Einstufige Clusterstichprobe

Die zufällige Auswahl einzelner Einheiten aus einer Grundgesamtheit (einfache Zufallsstichprobe) ist bei vielen Erhebungen nicht möglich, zeitlich zu aufwändig oder zu teuer. Stattdessen ist es häufig günstiger, sich nicht überlappende Gruppen (Cluster), wie z.B. Kindergärten, auszuwählen und innerhalb dieser Gruppen eine Vollerhebung durchzuführen (einstufige Clusterstichprobe).
Die einstufige Clusterstichprobe ist das klassische Stichprobenverfahren für zahnärztliche Untersuchungen in Kindergärten, wenn keine Vollerhebung geplant ist. Aus einer Liste aller Kindergärten in der Region, die in der Regel vorhanden ist, wird zufällig eine bestimmte Anzahl von Einrichtungen ausgewählt und dort werden alle anwesenden Kinder untersucht. Das ist organisatorisch einfach und wird in den Einrichtungen gut akzeptiert. Während Auswahleinheit und Beobachtungseinheit bei der einfachen Zufallsstichprobe (SRS) identisch sind, ist das bei der Clusterstichprobe (CRS) nicht der Fall. Hier sind Auswahleinheit (Kindergarten) und Beobachtungseinheit (Kinder) verschieden. Die Kinder jedes Kindergartens bilden eine eigenständige Gruppe (Cluster), sie interagieren untereinander und sind gemeinsamen Einflussfaktoren ausgesetzt (soziales Umfeld, Ernährungsangebote u.a.), welche zum Beispiel auch ihren Kariesbefall beeinflussen können. Daher sind die Untersuchungsmerkmale der Kinder eines Clusters nicht unabhängig voneinander. In der Regel führt diese Datenstruktur zu einer Erhöhung der Varianz, z.B. der geschätzten dmft-Mittelwerte oder der Anteile kariesfreier Kinder, gegenüber einer einfachen Zufallsstichprobe mit gleicher Kinderzahl und damit auch zu breiteren Konfidenzintervallen für diese Mittelwerte. Der Faktor, um den sich die Varianz vergrößert, wird als *Design Effekt* (DEFF) oder *Variance Inflation Factor* (VIF) bezeichnet und ist zum Beispiel für einen Mittelwert $\bar{x}$ durch folgenden Ausdruck definiert:

$$DEFF = Var_{CRS}(\bar{x}) / Var_{SRS}(\bar{x}) = (SE_{CRS} / SE_{SRS})^2$$

Dabei ist $Var_{SRS}(\bar{x})$ die Varianz des Mittelwertes, bei Behandlung der Daten als einfache Zufallsstichprobe. $Var_{CRS}(\bar{x})$ bezeichnet die Varianz des Mittelwertes unter Berücksichtigung der Clusterstruktur der Daten. Analoges gilt für die Standardfehler SE. Aus dieser Definition folgt die nützliche Gleichung:
$Var_{CRS}(\bar{x}) = Var_{SRS}(\bar{x}) \cdot DEFF$.
Auf den ersten Blick erscheint die Rechnung recht einfach. Möchte man zum Beispiel den mittleren dmft-Wert der Vorschulkinder einer Region aus einer vorhandenen Kindergartenstichprobe (Clusterstichprobe) schätzen, so berechnet man zuerst den arithmetischen Mittelwert über alle Kinder der Stichprobe und die Varianz des Mittelwertes $Var_{SRS}(\bar{x})$ ohne Berücksichtigung der Clusterstruktur der Daten. Dann muss man „nur noch“ mit DEFF multiplizieren und die Wurzel ziehen und erhält so den Standardfehler SE_{CRS} des Mittelwertes und damit auch das Konfidenzintervall des Mittelwertes der Clusterstichprobe. So einfach ist es aber leider nicht, denn DEFF ist normalerweise unbekannt. Zu seiner Berechnung benötigt man eine Angabe zur mittleren Kindergartengröße sowie den *Intracluster Correlation Coefficient* ICC als ein Maß für die Korrelation der Daten innerhalb der Kindergärten. Der ICC kann mit verschiedenen Ansätzen geschätzt werden, auf die wir im Rahmen dieses Beitrages nicht eingehen können. Allgemein kann er für verschiedene Variable oder unterschiedliche Grundgesamtheiten, aber auch für unterschiedliche Stichproben der gleichen Grundgesamtheit, jeweils andere Werte annehmen. So nimmt der ICC für den dmft-Mittelwert einer Stichprobe einen anderen Wert an, als dies zum Beispiel für die Anteilsschätzung aus der gleichen Stichprobe der Fall ist. Mit Hilfe der Gleichung $DEFF = 1 + (\bar{n}_g - 1) \cdot ICC$ schätzt man DEFF aus dem ICC und der gewichteten mittleren Größe der Kindergärten $\bar{n}_g$ in der Region. Dieser gewichtete Mittelwert $\bar{n}_g$ errechnet sich unter Verwendung der Gewichte $w_i = n_i / n$ nach der Formel
$\bar{n}_g = \sum w_i \cdot n_i$.
Dabei ist n_i die Anzahl der Kinder im *i*-ten Kindergarten (i = 1, ... ,m) und $n = \sum n_i$ die Gesamtzahl der Kinder in der Stichprobe. Je größer ein Kindergarten ist, desto größer wird dadurch sein Gewicht $w_i = n_i / n$. Summiert wird über alle r Kindergärten in der Stichprobe.
Bevor wir uns einem realen Datensatz mit 170 Kindergärten zuwenden, zunächst ein kleines Beispiel zur Berechnung des gewichteten Mittelwertes $\bar{n}_g$.

Beispiel: Gewichteter $\bar{n}_g$ und arithmetischer $\bar{n}$ Mittelwert für die Größe dreier Kindergärten (Kiga).

Kiga-Nr	Anzahl n_i der Kinder	n_i^2
K1	29	841
K2	49	2401
K3	75	5625
Summen	153	8867
Mittelwerte	$\bar{n} = 51$	$\bar{n}_g = 58$

Schätzung des dmft-Mittelwertes aus einem realen Datensatz ohne EK

Betrachten wir als Beispiel den realen Datensatz einer zahnärztlichen Untersuchung der Grundgesamtheit von M = 170 Kindergärten eines Landkreises mit N = 7978 Kindern im Alter von 3 bis 5 Jahren (siehe auch SRS_Teil_2). Für diese Grundgesamtheit sind der Mittelwert für den dmft = 1,669 und der Anteil kariesfreier Kinder = 62,4% in der Realität nicht bekannt, sie sollen durch die Werte einer 10%-Stichproben geschätzt werden. Obwohl der Auswahlsatz n/N größer ist als 0,05 verzichten wir vorerst auf eine Endlichkeitskorrektur (EK). Fehler infolge Fehldiagnosen der Zähne sollen ebenfalls unberücksichtigt bleiben.

Aus einer Liste der Kindergärten mit den Nummern 1 bis 170 wählen wir mittels Zufallsgenerator (Computer) oder Zufallszahlentabelle eine Stichprobe von 10% der Kindergärten, also 17 aus 170, aus und berechnen die Schätzung für den arithmetischen Mittelwert $\bar{x}$ = dmft-MW mit Konfidenzintervall.
Datei kiga_17v170_stipro.dta ***Anmerkung: Die Bezeichnung dmf entspricht dmft*** .

Berechnung des dmft-MW mit Konfidenzintervall ohne Clusterstruktur mit STATA:

```
. mean dmf

Mean estimation                 Number of obs    =       932
```

	Mean	Std. Err.	[95% Conf.	Interval]
dmf	1.635193	.1004774	1.438005	1.832382

Berechnung des dmft-MW mit Konfidenzintervall mit Clusterstruktur mit STATA:

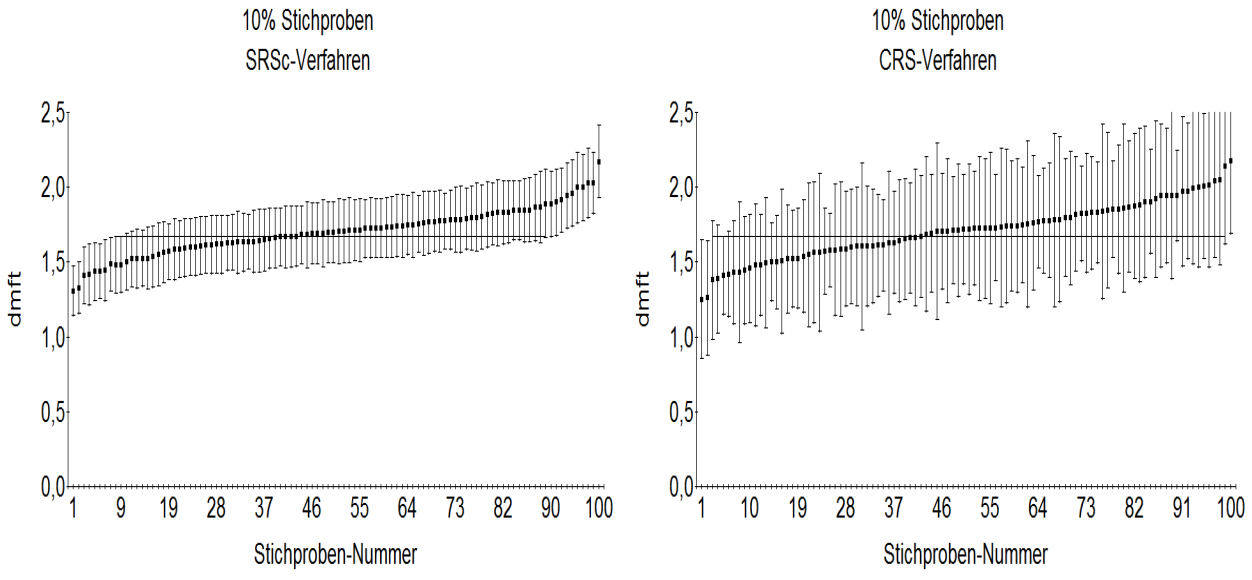
```
. mean dmf, vce(cluster kiganr)

Mean estimation                 Number of obs    =       932

                (Std. Err. adjusted for 17 clusters in kiganr)
```

	Mean	Robust Std. Err.	[95% Conf.	Interval]
dmf	1.635193	.1330155	1.353213	1.917173

Behandelt man die Clusterstichprobe wie eine einfache Zufallsstichprobe (SRS_C), bleibt also die Clusterstruktur der Daten unberücksichtigt, so erhält man in der Regel zu kleine Standardfehler und damit zu schmale Konfidenzintervalle, was zu einer Vergrößerung des Typ I-Fehlers führt (linke Grafik). Jetzt überdecken 18 Konfidenzintervalle von 100 Stichproben-Simulationen nicht den wahren Wert der Grundgesamtheit (wagerechte Linien).



Bei Berücksichtigung des Clusterdesigns wird das Konfidenzniveau von 5% im Mittel eingehalten (rechte Grafik). Allerdings führen die 100 unterschiedlichen DEFF-Werte zu größeren Konfidenzintervallen mit variablen Breiten.
Für die vorliegende Clusterstichprobe mit 932 Kindern kann aus den Standardfehlern (Std.Err.) der obigen STATA-Tabellen der DEFF berechnet werden

DEFF = (0,1330155 / 0,1004774)² = 1,75

Auch ohne Statistikprogramme (STATA, SPSS ua.) sind die Berechnungen in einer Tabellenkalkulation (Excel, OpenOffice Calc ua.) durchführbar.

1. Möglichkeit - Quotienten-Cluster-Schätzer für den dmft-MW (hier $\bar{x}$)

Öffnen Sie die Datei kiga_17v170_stipro.xls in einer Tabellenkalkulation (z.B. Excel) und bestimmen Sie Standardabweichung s und dmft-MW.

kiganr	alter	dmf	ng
2	3	0	1
2	3	0	1
2	4	0	1
2	5	4	0
2	4	0	1

Konfidenzintervall des dmft-MW ohne Clusterstruktur (SRS_C) mit Excel:
Mit der Standardabweichung s = 3,0674 , n = 932 , $\bar{x}$ = dmft-MW = 1,6352 erhält man nach der Formel:

$$\bar{x} \mp z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$ und $Z_{1-\alpha/2} = 1,96$

das Konfidenzintervall (1,438 ; 1,832) in Übereinstimmung mit der STATA-Tabelle oben.

Konfidenzintervall des dmft-MW mit Clusterstruktur (CRS) mit Excel:
Hier bietet sich zuerst der Quotienten-Cluster-Schätzer (QC) an, für den die Daten je Kindergarten zunächst aber zusammengefasst (aggregiert) werden müssen.
Mit dem Kommando: collapse (count) n=dmf (sum) sumdmf=dmf , by(kiganr) erzeugt STATA aus *kiga_17v170_stipro.dta* eine neue Datei mit dem Namen *kiga_17v170_stipro_aggr.dta* und den gewünschten Werten.

Zum Aggregieren der Werte direkt in der Excel-Tabelle helfen evtl. folgenden Formeln weiter:
ZÄHLENWENN(Suchbereich; Suchkriterium) (zum Beispiel =ZÄHLENWENN($A$2:$A$933;F2)) und
SUMMEWENN(Suchbereich; Suchkriterium [; Summenbereich]) (=SUMMEWENN($A$2:$A$933;F2;$C$2:$C$933)

kiganr	n_i	$\sum dmf_i$	$n_i \cdot \bar{x}$	$(\sum dmf_i - n_i \cdot \bar{x})^2$
2	59	65	96.4764	990.7629
3	42	67	68.6781	2.8160
5	33	75	53.9614	442.6240
6	67	100	109.5579	91.3540
38	57	122	93.2060	829.0944
62	58	105	94.8412	103.2013
81	49	77	80.1245	9.7622
87	75	133	122.6395	107.3405
92	72	181	117.7339	4002.5999
112	62	117	101.3820	243.9230
127	29	10	47.4206	1400.3011
128	65	108	106.2875	2.9325
133	38	73	62.1373	117.9975
142	68	114	111.1931	7.8786
150	55	113	89.9356	531.9659
159	57	33	93.2060	3624.7626
166	46	31	75.2189	1955.3092
			$\sum$	14464.6256

Der arithmetische MW bleibt beim Quotienten-Cluster Schätzer bestehen $\bar{x}_{QC}$ = dmft-MW = $\bar{x}$ = 1,6352 .
Die Varianz des Mittelwertes berechnet sich nach

$$\mathrm{Var}\left(\bar{x}_{QC}\right) = \frac{r}{n^2} \cdot \frac{1}{(r-1)} \cdot \sum_{i=1}^{r} \left(x_{i,T} - n_i \cdot \bar{x}_{QC}\right)^2$$

Dabei ist $x_{i,T} = \sum dmf_i$ die Summe der dmf - Werte im jeweiligen Kindergarten und r = 17 die Zahl der Cluster.
Setzt man für die Summe (roter Pfeil) den Wert aus der Tabelle ein, liefert die Formel $\mathrm{Var}(\bar{x}_{QC}) = 0{,}01769$ und damit den Standardfehler SE_{QC} (Std.Err.) = 0.1330 aus obiger STATA-Tabelle.
Mit $t_{16;0,05} = 2{,}12$ (die Freiheitsgrade richten sich hier nach der Clusterzahl) erhält man das Konfidenzinterv.

$(\bar{x}_{QC} - t_{16;0,05} \cdot SE_{QC} ; \bar{x}_{QC} + t_{16;0,05} \cdot SE_{QC}) = (1{,}353 ; 1{,}917)$ in Übereinstimmung mit der STATA-Tabelle.

2. Möglichkeit - Adjustierung mit DEFF

Eine weitere Möglichkeit der Varianzschätzung eröffnet die oben genannte Formel für $\mathrm{Var}_{CRS}(\bar{x})$:
$\mathrm{Var}_{CRS}(\bar{x}) = \mathrm{Var}_{SRS}(\bar{x}) \cdot DEFF$ durch Berechnung von $DEFF = 1 + (\bar{n}_g - 1) \cdot ICC$. Den hierzu benötigten ICC erhält man aus:

$$ICC = \frac{MSB - MSW}{MSB + (\bar{n} - 1) \cdot MSW}$$

Dabei ist $\bar{n}$ die mittlere Clustergröße, MSB die mittlere Quadratsumme zwischen den Clustern (Mean Square Between) und MSW die mittlere Quadratsumme innerhalb der Cluster (Mean Square Within). Zur Berechnung öffnen wir zunächst die Datei *kiga_17v170_stipro.xls* mit den 932 Einzeldaten in Excel und berechnen die Gesamtvarianz für alle Werte: $\mathrm{Var}(x) = \sum (x_i - \bar{x})^2 / (n-1)$. Multiplikation mit (n-1) ergibt die SST (Sum of Squares Total), die Summe der Abweichungsquadrate vom dmft - Mittelwert über alle Werte. Mit z.B. "varianz(c2:c933)" erhält man Var(x) = 9,4092 und SST = 9,4092·931 = 8759,9657 .

Es gilt: **SST = SSB + SSW** = Summe der Abweichungsquadrate zwischen den Clustern + Summe der Abweichungsquadrate innerhalb der Cluster, sowie **MSB = SSB / (r-1)** und **MSW = SSW / (n-r)** .
Die Berechnung von SSW ist recht aufwendig, daher ermitteln wir erst SSB und nutzen danach den oben genannten Zusammenhang für die "Sum of Squares". Dazu erweitern wir unsere Tabelle der aggregierten Daten um eine Spalte der Clustermittelwerte $\bar{x}_i = \sum dmf_i / n_i$ für i = 1 bis 17 und löschen die letzten zwei Spalten.

kiganr	n_i	$\sum dmf_i$	$\bar{x}_i$	$n_i \cdot (\bar{x}_i - \bar{x})^2$
2	59	65	1.1017	16.7926
3	42	67	1.5952	0.0670
5	33	75	2.2727	13.4128
6	67	100	1.4925	1.3635
38	57	122	2.1404	14.5455
62	58	105	1.8103	1.7793
81	49	77	1.5714	0.1992
87	75	133	1.7733	1.4312
92	72	181	2.5139	55.5917
112	62	117	1.8871	3.9342
127	29	10	0.3448	48.2862
128	65	108	1.6615	0.0451
133	38	73	1.9211	3.1052
142	68	114	1.6765	0.1159
150	55	113	2.0545	9.6721
159	57	33	0.5789	63.5923
166	46	31	0.6739	42.5067
	54,82		$\sum$	276.4407

$$SSB = \sum n_i \cdot \left(\bar{x}_i - \bar{x}\right)^2$$

Die Summe ergibt SSB = 276,4407 und daraus SSW = SST - SSB = 8759,9657 - 276,4407 = 8483,5249
Die obigen Formeln liefern: MSB = 276,4407 / 16 = 17,2775 und MSW = 8483,5249 / 915 = 9,2716
Diese Ergebnisse überprüfen wir mit dem eigentlich nicht vorhandenen Statistikprogramm STATA

```
. oneway dmf kiganr

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      276.440745      16   17.2775466      1.86     0.0204
 Within groups      8483.52492     915   9.27161193
------------------------------------------------------------------------
    Total           8759.96567     931   9.4092005
```

und kommen mit $\bar{n}$ = 54,82 jetzt zur Berechnung des ICC nach obiger Formel:
ICC = (17,2775 - 9,2716) / (17,2775 + 53,82·9,2716) = 0,0155

Mit $\bar{n}_g$ = 57,96 erhalten wir DEFF = 1 + ($\bar{n}_g$ - 1)·ICC = 1 + 56,96·0,0155 = 1,883 und somit aus der Formel
$Var_{CRS}(\bar{x}) = Var_{SRS}(\bar{x}) \cdot DEFF$ die Varianz des Mittelwertes unter Clusterdesign.
Für das Konfidenzintervall sind $\sqrt{(Var_{SRS} \cdot DEFF)} = \sqrt{(9,4092 \cdot 1,883)} = 4,2092$ und $\sqrt{n} = 30,5287$, so erhält man mit

$$\bar{x} \mp 1,96 \cdot \frac{\sqrt{Var_{SRS} \cdot DEFF}}{\sqrt{n}}$$ ein Konfidenzintervall

(1,6352 - 1,96·4,2092 / 30,5287 ; 1,6352 + 1,96·4,2092 / 30,5287) = **(1,3650 ; 1,9054)**
Im Vergleich mit dem Ergebnis aus STATA [(1,3532 ; 1,9171)] ist das ein etwas schmaleres Intervall.

3. Möglichkeit - Clusterlevel

Beim Clusterlevel-Schätzer für die Varianz werden die unterschiedlichen Größen und die Variationen innerhalb der Kindergärten nicht berücksichtigt. Trotzdem erhält man recht gute Intervallschätzungen, sofern die Clustermittelwerte und die Clustergrößen nicht zu stark korrelieren (Barnett S.141) . Analog der einfachen Zufallsstichprobe (SRS) sind hier Auswahleinheit (Cluster) und Beobachtungseinheit (Cluster) identisch und der jeweilige Clustermittelwert $\bar{x}_i$ das Merkmal. Man berechnet den Gesamtmittelwert $\bar{x}^*$ und die Varianz Var* für die r Cluster z.B. mit der Tabelle unter Punkt 2. aus

$$\bar{x}^* = \frac{1}{r}\sum_{i=1}^{r} \bar{x}_i \ , \ Var^* = \frac{1}{r-1} \cdot \sum_{i=1}^{r} \left(\bar{x}_i - \bar{x}^*\right)^2$$ und erhält mit SE* = √(Var*/r) den Standardfehler für das Konfidenzintervall. Es ergeben sich: Var* = 0,3619 und damit SE* = 0,1459 .
Mit Mit $t_{16;0,05}$ = 2,12 erhält man das Konfidenzintervall
(1,6352 - 2,12·0,1459 ; 1,6352 + 2,12·0,1459) = (1,3259 ; 1,9445)
Der Vergleich mit dem Ergebnis aus STATA [(1,3532 ; 1,9171)] zeigt ein etwas breiteres Intervall.

Korrektur der Konfidenzintervalle für endliche Grundgesamtheiten mit n / N > 0,05

SPSS, und auch andere Programme, liefern Konfidenzintervalle ohne Endlichkeitskorrektur (EK), die jedoch dann berücksichtigt werden muss, wenn n / N > 0,05.
Wie korrigiert man ?
Angenommen man erhält aus SPSS ohne EK: $\bar{x}$ = 1,6352 K.I. = **(1,3532 ; 1,9172)**
also $\bar{x}$ = 1,6352 ± 0,2820
Wegen **EK = 1 - n/N** = 1 - 932/7978 = 0,8832 folgt √EK = 0,9398
Die halbe Breite des K.I. (hier 0,2820) multipliziert mit √EK = 0,9398 ergibt das gesuchte Konfidenzintervall mit Endlichkeitskorrektur: $\bar{x}$ = 1,6352 ± 0,2820·0,9398 = 1,6352 ± 0,2650

Die Endlichkeitskorrektur bedingt eine Verringerung der Intervallbreite auf **(1,3702 ; 1,9002)**.
Eine Rechenkontrolle erfolgt mit STATA's Surveymodul:

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
dmf	1.635193	.1250047	1.370195	1.900191

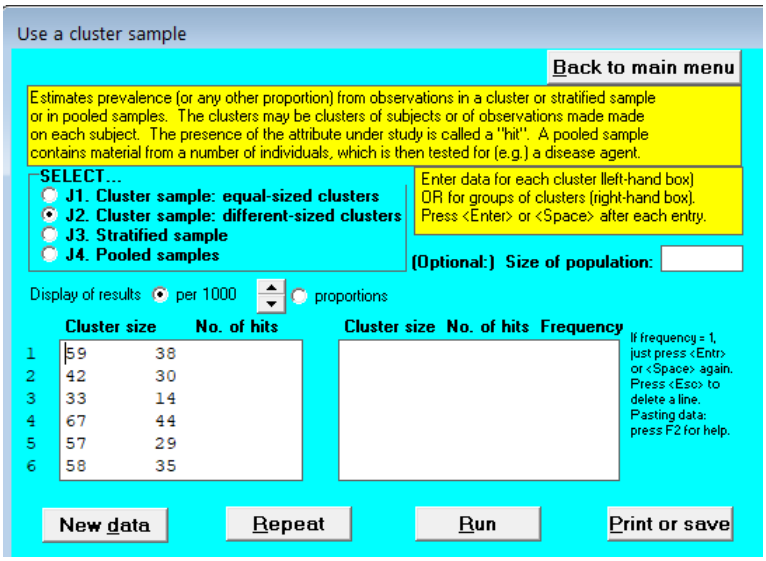
Schätzung des mittleren Anteils kariesfreier Kinder

Wir bleiben in diesem Abschnitt beim gleichen Datensatz und der gleichen Stichprobe, wollen aber die Berechnungen nicht mehr „zu Fuß“ durchführen, sondern das Programm WinPepi, Version 11.18, von J.H. Abramson verwenden, das unter www.brixtonhealth.com aus dem Internet geladen werden kann und unter fast allen Windows-Versionen funktioniert. Voraussetzung für die Schätzung des mittleren Anteils kariesfreier Kinder ist eine Umcodierung der Daten, die in der Datei *kiga_17v170_stipro.xls* bereits erfolgt ist. Dazu wurde eine neue Variable „ng“ (naturgesund) definiert und alle Kinder mit einem dmf = 0 erhielten die Codierung ng = 1 während alle mit einem dmft > 0 die Codierung ng = 0 erhielten. Jedes kariesfreie Kind wird in WinPepi als „Hit“ bezeichnet. Beispielsweise hat ein Kindergarten mit 43 Kindern und davon 28 kariesfreien somit 28 Hits.

Zunächst werden die Daten wieder aggregiert, z.B. mit dem STATA-Kommando:
collapse (count) n=ng (sum) sumng=ng , by(kiganr) Das sumng wird in der Tabelle wieder zu "ng".

Man erhält unten stehende Tabelle mit n = Zahl der Kinder in den Kindergärten ("Cluster size") und ng = Zahl der kariesfreien Kinder ("No. of hits"). Die beiden Spalten lassen sich direkt nach WinPepi kopieren, wobei darauf zu achten ist, dass nur Zahlen, nicht die Spaltenüberschriften, kopiert werden oder bei Kopie aus STATA die erste Zeile in WinPepi gelöscht wird.
Aufrufen von WinPepi / Describe / J2

kiganr	n	ng
2	59	38
3	42	30
5	33	14
6	67	44
38	57	29
62	58	35
81	49	33
87	75	50
92	72	39
112	62	36
127	29	25
128	65	38
133	38	19
142	68	50
150	55	34
159	57	50
166	46	38



Als Ergebnis erhält man verschiedene Schätzungen, von denen wir die nach Cochrane verwenden.

```
Prevalence = 64.59 per 100.
   By Cochran's procedure:
     95% C.I. = 58.97 to 70.21 per 100.
     S.E. = 0.0265
     Design effect = 2.86
Rate of homogeneity
    (intraclass correlation coefficient) = 0.041
```

Trägt man in "Size of population" noch den Wert 7978 ein, so wird auch die Endlichkeitskorrektur (EK) berücksichtigt und es ergibt sich ein etwas schmaleres `95% C.I. = 59.31 to 69.87 per 100.`

STATA liefert ganz ähnliche Ergebnisse für die Rechnung ohne EK:

	Proportion	Linearized Std. Err.	Logit [95% Conf.	Interval]
ng				
0	.3540773	.0264995	.3001064	.41204
1	.6459227	.0264995	.58796	.6998936

```
. estat effects, deff
```

	Proportion	Linearized Std. Err.	DEFF
ng			
0	.3540773	.0264995	2.85855
1	.6459227	.0264995	2.85855

und mit EK:

	Linearized		Logit	
	Proportion	Std. Err.	[95% Conf.	Interval]
ng				
0	.3540773	.0249036	.3032227	.408461
1	.6459227	.0249036	.591539	.6967773

Fazit: Wählt man als Elemente einer Stichprobe Kindergärten statt Kinder und wird dieses Clusterdesign der Stichprobe bei der Auswertung der Daten nicht berücksichtigt, so erhält man in der Regel für den geschätzten Mittelwert ein zu schmales Konfidenzintervall, das eine Genauigkeit vortäuscht, die tatsächlich nicht existiert. Damit vergrößert sich aber die Irrtumswahrscheinlichkeit von 5% auf einen deutlich höheren Wert, vielleicht 20%, so dass unter Umständen jede fünfte Stichprobe nicht repräsentativ ist für die Grundgesamtheit, die untersucht wurde. Diese Überlegungen gelten nicht nur für Kindergärten sondern z.B. auch für Schulen.

Mit dieser Stichprobe von 17 Kindergärten aus 170 Kindergärten erreicht man für den dmft-MW eine Präzision von etwa +/- 0,265 und für den Anteil etwa +/- 5,3 Prozentpunkte.

Fallzahlberechnung für Clusterstichproben (CRS)

Ohne EK erhält man für die Präzision e aus der Formel von Seite 5 oben:

$$e = 1{,}96 \cdot \frac{\sqrt{\mathrm{Var}_{SRS} \cdot DEFF}}{\sqrt{n}} \qquad \text{Umformen nach n ergibt} \quad n_{CRS} = \frac{(1{,}96)^2 \cdot \mathrm{Var}_{SRS} \cdot DEFF}{e^2} = n_{SRS} \cdot DEFF$$

Wird die Endlichkeitskorrektur berücksichtigt, erhält man mit N = 7978 folgenden Ausdruck (Ableitung analog SRS_Teil2, S. 8)

$$n_{CRS} = \frac{N}{1 + \dfrac{N \cdot e^2}{(1{,}96)^2 \cdot \mathrm{Var}_{SRS} \cdot DEFF}} \quad = \quad N / (1 + B)$$

Wird für den dmft-MW beispielsweise eine Präzision von etwa ± 10% angestrebt, so erhält man unter Verwendung der Ergebnisse der obigen kleinen Stichprobe (17 aus 170), die man z.B. als Pilotprojekt durchgeführt hat, mit e = 0,16 ; Var_{SRS} = 9,41 ; DEFF = 1,75 und der bekannten mittleren Belegung der Kindergärten $\overline{n}$ = 7978 / 170 = 46,9 ≈ 47

ohne EK: n_{CRS} = (1,96² · 9,41 · 1,75) / 0,16² = 2471 Kinder , somit etwa 2471 / 47 = **53 Kindergärten**

mit EK : B = (7978 · 0,16²) / (1,96² ·9,41 · 1,75) = 3,22845 , n_{CRS} = 7978 / 4,22845 = 1887 (**40 Kigä)**

Doch Vorsicht ! Var_{CRS} weist im Gegensatz zu Var_{SRS} deutliche Schwankungen auf (siehe Seite 2), denn jede Stichprobe liefert nicht nur einen etwas anderen dmft-MW, sondern auch einen jeweils anderen Wert für DEFF. Ein Wert von DEFF = 1,9 z.B. ergibt ein n_{CRS} = 2683 und damit etwa 57 Kindergärten. Ähnliche Rechnungen führten zur Stichprobe **kiga_57.sav** (siehe SRS_Teil2), für die folgende Schätzungen für den dmft-MW mit den Formeln in diesem Beitrag nachvollziehbar sind:

	dmft-MW	**Konfidenzintervall**	**Präzision**
Quotienten-Cluster-Schätzer	1,8146	(1,6722 ; 1,9569)	± 0,1424
Adjustierung	1,8146	(1,6771 ; 1,9521)	± 0,1375
Cluster-Level	1,8146	(1,6710 ; 1,9582)	± 0,1436

Für eine weitergehende Beschäftigung mit diesem Thema wird folgendes Buch empfohlen:
Stichproben von Göran Kauermann und Helmut Küchenhoff, Springer-Verlag 2010.