Schätzung des mittleren dmft - Wertes mittels Clusterstichproben aus einer regionalen Grundgesamtheit von Kindergärten - ein Beispiel

(Skalenniveau des dmft: ordinal / pseudo - metrisch)

Für die Präzision der Schätzung erwarten wir etwa ± 10% bei einem Stichprobenumfang von 50 Kindergärten. Es wird der übliche dmft-MW= ∑dmf / n geschätzt (siehe Textende). Eine Auswertungen quantitativer Daten (dmft, Körpergewicht, Blutzucker u.a.) aus Clusterstichproben sind mit dem Programm WinPepi nicht möglich.

Hier bieten sich zur Schätzung des mittleren dmft - Wertes aus zahnärztlichen Daten, welche zum Beispiel durch Jugendzahnärzte in Kindergärten erhoben wurden, das kostenfreie Programm **EPI INFO** oder das kommerzielle Statistikprogramm **STATA** (siehe Rubrik „STATA - Intro“) an. Der Quotienten-Cluster-Schätzer (**QC-Schätzer**) zur Rechnung per Hand kann auch in einer Tabellenkalkulation berechnet werden und liefert STATA-Ergebnisse.

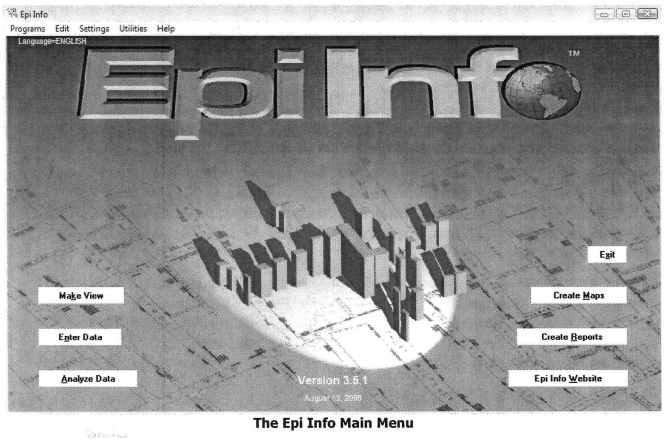
Die Daten können als Textdatei (*.txt), als dBase IV oder auch im Excel - Format vorliegen. Sie müssen folgendermaßen angeordnet sein: Jede Zeile repräsentiert ein Kind und jede Spalte eine Variable. Weitere Angaben, die für das folgende **Beispiel** notwendig sind:

- Größe der Studienpopulation, d.h., Anzahl gemeldeter 3-5 jähriger Kinder in allen Kindergärten der Region (hier 7978) und die Anzahl in der jeweils vorgesehenen Altersgruppe (hier 1710 Dreijährige, 3077 Vierjährige, 3191 Fünfjährige). 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. Die Anteile der jeweiligen Altersgruppen in der Studienpopulation lassen sich näherungsweise aus den Anteilen der randomisierten Stichprobe schätzen. Das funktioniert aber nur, wenn die Kindergärten in der Stichprobe wirklich zufällig ausgewählt wurden und die Stichprobe groß genug ist (z.B. 50 Kindergärten).
- Zahl untersuchter Kinder pro Kindergarten und je Altersgruppe,
- Wenn die Untersuchungsdaten als Textdatei *.txt (oder dBase, Excel) vorliegen sollen, müssen sie aus dem jeweils verwendeten Erfassungsprogramm (Octoware, ISGA u.a.) exportiert werden. Dies funktioniert ggf. erst auf Nachfrage beim Softwareanbieter.

Beispielrechnung mit Epi Info für die Altersgruppe der 3 - 5 Jährigen:

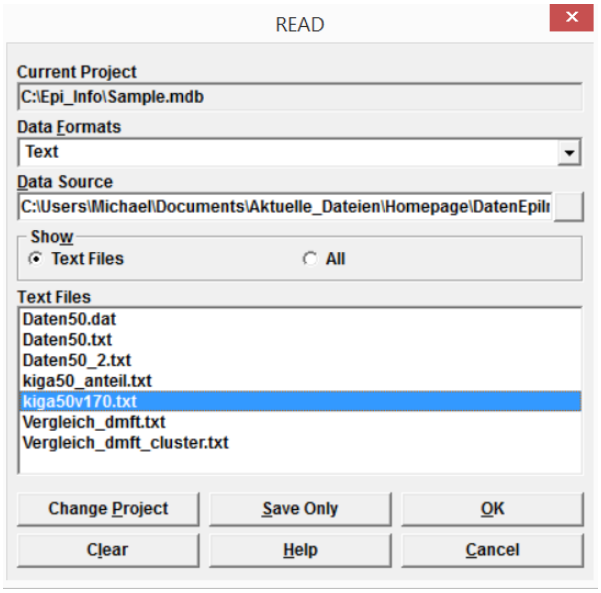
Laden Sie die Datei **kiga50v170.txt** herunter. Starten Sie EPI-INFO 3.5.4.und wählen Sie im Startmenü das Programm „Analyze Data“ (Abb. links, links unten). Diese ältere Version erhält man unter: **www.cdc.gov/epiinfo/support/downloads/prevversions.html**

© 2024 M. Herzog



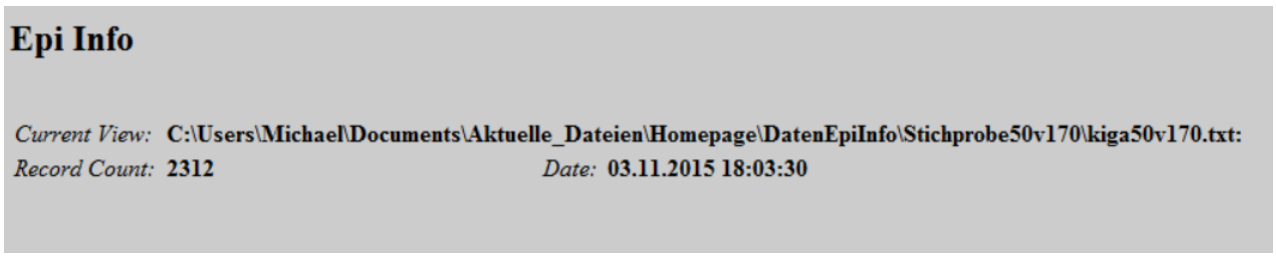
The Epi Info Main Menu

Startmenü für Epi Info 3.5.1 / 3.5.4



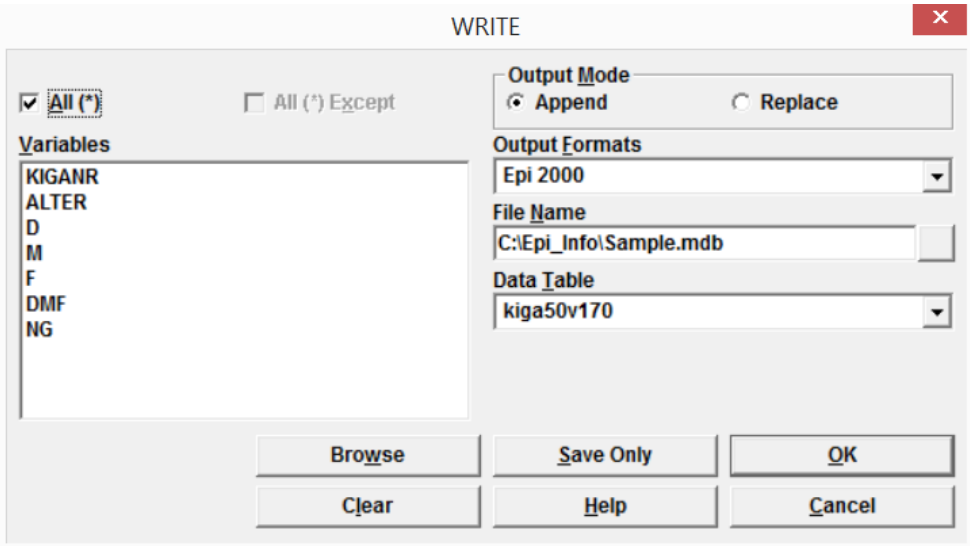
Auswahlfenster

Unter „Data - Read (Import)“ (linkes Menü im Programm *Analysis*) öffnet sich ein Auswahlfenster (siehe Abb. rechts) und Sie wählen unter „Data Formats“ den Eintrag „Text“. Unter „Data Source“ wählen Sie nach Klick auf den rechten Button die Datei **kiga50v170.txt** aus dem Download-Verzeichnis (oder dem Ort, an dem Sie diese Datei nach dem Downloaden gespeichert haben). Den sich jetzt öffnenden Editor schliessen Sie wieder und bestätigen im Auswahlfenster mit OK. Das nächste Auswahlfenster „Filespec“ bestätigen Sie ohne Änderungen ebenfalls mit OK.
Es erscheint im rechten oberen Fenster die Meldung, dass die Daten eingelesen wurden (*Record Count:* 2312) und einige andere Angaben.



Um mit der Datei auch später noch arbeiten zu können, sollte sie im Epi 2000 Format in EPI-INFO gespeichert werden.

© 2024 M. Herzog

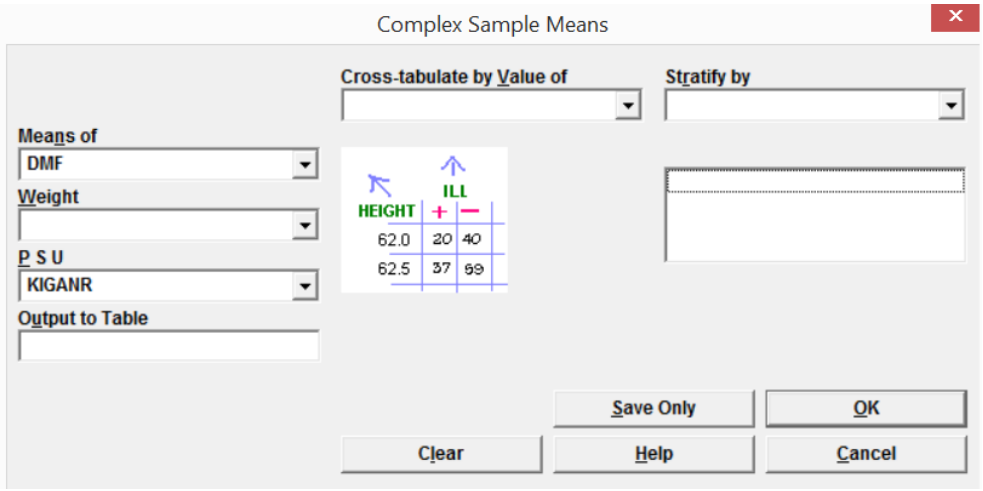


Hierzu rufen Sie im Programm *Analysis* im linken Menü „Write (Export)“ auf. Das Output Format Epi 2000 ist im Auswahlfenster schon vorgegeben, als „File Name“ wählen Sie nach Klick auf das rechte kleine Quadrat „Sample.mdb“ und in „Data Table“ schreiben sie den Namen der Datei **kiga50v170**. Sie haben außerdem noch die Möglichkeit, einige nicht interessierende Variable zu entfernen, um dadurch beispielsweise die Daten zu anonymisieren. Danach bestätigen Sie mit OK. Jetzt können Sie die Daten jederzeit mit „Read“ einlesen (siehe oben), wobei unter „Data Source“ stehen sollte C:\Epi_Info\Sample.mdb und der Button „All“ aktiviert sein muss. Jetzt sind die Daten erfolgreich eingelesen und können mit dem Befehl „Statistics - List“ (ohne weitere Änderungen) und mit OK, angesehen werden. Jeder Datensatz (entspricht einem Kind) beinhaltet die Nummer des Kindergartens (KIGANR), das Alter, Angaben zum dmft und eventuell andere individuelle Befunde. Ist der dmft = 0, zeigt die Variable NG (naturgesund) den Eintrag Yes (für 1), ansonsten No (für 0), was aber keinen Einfluss auf die Rechnungen hat. Wählt man beispielsweise „Advanced Statistics“ und unter „Complex Sample Frequencies“ unter "Frequency of" die Variable NG und bei "PSU" die Variable KIGANR, so erhält man nach OK den schon bekannten Anteil kariesfreier Kinder von 62,11%.

KIGANR	ALTER	D	M	F	DMF	NG
2	3	5	0	1	6	No
2	3	1	0	0	1	No
2	3	0	0	1	1	No
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes
2	3	0	0	0	0	Yes

© 2024 M. Herzog

Für die Berechnung (Schätzung) des mittleren dmft-Wertes bei Clusterstichproben wählt man unter „Advanced Statistics - Complex Sample Means“ (siehe Abb. unten) bei „Means of“ die Variable DMF und unter „PSU“ die Variable KIGANR. Nach OK wird eine Tabelle ausgegeben mit dem mittleren dmft von 1,713 und dem Konfidenzintervall (1,537 ; 1,889), oder analog **1,713 ± 0,176**. „PSU“ bedeuted „Primary Sampling Unit“ und bezieht sich auf die Elemente, die ausgewählt wurden - hier die Kindergärten.



	DMF						
	Count	Mean	Std Error	Confidence Limits		Minimum	Maximum
				Lower	Upper		
TOTAL	2312	1,713	0,088	1,537	1,889	0,000	20,000

Wichtige Anmerkung: Wählt man "Statistics - Means" und setzt unter „Means of“ die Variable DMF, so erhält man folgendes Ergebnis **ohne Berücksichtigung der Clusterstruktur** der Daten. Der Std Error (SE) ergibt sich aus SE = Std Dev / √2312
SE = 3,1060 / 48,083261 = 0,0645963 und hieraus ein Konfidenzintervall von 1,586 bis 1,840 (±0,127) . 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.

Obs	Total	Mean	Variance	Std Dev
2312	3960,0000	1,7128	9,6475	3,1060

Ergebnis mit „Means“ im Menü „Statistics“

Achtung :

Leider berücksichtigt Epi Info nicht die häufig wichtige Endlichkeitskorrektur (!)

© 2024 M. Herzog

$\left(1-\frac{n}{N}\right)$. Hier bedeuten ***n*** die Zahl der Kinder in der Stichprobe und ***N*** die Zahl in der Studienpopulation (Grundgesamtheit). Das Programm rechnet unter der Annahme, dass die Stichprobe klein und die Grundgesamtheit sehr groß ist, zum Beispiel ***n*** = 100 und ***N*** = 1.000.000. Dann wäre ***n / N*** = 0,0001 und die Endlichkeitskorrektur (EK) bis auf einen vernachlässigbaren Fehler gleich 1 (genau 0,9999).
Mit ***n*** = 2312 und ***N***= 7978 erhält man aber für EK = (1 - 2312 / 7978) = 0,71.
Daraus die Wurzel ergibt 0,843. Der Wert für die Genauigkeit ±0,176 ist noch mit dieser Zahl 0,843 zu multiplizieren, um das richtige Konfidenzintervall angeben zu können, was demnach lautet 1,713 ± (0,176 * 0,843) => **1,713 ± 0,148** oder analog **(1,565 ; 1,861)** statt (1,537 ; 1,889).
Mit der Endlichkeitskorrektur wird das Konfidenzintervall etwas schmaler. Die Anforderung an die Genauigkeit von ± 10% ist erfüllt.

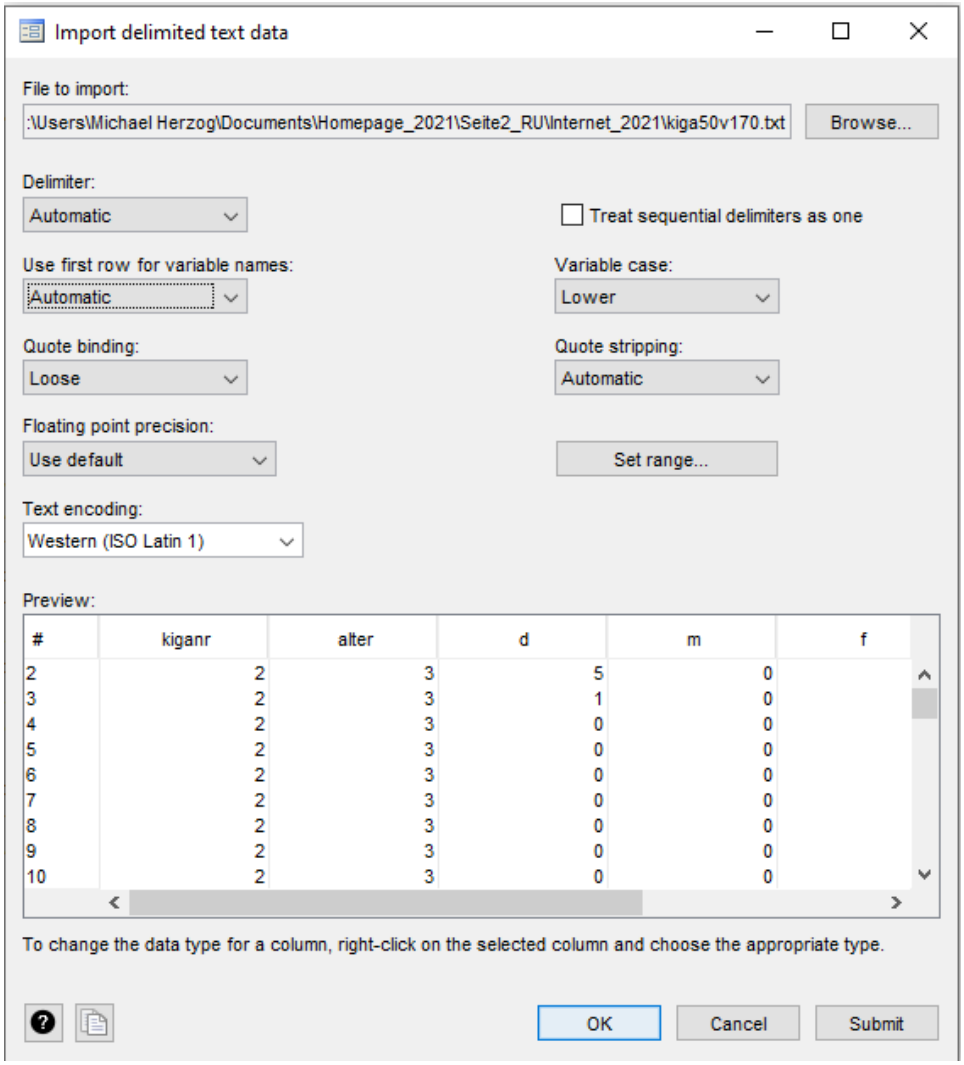
Der aus der Clusterstichprobe mit 50 Kindergärten geschätzte mittlere dmft-Wert der Kinder im Alter von 3 bis 5 Jahren in der Grundgesamtheit von 170 Kindergärten beträgt 1,713. Der wahre Wert von 1,67 in dieser Grundgesamtheit (den wir aber nicht kennen) liegt mit 95%-iger Wahrscheinlichkeit im Intervall zwischen 1,565 und 1,861 (mit EK). Die Genauigkeit der Schätzung kann angegeben werden mit ±0,148 . Die Verteilung der dmft-Einzelwerte ist stark rechtsschief. Der dmft-Mittelwert kennzeichnet nicht das Zentrum der Verteilung, ist aber als Zufallsvariable für Stichprobenumfänge n > 30 (besser 100) asymptotisch normalverteilt (Zentraler Grenzwertsatz) und somit die Mittelwertberechnung gerechtfertigt.
Der DEFF wird bei Epi Info nur im Modul „Complex Sample Frequencies“ ausgegeben. Dieser Wert ist für Berechnungen im Modul „Complex Sample Means“ **nicht** verwendbar. Stattdessen muss man ihn über die Standardfehler (SE) des Mittelwertes mit und ohne Clusterdesign per Hand berechnen. Die Endlichkeitskorrektur bleibt dabei unbeachtet.
Für den SE ohne Clusterdesign liefert Epi Info: 0,0646 und für den SE mit Clusterdesign 0,088 . Hieraus erhält man den DEFF zu DEFF = (0,088 / 0,0646)² = 1,856 .
STATA liefert hier DEFF = 1,84276. Der geringe Unterschied ist rundungsbedingt.

Altersadjustierung:
Für eine Aussage zum dmft 3 - 5 jähriger Kinder in der Bevölkerung fehlt noch eine Altersadjustierung, die in entsprechender Weise wie bei der Anteilsschätzung vorzunehmen ist.

© 2024 M. Herzog

Berechnung mit STATA für die Altersgruppe der 3 - 5 Jährigen:

Laden Sie die Datei **kiga50v170.txt** herunter. Starten Sie STATA und wählen Sie unter "File/Import" - Text data (delimited, ...) aus.



Mit "Browse" suchen Sie nach der heruntergeladenen Datei kiga50v170.txt, die Sie mit Tabulator-Trennzeichen gespeichert hatten. Im Preview sehen Sie die Daten. Mit OK werden diese ins Programm importiert. Jetzt speichern Sie die Datei im STATA - Format **kiga50v170.sta** zur späteren weiteren Bearbeitung. Für die Endlichkeitskorrektur (fpc) wird die Variable fpc = n / N = 2312 / 7978 = 0,289797 und für sampling weight pw = 7978 / 2312 hinzugefügt.

Für die Berücksichtigung der Endlichkeitskorrektur verwendet man in STATA Survey-Kommandos. Zuerst werden Setup-Informationen eingegeben. Als "primary sampling unit" (SU) gibt man "kiganr" an und als finite population correction - "fpc". Ausführliche Beschreibungen zu den Berechnungen findet man im STATA Survey Data Reference Manual unter www.stata.com .

© 2024 M. Herzog

```
. 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 : mean dmf
(running mean on estimation sample)

Survey: Mean estimation

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

	Linearized			
	Mean	Std. Err.	[95% Conf.	Interval]
dmf	1.712803	.0738989	1.564297	1.861308

Über ein weiteres Kommando erhält man DEFF = 1,84276 .

```
. estat effects, deff
```

	Linearized		
	Mean	Std. Err.	DEFF
dmf	1.712803	.0876894	1.84276

Wir sehen eine gute Übereinstimmung mit Epi Info.

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, $y_{i,T}$(dmf) = Summe der dmf-Werte im i-ten Kindergarten, berechnet man je Kiga: $n_i * \overline{y}$ und $(y_{i,T} - n_i * \overline{y})^2$ mit $\overline{y}$ = 1,7128 .

Tab. Ausschnitt aus der Tabelle

kiganr	$y_{i,T}$ (dmf)	n_i	$n_i * \overline{y}$	$(y_{i,T} - n_i * \overline{y})^2$
2	65	59	101.05536	1299.98922
7	146	61	104.48097	1723.82995
9	76	52	89.06574	170.71366

Nach Aufsummieren der rechten Spalte erhält man unter Verwendung der Formel im Text von "Schätzung_Anteil" (S. 5) für die Var($\overline{y}_{QC}$) = 0,007689 und nach einer Multiplikation mit fpc = 0,71 für den SE_{fpc} = 0,0739 .
Damit ergibt sich für das Konfidenzintervall C.I. = (1,57 ; 1,86) in guter Übereinstimmung mit den beiden anderen Rechnungen.

Anmerkung zum dmft-Mittelwert:

In einer Population von n Kindern gibt es derzeit (in Mitteleuropa) eine Anzahl n_g Kinder mit naturgesunden Gebissen und eine Zahl n_k Kinder mit Karieserfahrung (KE). Somit ist $n = n_g + n_k$.
Der Prävalenzbegriff bezieht sich auf Krankheit in einer Population, hier Zahnkaries, und ist definiert als:
p_{nk} = *(Zahl Kinder mit KE) / (Zahl der Kinder unter Risiko zu einem bestimmten Zeitpunkt) .*
Da in der Regel alle Kinder dem Risiko der Karies ausgesetzt sind, ist $p_{nk} = n_k / n$.
Bei der Schätzung des "Anteils kariesfreier Kinder" in Abschnitt A) handelt es sich um eine Schätzung der "Prävalenz der Zahngesunden" und es gilt: $p_{ng} = 1 - p_{nk}$.

Die Summe der dmft - Werte pro Kind ist ein Maß für dessen individuelle Karieslast, d.h., n_k Kinder mit dmft > 0 weisen i.d.R. unterschiedlichen Kariesbefall auf, wohingegen eine Anzahl n_g mit jeweils dmft = 0 in der betrachteten Population keinen Kariesbefall zeigen. Dennoch wird heute in vielen Publikationen die Summe der dmft - Werte zur Berechnung des Mittelwertes dmft-MW auf alle Kinder n der Population bezogen: dmft-MW = ∑dmft / n und vermutlich deshalb gelegentlich als Kariesprävalenz bezeichnet. Es handelt sich jedoch eher um eine mittlere Schwere der Karieslast pro Kind, die bei solcher Rechnung von den Kindern mit KE auf alle Kinder der Population, auch die Gesunden, verteilt wird. Ein solcher mittlerer dmft - Wert ist wenig aussagekräftig und Informationen zur tatsächlichen Karieslast der Kinder mit KE gehen so verloren.

Zur Veranschaulichung soll noch ein Beispiel aus der Dermatologie dienen. Eine Studie mit 10339 Probanden zum Befall von Nagelpilz an den Zehennägeln ergab 1998 eine Prävalenz von etwa 12%. Bei 1285 Probanden wurde Nagelpilz festgestellt (Dt Ärztebl 2000; 97: A 1984–1986 [Heft 28–29]). In der Regel waren nicht alle Zehen befallen. Angenommen, die mittlere Zahl befallener Zehen pro Probanden mit Nagelpilz wäre in dieser Studie drei, dann wären 3 x 1285 = 3855 Zehen insgesamt befallen gewesen. Ein Bezug auf die Gesamtzahl der Studienteilnehmer ergäbe 3855 / 10339 = 0,37. Wäre die Studie repräsentativ für Deutschland, hätte somit jeder Dritte in der Bevölkerung eine Zehe mit Nagelpilz und wäre daher behandlungsbedürftig.

© 2024 M. Herzog

Für die Kariesepidemiologie scheinen daher folgende Kenngrößen nützlich:

Prävalenz der Gesunden: $p_{ng} = 1 - p_{nk} = 1 - n_k / n$

Mittlerer Kariesbefall der Kinder mit KE: dmft°-MW = $\sum$dmft / n_k

Für den Datensatz **kiga50v170.sta** ergibt sich folgendes Ergebnis für dmft°-MW

```
. drop if dmf==0
(1,436 observations deleted)

. 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 : mean dmf
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       1        Number of obs    =        876
Number of PSUs   =      50        Population size = 3,022.8061
                                  Design df        =         49
```

	Linearized			
	Mean	Std. Err.	[95% Conf.	Interval]
dmf	4.520548	.1036534	4.312249	4.728847

```
. estat effects, deff
```

	Linearized		
	Mean	Std. Err.	DEFF
dmf	4.520548	.1036534	1.03749

Interpretation: Kinder dieser Population mit KE weisen im Mittel 4,5 befallene Zähne auf.

Durch die veränderte Clustergröße reduziert sich der DEFF auf 1,04.

© 2024 M. Herzog