Zurück zum Beispiel (kiga_57.sav)

Ziel war ein Bericht über die Zahngesundheit der 3 - 5-Jährigen (siehe Beitrag SRS_Teil_2) in der gesamten Region (QP).
Aus der Stichprobe (kiga_57.sav) ergaben sich folgende Schätzungen (mit EK):

Schätzung des naturgesunden Anteils in der Studienpopulation (alle Kigä):
$p = 61{,}95\%\ (60{,}46\%\ ;\ 63{,}44\%)_{EK}\quad \pm 1{,}49\%$

Schätzung des mittleren Kariesbefalls in der Studienpopulation:
$\text{MW-dmf} = 1{,}81\ (1{,}71\ ;\ 1{,}92)_{EK}\quad \pm 0{,}10$

Die Schätzungen beziehen sich bisher auf die Studienpopulation, sind aber etwas ungenau, da sich die Anteile der 3-Jährigen zwischen SP und ST unterscheiden. Schätzungen für die Quellenpopulation QP erfolgten bisher nicht.

Eine Adjustierung auf die Quellenpopulation, erreicht man z.B. durch eine a posteriori Schichtung (im Nachhinein). Geschichtet wird hier nach Alter (3, 4, 5). In der folgenden Tabelle ist die Anzahl der Kinder in der jeweiligen Schicht abhängig von der Population QP, SP oder ST angegeben. In den Zeilen darunter findet man die relativen Schichtgrößen und schließlich die Stichprobenwerte.

A posteriori Schichtung Schätzung des Mittelwertes (MW-dmf$_{post}$)

	3	4	5	Gesamt
QP	3100	2900	3000	9000
SP	1710	3077	3191	7978
ST	482	1037	1188	2707
WQ	0,34444	0,32222	0,33333	1
WSP	0,21434	0,38569	0,39997	1
WST	0,17806	0,38308	0,43886	1
MW-dmf	1,15353	1,77049	2,12121	1,81455
Var-dmf	7,79343	11,16542	11,71402	10,9183
p	0,74481	0,62777	0,56145	0,6195
Var-p	0,19046	0,2339	0,24643	0,23581

Quellenpopulation QP

Studienpopulation SP

Stichprobe ST

N_h / N für QP

N_h / N für SP

n_h / n für ST

Stichprobenwerte

Es müssen nicht die absoluten N_h sondern nur die relativen N_h / N Schichtgrößen WQ, WSP und WST bekannt sein.

$$\overline{dmf}_{post} = \sum_{h=3}^{5} \frac{N_h}{N} \cdot \overline{dmf}_h$$

Beim Vergleich der relativen Schichtgrößen WQ und WST erkennt man, dass 3-Jährige in der Stipro deutlich unterrepräsentiert sind. Somit resultiert in der Stipro ein größerer Gesamt-MW für den dmf, da 3-Jährige den kleinsten dmf-Mittelwert haben. Dieser Effekt wirkt sich, wenn auch geringer, auch in der Studienpopulation (Grundgesamtheit) aus, da der Anteil der 3-Jährigen auch hier (WSP) nicht dem in der Quellenpopulation entspricht.
Bei dieser Hochrechnung wird vorausgesetzt, dass die Mittelwerte in den Schichten der Stichprobe mit denen in der Studienpopulation bzw. der Quellenpopulation übereinstimmen, die Stichprobe also repräsentativ ist in den Schichten.

A posteriori Schichtung Schätzung des Mittelwertes (MW-dmf_{post})

	3	4	5	Gesamt
QP	3100	2900	3000	9000
SP	1710	3077	3191	7978
ST	482	1037	1188	2707
WQ	0,34444	0,32222	0,33333	1
WSP	0,21434	0,38569	0,39997	1
WST	0,17806	0,38308	0,43886	1
MW-dmf	1,15353	1,77049	2,12121	1,81455
Var-dmf	7,79343	11,16542	11,71402	10,9183
p	0,74481	0,62777	0,56145	0,6195
Var-p	0,19046	0,2339	0,24643	0,23581

Quellenpopulation QP
Studienpopulation SP
Stichprobe ST
N_h / N für QP
N_h / N für SP
n_h / n für ST

Stichprobenwerte

$$\overline{dmf}_{post} = \sum_{h=3}^{5} \frac{N_h}{N} \cdot \overline{dmf}_h$$

Für die Stichprobe erhält man das bekannte Ergebnis 1,81. Wegen der günstigeren relativen Schichtgrößen in der Studienpopulation (Grundgesamtheit) ergibt sich ein etwas kleinerer Mittelwert von 1,78. Dieser Wert stimmt i.d.R. nicht mit dem wahren Wert der Grundgesamtheit überein, sondern ist eine weitere Schätzung aufgrund der Hochrechnung.
Die Hochrechnung für die Quellenpopulation liefert den Wert 1,67.

MW-dmf(ST) = 0,178·1,15 + 0,383·1,77 + 0,439·2,12 = 1,8146 MW Stichprobe

MW-dmf_{post}(SP) = 0,2143·1,15 + 0,3857·1,77 + 0,4·2,12 = 1,7785 MW Studienpopulation

MW-dmf_{post}(QP) = 0,344·1,15 + 0,322·1,77 + 0,333·2,12 = 1,6749 MW Quellenpopulation

A posteriori Schichtung Schätzung der Varianz des Mittelwertes

	3	4	5	Gesamt
QP	3100	2900	3000	9000
SP	1710	3077	3191	7978
ST	482	1037	1188	2707
WQ	0,34444	0,32222	0,33333	1
WSP	0,21434	0,38569	0,39997	1
WST	0,17806	0,38308	0,43886	1
MW-dmf	1,15353	1,77049	2,12121	1,81455
Var-dmf	7,79343	11,16542	11,71402	10,9183
p	0,74481	0,62777	0,56145	0,6195
Var-p	0,19046	0,2339	0,24643	0,23581

Quellenpopulation QP
Studienpopulation SP
Stichprobe ST

N_h / N für QP
N_h / N für SP
n_h / n für ST

Stichprobenwerte

Bei der a posteriori Schichtung ist die Aufteilung der Stipro auf die Schichten zufällig, denn sie wurde nicht vorher festgelegt, sondern ergibt sich erst im Nachhinein.

$$\overline{\mathrm{dmf}}_{post} \mp 1{,}96 \cdot \sqrt{s^2(\overline{x})_{post}}$$

$$s^2(\overline{x})_{post} = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \cdot \frac{s_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right)$$

Für die Berechnung der Varianz des Mittelwertes $s^2(\overline{x})_{post}$ sind die Endlichkeits-korrekturen der Altersklassen erforderlich. Der EK-Wert für die 3-Jährigen der QP lautet $(1 - n_3 / N_3) = (1 - 482/3100) = 0{,}8445$.

Berechnung der EK - Werte:

$(1 - n_3 / N_3) = 0{,}84$

$(1 - n_4 / N_4) = 0{,}64$

$(1 - n_5 / N_5) = 0{,}60$

A posteriori Schichtung Schätzung der Varianz des MW-dmf_{post}

	3	4	5	Gesamt
QP	3100	2900	3000	9000
SP	1710	3077	3191	7978
ST	482	1037	1188	2707
WQ	0,34444	0,32222	0,33333	1
WSP	0,21434	0,38569	0,39997	1
WST	0,17806	0,38308	0,43886	1
MW-dmf	1,15353	1,77049	2,12121	1,81455
Var-dmf	7,79343	11,16542	11,71402	10,9183
p	0,74481	0,62777	0,56145	0,6195
Var-p	0,19046	0,2339	0,24643	0,23581

Berechnung der EK - Werte:

$(1 - n_3 / N_3) = 0{,}84$

$(1 - n_4 / N_4) = 0{,}64$

$(1 - n_5 / N_5) = 0{,}60$

$(1 - n / N) =$ **0,81283**

Erinnerung: Die Var-dmf aller 2707 Einzelwerte in obiger Tabelle (10,9183) ohne EK ergibt sich aus den Einzelvarianzen nach Folie 6.
Für die geschätzte Varianz des geschätzten Gesamtmittels bei a post. Schichtung ergibt sich nach der Formel von Folie 4:

$s^2(\overline{x})_{post} = (0{,}344)^2 \cdot 0{,}84 \cdot 7{,}79/482 + (0{,}322)^2 \cdot 0{,}64 \cdot 11{,}17/1037 + (0{,}3333)^2 \cdot 0{,}60 \cdot 11{,}71/1188$
$= 0{,}00300$ ----> SE = 0,05477 geschätzter SE in der Quellenpopulation mit EK

$\text{C.I.}_{post} = 1{,}67489 \pm 1{,}96 \cdot 0{,}05477$ ---> (1,56754 ; 1,78224) (± 0,10735)

Resultat:

MW-dmf_{post} = 1,67 ± 0,11 (1,57 ; 1,78) im Vergleich zur Rechnung ohne a post. Schichtung
MW-dmf = 1,81 ± 0,10 (1,71 ; 1,92)
Korrektur des systematischen Fehlers (3-Jährige unterrepräsentiert), etwa gleiche Präzision.
Zur Vermeidung von Rundungsfehlern: Rechnungen in Tabelle oder Statistikprogramm.

Mittelwert und Varianz bei aggregierten Daten einer Stichprobe

Nr.	Wert		
1	4		
2	2		
3	3		
4	5		$n_1 = 8$
5	6		
6	4	MW	4.87500
7	7	Var	4.12500
8	8	SD	2.03101
9	5		
10	2		
11	4		
12	5		
13	3		$n_2 = 12$
14	3		
15	6	MW	4.75000
16	5	Var	2.93182
17	8	SD	1.71226
18	7		
19	4		
20	5		
MW_G	4.80000		
Var_G	3.22105		
SD_G	1.79473		
n_G	20		

Wenn nur Mittelwert, Standardabweichung und Fallzahl mehrerer (k) Meßreihen bekannt sind und jeweils ein Gesamtwert gewünscht wird (Gesamtmittel $\overline{x}_G$, Gesamtvarianz Var_G, Gesamtzahl n).

$$\overline{x}_G = \sum_{i=1}^{k} \frac{n_i}{n} \cdot \overline{x}_i \qquad n = \sum_{i=1}^{k} n_i$$

Varianzzerlegung: interne Varianz externe Varianz

$$Var_G = s_G^2 = \frac{1}{n-1} \cdot \left[\sum_{i=1}^{k} (n_i - 1) \cdot s_i^2 + \sum_{i=1}^{k} n_i \cdot (\overline{x}_i - \overline{x}_G)^2 \right]$$

Beispiel: $\overline{x}_G$ = 8/20·4,875 + 12/20·4,75 = 4,8

Var_G = 1/19·[(7·4,125 + 11·2,93182) + (8·0,075² + 12·0,05²)

= 3,22105

Für Grundgesamtheiten gilt:

$$\sigma^2 = \frac{1}{N} \cdot \left[\sum N_i \cdot \sigma_i^2 + \sum N_i \cdot (\mu_i - \mu)^2 \right]$$