (e.g.) a disease agent.

SELECT...

- ☐ J1. Cluster sample: equal-sized clusters
- ☒ J2. Cluster sample: different-sized clusters
- ☐ J3. Stratified sample
- ☐ J4. Pooled samples

Enter data for each cluster (left-hand box) OR for groups of clusters (right-hand box). Press <Enter> or <Space> after each entry.

(Optional:) Size of population: 7978

Display of results: ☒ per 100 ☐ proportions

Cluster size	No. of hits	Cluster size	No. of hits	Frequency
1	59	38		
2	61	30		
3	52	35		
4	50	35		
5	58	29		
6	22	15		

If frequency = 1, just press <Enter> or <Space> again. Press <Esc> to delete a line. Pasting data: press F2 for help.

Use PgUp, PgDn, or Arrow keys

Weitere notwendige Angaben hierfür beschränken sich auf:

- Größe der Gesamtpopulation (in WinPepi: **Size of population**) der 3 - 5 Jährigen, entspricht der Summe der gemeldeten Kinder je Altersgruppe in allen Kindergärten der Region (hier 1710 Dreijährige, 3077 Vierjährige, 3191 Fünfjährige, zusammen 7978). Wenn nur die Gesamtzahl der 3-6 Jährigen Kindergartenkinder in der Region verfügbar ist, so wird hiervon 1/6 abgezogen und man erhält näherungsweise die Anzahl der 3-5 Jährigen. Falls die altersspezifischen Zahlen nicht verfügbar sind sondern nur die Gesamtzahl der Kinder, müssen sie aus den Anteilen der Stichprobe geschätzt werden (Anteil 3 Jähriger in der Stichprobe x Gesamtzahl = Zahl 3 Jähriger in der Population).
- untersuchte Kinder pro Kindergarten (in WinPepi: **cluster size**) und
- davon Zahl der Kinder mit naturgesunden Gebissen ($\sum ng$ = **No. of hits**)

Beispiel für die Altersgruppe der 3 - 5 Jährigen: Laden Sie die Datei **kiga50agg.txt** herunter. Diese Textdatei zeigt die Untersuchungsergebnisse aus einer Stichprobe von 50 zufällig ausgewählten Kindergärten. Sie beinhaltet drei Spalten, die Nummer der Kindergärten „kiganr“, die Zahl der jeweils dort untersuchten Kinder „anzahl“ und die Zahl derjenigen mit naturgesunden Zähnen „Σng“. Eine nähere Beschreibung der Studienpopulation für diese Clusterstichprobe findet man im Abschnitt „Planung“. Zur Eingabe der Daten in WinPepi gibt es zwei Möglichkeiten:

1. Öffnen der Datei in einem Editor und Ausdrucken der Werte. Jedem Kindergarten (**kiganr**) sind die Zahl der untersuchten Kinder (**anzahl = cluster size**) und die davon kariesfreien (**Σng = No. of hits**) zugeordnet. Eingabe der Daten per Hand erfolgt in das linke Feld des WinPepi - Moduls. Dieses Vorgehen ist dann relevant, wenn aggregierte Ergebnisse auf Kindergartenenebene bereits existieren, aber nicht zum Kopieren in WinPepi geeignet sind.
2. Öffnen der Datei in einem Editor und Kopieren des Inhaltes in eine beliebige Tabellenkalkulation. Gleichzeitiges Öffnen des WinPepi-Moduls Describe: J2 - cluster sample: different-sized clusters. Kopieren der Daten (ohne Überschrift) aus den zwei Spalten „anzahl“ und „Σng“ der Tabellenkalkulation in das linke Feld von WinPepi (siehe Abb. oben). Dort muss der Cursor vorher in der ersten Zeile links stehen. Die Struktur der Daten muss nach dem Kopieren erhalten bleiben und es müssen 50 Datensätze erkennbar sein.

Die Größe der Gesamtpopulation der 3 - 5 Jährigen, die hier 7978 Kinder beträgt, wird bei „**Size of Population**“ eingetragen.

Starten mit „**Run**“. Das Ergebnis für die Methode nach Cochran lautet:

```
Prevalence = 62.1 per 100.  
  
By Cochran's procedure:  
  
90% C.I. = 59.8 to 64.4 per 100.  
95% C.I. = 59.4 to 64.9 per 100.  
99% C.I. = 58.4 to 65.8 per 100.  
  
S.E. = 1.4 per 100.  
Design effect = 2.59
```

und im unteren Teil der Ausgabe

```
Rate of homogeneity
(intraclass correlation coefficient) = 0.036

Finite population correction = 0.710

Mean cluster size          = 46.240
Adjusted mean cluster size = 46.138
```

Ergebnis: Anteil kariesfreier Kinder (3 - 5) in der Stichprobe (**Prevalence**): **p** = 62,1%

Das 95%-Konfidenzintervall beträgt unter Berücksichtigung der Clusterstruktur und der

Endlichkeitskorrektur: [59,4% ; 64,9%]

Standardfehler: 0,014

Design effect (variance inflation factor): 2,59

ICC: 0,036

und einige weitere Angaben. Man erkennt, dass die Endlichkeitskorrektur von 0.71 durchaus zu berücksichtigen ist.

Der aus der Clusterstichprobe mit 50 Kindergärten geschätzte Anteil kariesfreier Kinder im Alter von 3 bis 5 Jahren in der Studienpopulation (Grundgesamtheit) von 170 Kindergärten beträgt 62,1%. Der wahre Anteil (62,4%, den wir aber nicht kennen!) in dieser Grundgesamtheit liegt mit 95%-iger Wahrscheinlichkeit im Intervall zwischen 59,4% und 64,9%. Die Genauigkeit der Schätzung kann angegeben werden mit $\pm 2,75\%$ Punkten und erfüllt damit die Anforderung an die Genauigkeit von $\pm 3\%$.

Für eine Aussage zum Anteil kariesfreier Kinder (3 - 5) in der Bevölkerung fehlt hier noch eine **Altersadjustierung** (siehe unten), denn die Alterszusammensetzung in den Kindergärten entspricht nicht genau der in der Bevölkerung.

Wichtige Anmerkung:

Ohne Berücksichtigung der Clusterstruktur der Daten aber mit Endlichkeitskorrektur erhält man ein Konfidenzintervall von 60,43% bis 63,76% , d.h., eine Genauigkeit von $\pm 1,66\%$ Punkten. Das Konfidenzintervall nach dieser Rechnung ist schmaler und die Genauigkeit scheinbar größer. Leider ist diese Rechnung und damit auch das Konfidenzintervall falsch, da die Clusterstruktur der Daten nicht berücksichtigt wurde.

Berechnung mit STATA

Hierzu verwendet man die nichtaggregierten Daten kiga50v170.txt , die man mit *import delimited "Pfadeigabe"* importiert. Danach bildet man zwei neue Variable "fpc" und "pw".

Endlichkeitskorrektur $fpc = 2312 / 7978$ und sampling weight $pw = 7978 / 2312$. Mit *svyset* wird zuerst das Surveydesign angegeben . Mit *svy linearized : proportion ng* anschließend die Berechnung gestartet. Ausführliche Beschreibungen zu den Berechnungen findet man im STATA Survey Data Reference Manual unter www.stata.com .

```
. svyset kiganr, fpc(fpc) weight(pw) vce(linearized) singleunit(missing)
```

```
      pweight: <none>
      VCE: linearized
Single unit: missing
  Strata 1: <one>
      SU 1: kiganr
      FPC 1: fpc
  Weight 1: pw
```

```
. svy linearized : proportion ng
(running proportion on estimation sample)
```

Survey: Proportion estimation

```
Number of strata =      1      Number of obs   =      2,312
Number of PSUs   =      50      Population size = 7,977.9998
                                   Design df      =          49
```

		Linearized		Logit	
		Proportion	Std. Err.	[95% Conf. Interval]	
ng	0	.3788927	.0136918	.3517927	.4067704
	1	.6211073	.0136918	.5932296	.6482073

Den DEFF erhält man direkt anschließend mit dem Kommando:

```
. estat effects, deff
```

		Linearized		
		Proportion	Std. Err.	DEFF
ng	0	.3788927	.0136918	2.59211
	1	.6211073	.0136918	2.59211

Ergebnis: Der Anteil kariesfreier Kinder (3 - 5) in der Stichprobe beträgt 62,1% mit einem 95%-Konfidenzintervall unter Berücksichtigung der Clusterstruktur und der Endlichkeitskorrektur: C.I. = [59,3% ; 64,8%], einem Standardfehler von 0,0137 und einem Design ef-

fect (DEFF) von 2,59 in guter Übereinstimmung mit den Werten von WinPepi.

Berechnung per Hand

Hier sollte der **Quotienten-Cluster-Schätzer** (QC - Schätzer) für die Varianz bevorzugt werden, den man leicht mit einer Tabellenkalkulation berechnen kann. Nähere Angaben findet man in der Rubrik "Statistik im ÖGD". Ausgehend von den aggregierten Daten für die Kindergärten mit n_i = Anzahl der Kinder im i-ten Kiga, davon $y_{i,T}(=\sum ng)$ = Zahl der kariesfreien Kinder, berechnet man je Kiga: $n_i \cdot \bar{y}$ und $(y_{i,T} - n_i \cdot \bar{y})^2$.

Tab. Ausschnitt aus der Tabelle

kiganr	n_i	$y_{i,T} (\sum ng)$	$n_i \cdot \bar{y}$	$(y_{i,T} - n_i \cdot \bar{y})^2$
2	59	38	36,64533	1,83513
7	61	30	37,88754	62,21334
9	52	35	32,29758	7,30309
12	50	35	31,05536	15,56016
13	58	29	36,02422	49,33969
19	22	15	13,66436	1,78393
20	24	16	14,90657	1,19558
21	57	36	35,40311	0,35627
25	29	18	18,01211	0,00015
.	.	.	.	.
.	.	.	.	.

Unter Verwendung der Formeln nach Aufsummieren der entsprechenden Spalten

$$\bar{y}_{QC} = \frac{\sum_{i=1}^r y_{i,T}}{\sum_{i=1}^r n_i} = \bar{y}$$

$$\text{Var}(\bar{y}_{QC}) = \frac{r}{n^2} \cdot \frac{1}{(r-1)} \cdot \sum_{i=1}^r (y_{i,T} - n_i \cdot \bar{y})^2 ; \quad n = \sum n_i \quad (i = 1 \text{ bis } r)$$

gelangt man zum "mittleren" Anteil $\bar{y} = p = 0,6211$ und dessen Varianz $\text{Var}(\bar{y}_{QC})$. Multiplikation der Varianz mit der Endlichkeitskorrektur $fpc = 0,71$ und Wurzelziehen ergibt den obigen Standardfehler $SE_{fpc} = 0,0136918$ zur Berechnung des Konfidenzintervalls C.I. = [59,4% ; 64,8%] in guter Übereinstimmung mit den anderen Berechnungen.

Altersadjustierung:

Das obige Ergebnis bezieht sich auf alle untersuchten Kinder von 3 bis 5 Jahren in Kindergärten. Da der jeweilige Anteil an 3, 4, und 5 Jährigen in der untersuchten Stichprobe nicht dem der regionalen Bevölkerung entspricht, ist die Aussage, dass der Anteil der kariesfreien 3 bis 5 Jährigen in der Bevölkerung 62,1% beträgt, nicht korrekt. Hier ist eine Adjus-

tierung erforderlich. Zunächst benötigt man dazu die Anteile kariesfreier Kinder in den einzelnen Altersgruppen. Das Vorgehen wird am Beispiel der 3 Jährigen gezeigt.

Laden Sie die Datei **kiga50agg_3.txt** herunter. Sie hat die gleiche Struktur, wie **kia50agg.txt**, beinhaltet aber nur Daten von 3 Jährigen. Kopieren Sie die Daten in eine Tabellenkalkulation. Gleichzeitig öffnen Sie das WinPepi-Modul und kopieren die Daten (ohne Überschrift) der zwei Spalten „anzahl“ und „Σng“ aus der Tabellenkalkulation in das linke Feld von WinPepi. Dort muss der Cursor vorher in der ersten Zeile links stehen. Die Struktur der Daten muss nach dem Kopieren erhalten bleiben und es müssen 50 Datensätze erkennbar sein. Unter „Size of Population“ wird jetzt 1710 eingetragen, die Anzahl der gemeldeten 3 Jährigen in der Grundgesamtheit aller Kindergärten.

Ergebnis: Anteil kariesfreier 3 Jähriger in der Stichprobe (**Prevalence**): **p** = 73,9%

Konfidenzintervall unter Berücksichtigung der Clusterstruktur und der Endlichkeitskorrektur:

[70,0% ; 77,8%]

Standardfehler: 0,019

Design effect (DEFF): 1,43

ICC: 0,065

Analoges Vorgehen liefert die Ergebnisse für die zwei anderen Altersgruppen. Alle Ergebnisse sind in der folgenden Tabelle zusammengestellt und können mit Hilfe der Dateien kiga50agg_3.txt, kiga50agg_4.txt und kiga50agg_5.txt nachgerechnet werden.

Alter	<i>n</i>	<i>P</i> (%)	<i>ICC</i>	<i>DEFF</i>	Konfidenzintervall Ober- Untergrenze (%)		Präzision (%)
3-5	2312	62,1	0,036	2,59	59,4	64,9	± 2,75
3	517	73,9	0,065	1,43	70,0	77,8	± 3,90
4	885	61,2	0,029	1,65	57,7	64,8	± 3,55
5	910	56,3	0,045	1,90	52,4	60,1	± 3,85

ICC = intraclass correlations coefficient, DEFF = variance inflation factor (design effect)

Nähere Erläuterung zu Design effect (DEFF) findet man unter "Statistik im ÖGD".

Zur Adjustierung benötigt man die Bevölkerungszahlen der 3,- 4,- und 5 Jährigen, die von den statistischen Landesämtern zur Verfügung gestellt werden. Angenommen, im Jahr der Untersuchung seien 3353 Dreijährige, 3378 Vierjährige und 2998 Fünfjährige in diesem Landkreis gemeldet, zusammen 9729. Unter der Annahme gleicher kariesfreier Anteile in der Stichprobe und der Bevölkerung ergeben sich hier:

$3353 \times 0,7389 + 3378 \times 0,6124 + 2998 \times 0,5626 = 6233$ kariesfreie Kinder von insgesamt

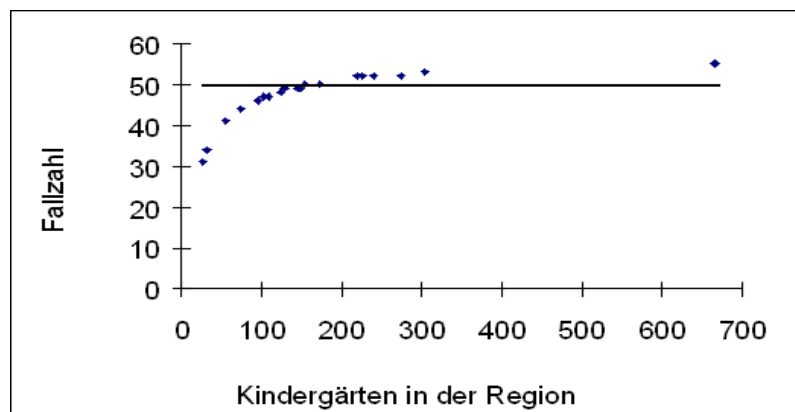
9729 in der Bevölkerung. Es resultiert ein Anteil von 64,1%. Somit unterschätzt die repräsentative Clusterstichprobe mit 62,1% den tatsächlichen Anteil um 2%. Die genannte Annahme kann aufgrund der Stichprobengröße (50 von 170 Kindergärten), der Zufallsauswahl der Kindergärten (Repräsentativität) und der Tatsache, dass mehr als 80% der Vorschulkinder einen Kindergarten besuchen, gerechtfertigt werden.

Daumenregel für Vorschulkinder bei einer gewünschten Präzision von etwa $\pm 3\%$:

Zahnärztliche Reihenuntersuchungen in Landkreisen oder Städten mit über 100 Kindergärten sollten 50 bis 60 Kindergärten pro Schuljahr umfassen. Für Großstädte können es auch 80 bis 90 sein, um eine höhere Präzision zu erreichen. Für Regionen mit deutlich weniger als 100 Kindergärten wäre ein Zusammenschluss mit anderen Regionen oder eine Vollerhebung sinnvoll.

Diese Daumenregel gibt nur eine grobe Orientierung und sollte nur dort angewendet werden, wo keine früheren Untersuchungsdaten existieren (was praktisch nirgends in Deutschland der Fall ist). Die mittlere Belegung der Kindergärten lassen sich aus der normalerweise vorhandenen Liste in den Gesundheitsämtern leicht in einer Tabellenkalkulation berechnen, ebenso die Größe der Studienpopulation. Als DEFF - Wert für die Fallzahlplanung von Schätzungen des Anteils naturgesund eignet sich 2,5. Mit diesen Angaben ist eine Berechnung mit dem Programm WinPepi leicht möglich.

Die folgende Abbildung simuliert erforderliche Fallzahlen (Zahl der Kindergärten) für Regionen mit unterschiedlicher Zahl der Kindergärten unter der Annahme gleicher Werte für die mittlere Belegung der Kindergärten mit 3 - 5 Jährigen in allen Regionen (47), eine vermutete Prävalenz kariesfreier Kinder von 0,6 (60%), eine gewünschte Genauigkeit von 0,03 (3%) und gleichem DEFF = 2,5. Obwohl diese Annahmen von der Realität sicher abweichen, erkennt man einen deutlichen Sättigungseffekt.



Nachtrag zur Fallzahlplanung

In der obigen Tabelle, rechte Spalte, sieht man, dass die Präzision für die einzelnen Altersgruppen geringer ist, als die gewünschten 3%. Verantwortlich dafür ist die Fallzahlplanung, die lediglich für die gesamte Stichprobe der 3-5 Jährigen und nicht für die einzelnen Altersgruppen durchgeführt wurde. Wollte man beispielsweise für die 5 Jährigen eine separate Fallzahlplanung vornehmen mit einer mittleren Belegung von 19, einer gewünschten Präzision von 3%, einer Größe der Studienpopulation von 3191, einem vermuteten Anteil kariesfreier Kinder von 60% und einem DEFF von 1,9, dann lieferte das Programm WinPepi für „Simple random sample“ den Wert 776. Multiplikation mit DEFF = 1,9 ergibt 1474. Geteilt durch 19 ergibt die gesuchte Zahl der Kindergärten von 78, d.h., es müssten deutlich mehr als 50 Kindergärten untersucht werden, um die gewünschte Präzision von 3% auch in der Subgruppe der 5 Jährigen erwarten zu können.

Doch dieser Umfang an Untersuchungen ist aus verschiedenen Gründen nicht immer möglich und daher die Frage interessant, wie sich die Präzision bei kleinerer Fallzahl ändert.

Für die Berechnung der Fallzahl n findet man (ohne Herleitung)

$$n = \frac{N}{1 + \frac{e^2 \cdot N}{(1,96)^2 \cdot p \cdot (1 - p)}}$$

Dabei ist N der Umfang der Grundgesamtheit, e die gewünschte Präzision und p der Anteil kariesfreier Kinder. Die Zahl der Kindergärten ergibt sich aus $n \cdot \text{DEFF} / 47$.

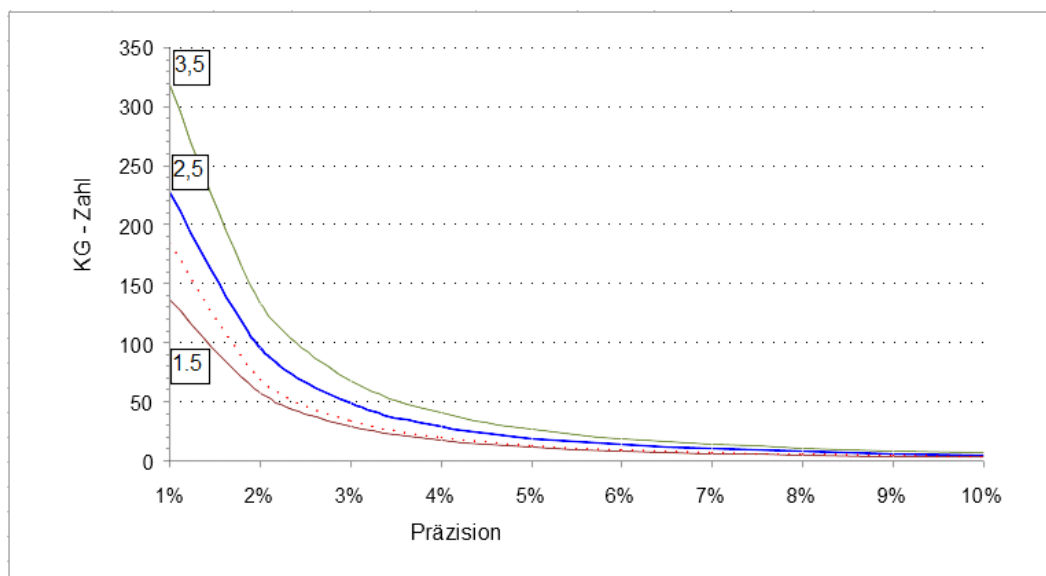


Abb. Fallzahl (Kindergärten) in Abhängigkeit von der Präzision $\pm x\%$

Für $p = 60\%$, $N = 7978$ und eine gewünschte Präzision von $\pm 3\%$ ist, wie oben ausgeführt, eine Stichprobe von 50 aus 170 Kindergärten erforderlich, sofern $\text{DEFF} = 2,5$ beträgt. Für

DEFF = 1,5 sind es 30 und für DEFF = 3,5 etwa 70 Kindergärten. Höhere Genauigkeiten von 1% oder 2% sind nur mit entsprechend größerer Fallzahl realisierbar.

Gibt man sich mit einer Präzision von $\pm 6\%$ zufrieden, müssten bei vorliegenden Rahmenbedingungen je nach DEFF nur 8, 13 oder 18 Kindergärten untersucht werden. Mit abnehmender Präzision sinkt auch der Einfluss von DEFF auf die Fallzahl.

Mit kleinerer Varianz (z.B. $p \cdot (1-p) = 0,16$ - rot gepunktete Linie im Diagramm) verringert sich bei gleichem DEFF = 2,5 die Fallzahl.

Diese Betrachtungen zur Fallzahl für die Schätzung des Anteils kariesfreier Kinder beziehen sich nur auf die einstufige Clusterstichprobe, wie man sie bei der zahnärztlichen Untersuchung in Kindergärten vorfindet.