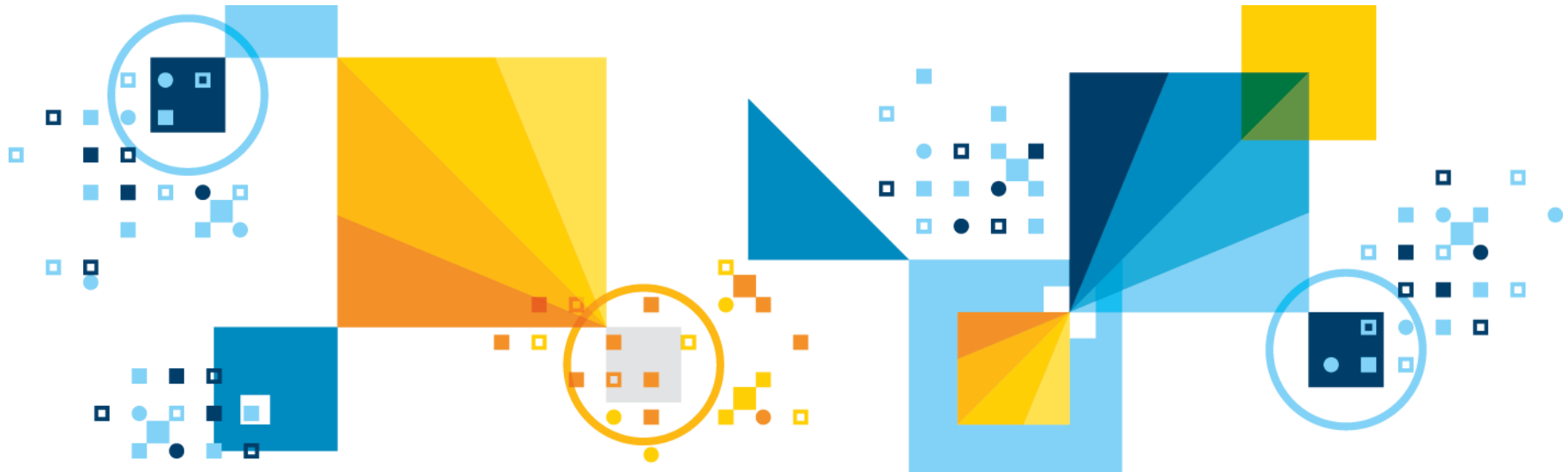


Jon K Peck

Rim Weighting with IBM® SPSS® Statistics: Theory and Practice



Agenda

- Why we weight
- Raking weights with the SPSSINC RAKE extension command
- Practical issues with raking
- Using rake-weighted data in statistical procedures
- Summary

PURPOSE OF WEIGHTING

Purposes for weighting - I

- Making a sample representative of a population for analysis
 - Estimating population quantities, e.g., counts, percentages, means
 - Causal modeling, e.g., regression
 - Correcting sample variable distributions (categorical weighting)
 - Correcting sample variable distributions (scale weighting)
 - Adjusting for randomness, sampling schemes and post stratification
 - Deliberate over- or under- sampling
 - Stratified, clustered, multi-stage, etc
 - Requires different algorithms as well as weights when calculating variances
 - Random nonresponse adjustments
 - Samples representing wrong population (noncoverage or overcoverage)
 - convenience samples
 - Reducing selection bias

Purposes for weighting - II

- Statistically correcting modeling assumptions
 - E. g. Heteroscedasticity of regression error terms
- Assigning importance weights to cases
 - E. g., influence case weighting in TREES
 - Overweight cases of particular interest, e.g., maximize retention probability of most profitable customers
- For computational efficiency
 - A case represents n identical cases (replication weight)
- NOT good for
 - Standard weighting does not address measurement or coding errors
 - E. g., underreporting income from capital gains
 - Missing values for respondents (item nonresponse)
 - Imputation techniques address missing values
 - Can adjust control totals for missing values
 - Modeling may require more complex weighting schemes

Reasons not to weight

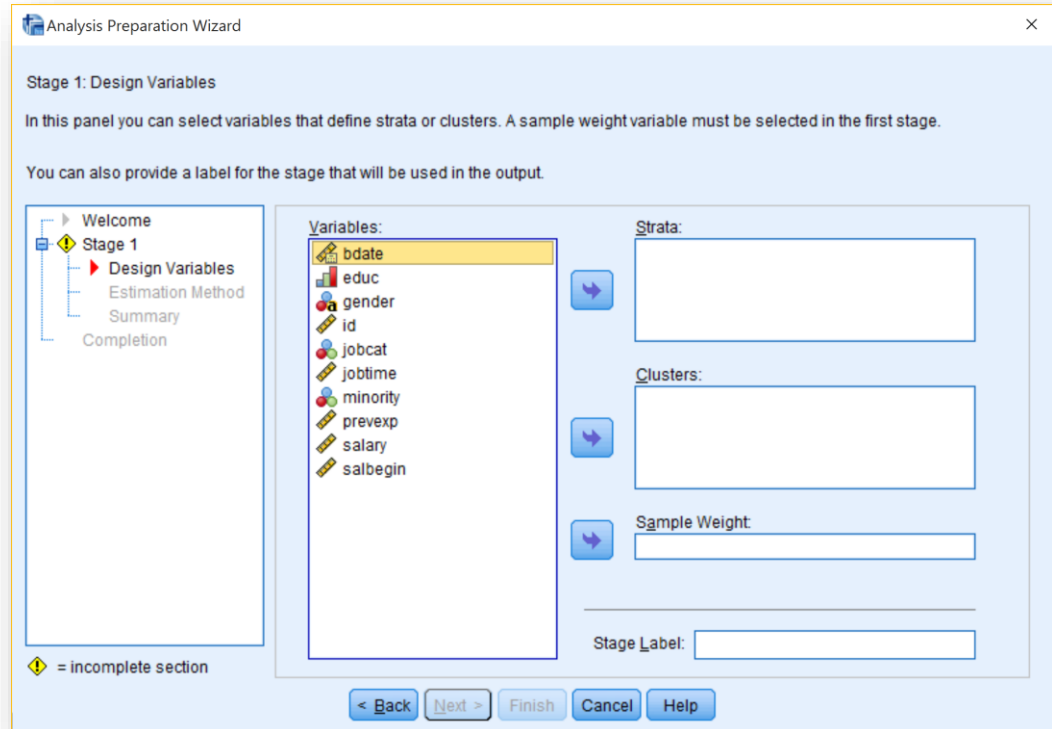
- Complexity
- Models may already correct for selection bias
 - Tobit regression
 - Heckman regression
 - Survival problems
 - Cox regression
- Increased variance with large weights
 - Unstable estimates
 - Bias vs variance tradeoffs
- “Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratios, and standard errors are tricky even with simple weighted means....Contrary to what is assumed by many theoretical statisticians, survey weights are not in general equal to inverse probabilities of selection but are typically constructed based on a combination of probability calculations and nonresponse adjustments.”
 - Andrew Gelman, Struggles with Survey Weighting and Regression Modeling, *Statistical Science*, 2007

There are several types of weights

- Types of weights
 - Frequency or replication weights
 - Probability of selection weights
 - Sample adjustment weights
 - Importance weights
 - Modeling weights
- All of these apply to some procedures in Statistics
- It is important to understand how a particular procedure interprets weights
- The implications are different depending on the statistic

Complex sample designs require special procedures

- Complex Samples option procedures supports multistage, clustered or unequally sampled designs of many types
- If designing a survey
 - *Analyze > Complex Samples > Select a Sample*
- If using an already constructed sample
 - *Analyze > Complex Samples > Prepare for Analysis*
- Creates a plan file describing design
- Apply appropriate CS procedure using the plan file (*Analyze > Complex Samples > ...*)
 - Includes frequencies, descriptives, crosstabs, linear models, logistic regression and others
- CS procedures are not designed for poststratification weighting




The screenshot shows the 'Analysis Preparation Wizard' window, specifically 'Stage 1: Design Variables'. The window has a title bar with a close button. Below the title bar, the text reads: 'Stage 1: Design Variables' and 'In this panel you can select variables that define strata or clusters. A sample weight variable must be selected in the first stage.' Below this, it says: 'You can also provide a label for the stage that will be used in the output.'

On the left, there is a tree view with the following items: 'Welcome', 'Stage 1' (selected), 'Design Variables' (sub-item of Stage 1), 'Estimation Method', 'Summary', and 'Completion'. A yellow warning icon is next to 'Stage 1'. Below the tree view, a legend indicates that a yellow triangle icon means '= incomplete section'.

The main area is divided into two sections. The left section is titled 'Variables:' and contains a list of variables: 'bdate', 'educ', 'gender', 'id', 'jobcat', 'jobtime', 'minority', 'prevexp', 'salary', and 'salbegin'. The 'bdate' variable is selected and highlighted in yellow. To the right of this list are three empty text boxes labeled 'Strata:', 'Clusters:', and 'Sample Weight:', each with a blue arrow button pointing to it. Below these boxes is a 'Stage Label:' text box. At the bottom of the window are five buttons: '< Back', 'Next >', 'Finish', 'Cancel', and 'Help'.

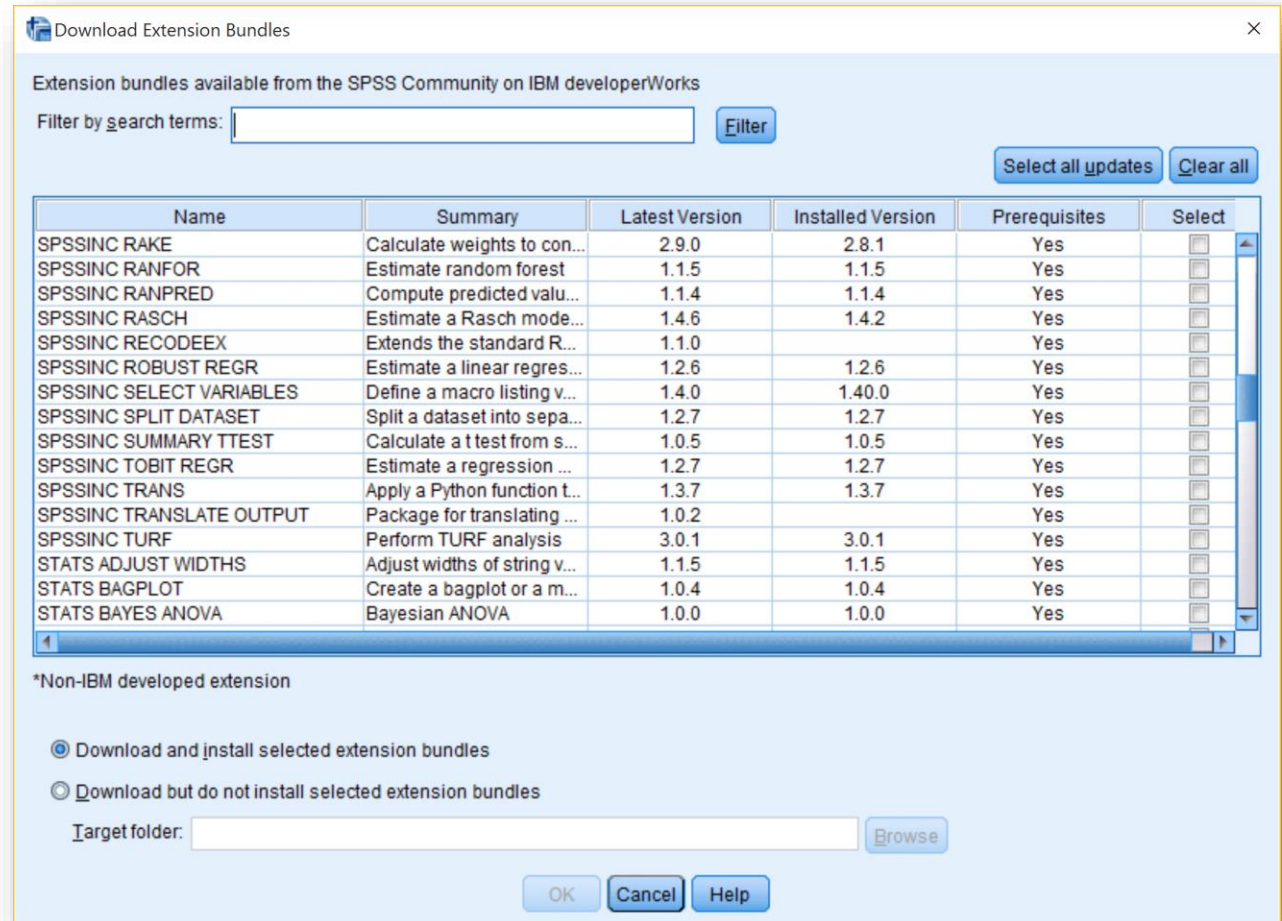
THE SPSSINC RAKE EXTENSION COMMAND

Extension commands add functionality to IBM SPSS Statistics

- Provide standard Statistics syntax for Python, R, or Java programs
 - Most IBM-created extension command names start with SPSSINC or STATS
 - Recognize by menu icon  Rake Weights...
- Work like built-in commands
- Usually include a dialog box built with the Custom Dialog Builder
- By convention, *COMMAND-NAME /HELP* displays the syntax help (F1 in V23+)
- Require Python and/or R Essentials as appropriate
- Install from *Utilities* or *Extensions* menu
 - Over 110 are available and the number continues to grow
 - Most not tied to a particular Statistics version
 - Free
- Consider using environment variables to specify where to install
 - May need to install using Run As Administrator
 - Sharing
 - Updating to new Statistics versions
 - My settings
 - SPSS_CDIALOGS_PATH=C:\dlgcommon
 - SPSS_EXTENSIONS_PATH=C:\extcommon; C:\extcommon18
 - On Windows, environment variables are set via Control Panel > System > ...

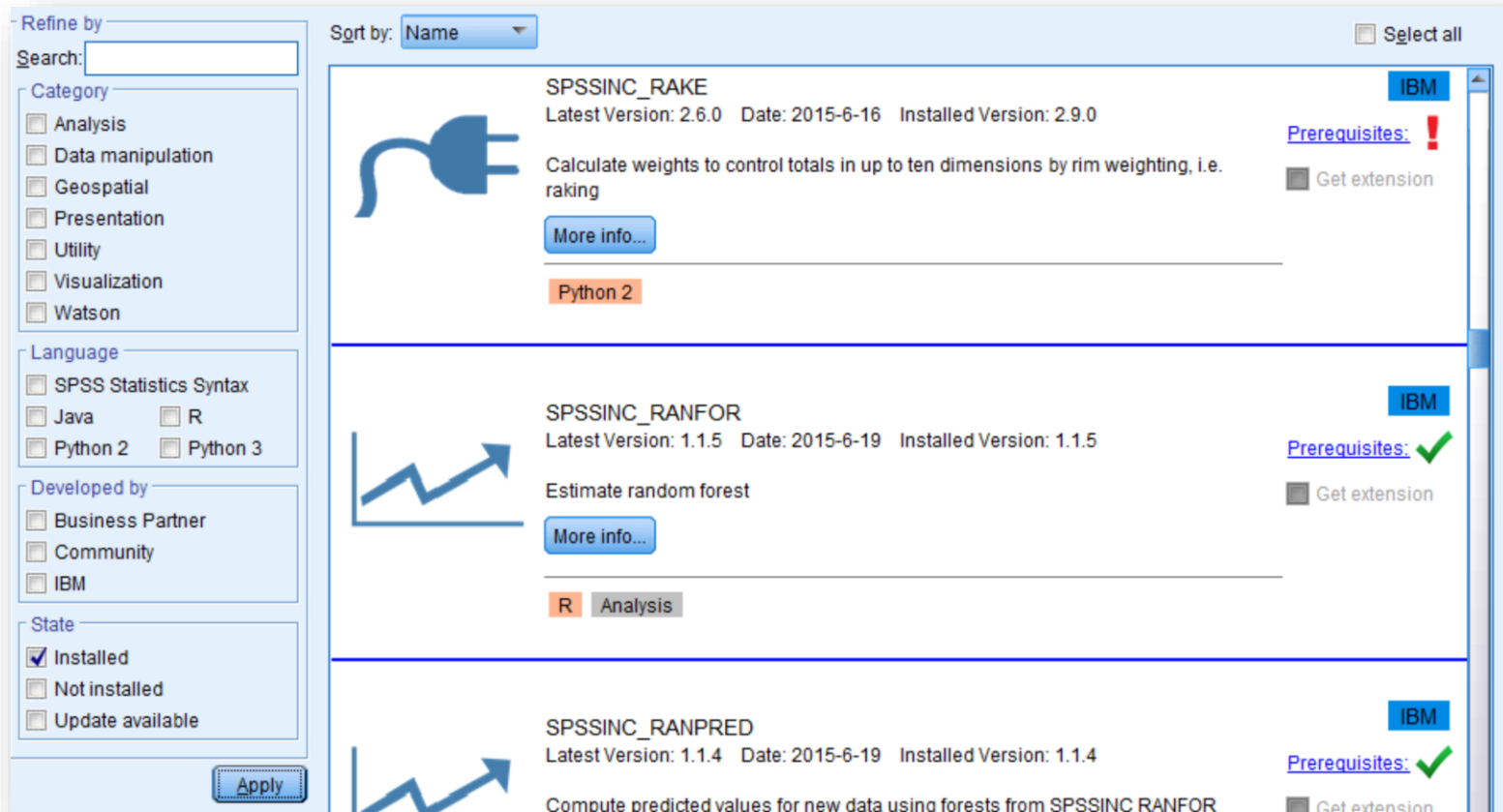
In Statistics 22-23 *Utilities > Extension Bundles > Download and Install Extension Bundles*

- May require admin permissions
- Start Statistics using Run As Administrator (Windows) if possible



In Statistics 24+ *Extensions > Extension Hub*

- Also available directly from website



The screenshot displays the IBM Extension Hub interface. On the left, there is a sidebar with filters for refining search results. The main area shows a list of extensions, sorted by Name. Each extension entry includes an icon, the extension name, version information, a description, a 'More info...' button, and a 'Get extension' button. The extensions are categorized by language and developed by IBM.

Refine by

Search:

Category

- ☐ Analysis
- ☐ Data manipulation
- ☐ Geospatial
- ☐ Presentation
- ☐ Utility
- ☐ Visualization
- ☐ Watson

Language

- ☐ SPSS Statistics Syntax
- ☐ Java
- ☐ R
- ☐ Python 2
- ☐ Python 3

Developed by

- ☐ Business Partner
- ☐ Community
- ☐ IBM

State

- ☒ Installed
- ☐ Not installed
- ☐ Update available

Sort by: Name

☐ Select all

Extension Name	Latest Version	Date	Installed Version	Prerequisites	Get extension
SPSSINC_RAKE	2.6.0	2015-6-16	2.9.0	Prerequisites: !	<input type="checkbox"/>
SPSSINC_RANFOR	1.1.5	2015-6-19	1.1.5	Prerequisites: ✓	<input type="checkbox"/>
SPSSINC_RANPRED	1.1.4	2015-6-19	1.1.4	Prerequisites: ✓	<input type="checkbox"/>

The appropriate weighting technique depends on the information you have

- Assume a 2-dimensional set of cells defined by crossing categorical variables A and B with categories A_i and B_j . The cells are C_{ij} . This extends to any number of dimensions.
- If you know the joint distribution in the target population of A and B
 - Weight each cell, C_{ij} as $\frac{N_{ij}}{n_{ij}} * K$
where N's are population counts and n's are sample counts, and K normalizes the weight to sum to the actual sample size
 - Number of cells grows very rapidly – demands a lot of information
 - Small cell counts vulnerable to sampling variability
- If you know only the marginal distributions
 - Consider raking aka rim weighting (multi-dimensional poststratification) aka iterative proportional fitting
 - Assumes probabilities and weights are function of marginals only. No interaction
- If you have partial information about the joint distribution
 - Convert joint distribution to one dimension and rake with that and the remaining marginals
 - May need to further collapse categories

There are other approaches to weighting besides raking

▪ Inverse Propensity Score

- Estimate the propensity/probability of being in the sample using sample and population data
- How to estimate propensity?
 - E.g., regress cell counts on population counts in k dimensions
 - Regression on totals (cells)

$$\log(n_{ij}) = \beta_1 X_{1i} + \beta_2 X_{2j} + \dots$$

- Weight by inverse of propensity
- Need to bound weights as $p \rightarrow 0$
- Scale sum of weights to actual sample size


▪ Maximum entropy

- Use the known marginals and other information and make weights that are consistent with the information but are otherwise as uniform as possible (maximize entropy)
 - Requires solving nonlinear constrained maximization problem
- If information is only marginal distributions, converges to raking
- If information is joint distribution, converges to joint probability weighting

Raking adjusts cell counts so that marginal totals match control totals

Originally due to W. Edwards Deming and Frederick F. Stephan,
On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, Annals of Mathematical Statistics, Vol 11, Number 4, 1940

"There are situations in sampling wherein the data furnished by the sample must be adjusted for consistency with data obtained from other sources or with deductions from established theory."



$$\begin{aligned} (1) \quad \widehat{N^{(1)}}_{ij} &= \frac{n_{ij} N_{i+}}{n_{i+}} && \text{(row)} \\ (2) \quad \widehat{N^{(2)}}_{ij} &= \widehat{N^{(1)}}_{ij} N_{+j} / \widehat{N^{(1)}}_{+j} && \text{(column)} \end{aligned}$$

where + indicates summation over that index:

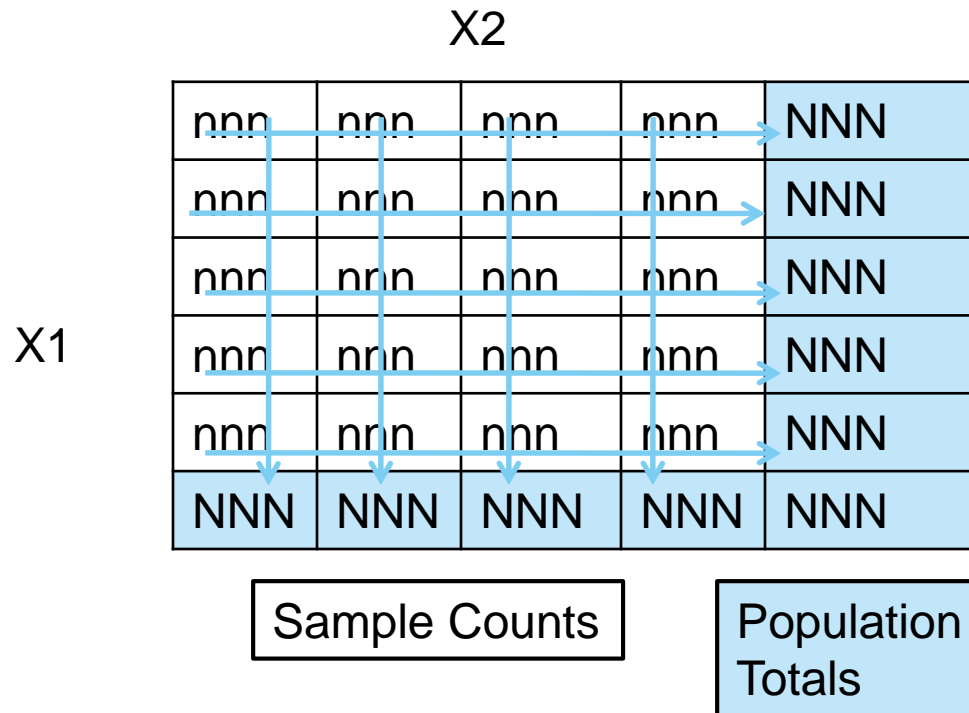
$$N_{i+} = \sum_j N_{ij} \text{ etc}$$

Lowercase symbols refer to the sample

Uppercase symbols refer to the population

Superscripts refer to the iteration

The iterations repeat in a raking pattern in 1 to M dimensions



- Iteration is repeated until convergence is achieved

Raking is equivalent to fitting a loglinear model

Loglinear model:

$$N_{ij} = a_i b_j n_{ij}$$

written as probabilities, $\pi_{ij} = a_i b_j p_{ij}$ where

π_{ij} and p_{ij} are population and sample probabilities, respectively)

$$\log\left(\frac{\pi_{ij}}{p_{ij}}\right) = \log(a_i) + \log(b_j) + \varepsilon_{ij}$$

hence SPSSINC RAKE uses GENLOG procedure to fit the model

- Loglinear models are fit by Iterative Proportional Fitting (IPF)
- Observed counts are assumed to be independent Poisson variables
- Fit by MLE using Newton-Raphson algorithm
- You can expose details of GENLOG portion of output in SPSSINC RAKE if necessary and adjust fitting parameters
 - Don't do this for really big problems!

Sample balance measured by the *rim weighting efficiency*

rim weighting efficiency

- If no input weights, this reduces to

W_i = input weight for case i

R_i = rim weight for case i

$$E = 100 * \frac{(\sum_i R_i)^2}{N \sum_i R_i^2}$$

$$E = 100 * \frac{(\sum_i W_i R_i)^2}{\sum_i W_i \sum_i W_i R_i^2}$$

- N is the total number of cases
- The square of the sum of the weights / ($N * \text{sum of squared weights}$)
- Equals 100 if sample is perfectly balanced
- Defines effective base
- Compares sample with a perfect random sample assuming that the optimal sample is proportionate
- If subgroup variances differ, this might not be optimum
- There is no magic cutoff value

$\text{cumsum}(R) = N$
if weights are
normalized

Example: 1 male, 10 females balanced to 50/50

	numericGender	weight
1	2	5.50
2	1	.55
3	1	.55
4	1	.55
5	1	.55
6	1	.55
7	1	.55
8	1	.55
9	1	.55
10	1	.55
11	1	.55

Sample Balance Based on Variables: numericGender

Balance

Sample Balance	33.058
----------------	--------

Raked Weights

numericGender	Category Rake Weight	Unweighted Case Count
1.0	.550	10.000
2.0	5.500	1.000

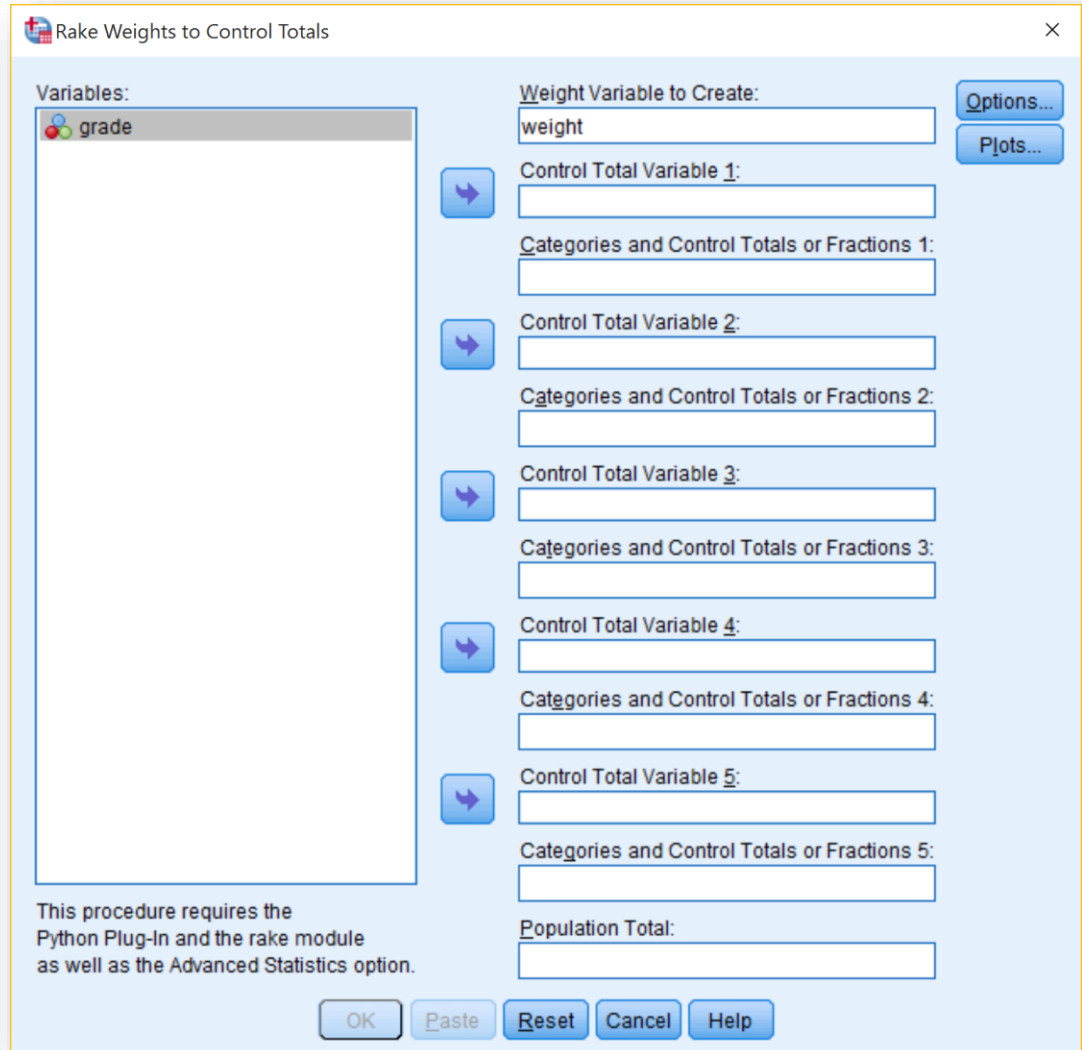
numerator	$(5.5 + 10 * 0.55)^2$	121.000
denominator	$11 * (5.5^2 + 10 * 0.55^2)$	366.025
E		0.331

- Would you trust this sample?
- Consider estimating a mean (Std error of mean depends on variance of sum)

$$Var(w_1X_1 + w_2X_2 + \dots) = (w_1^2 + w_2^2 + \dots)Var(X) = w^2 * (1 + 1 + \dots)Var(x)$$

Rim weighting (raking) is done using the SPSSINC RAKE extension command

- *Data > Rake Weights* or SPSSINC RAKE procedure
- This presentation is based on version 2.9.0 of the extension
- Some functionality of the procedure is not included in the dialog box
- SPSSINC RAKE is an extension command



Rake Weights to Control Totals

Variables:

grade

Weight Variable to Create:
weight

Options...
Plots...

Control Total Variable 1:
[]

Categories and Control Totals or Fractions 1:
[]

Control Total Variable 2:
[]

Categories and Control Totals or Fractions 2:
[]

Control Total Variable 3:
[]

Categories and Control Totals or Fractions 3:
[]

Control Total Variable 4:
[]

Categories and Control Totals or Fractions 4:
[]

Control Total Variable 5:
[]

Categories and Control Totals or Fractions 5:
[]

Population Total:
[]

This procedure requires the Python Plug-In and the rake module as well as the Advanced Statistics option.

OK Paste Reset Cancel Help

SPSSINC RAKE is the extension command for raking

- Requires the Advanced Statistics option because it uses GENLOG
- SPSSINC RAKE /HELP displays the syntax help in the Viewer
 - Extension commands are not included in the Command Syntax Reference
 - Might need to update extension.py file from website
- Variable names in extension commands are CaSe SeNsItIvE
- Dialog box appears on the Data menu

+ SPSSINC RAKE Extension Command

Calculate case weights so that weighted counts match control totals in up to ten dimensions

Syntax:

```
SPSSINC RAKE DIM1=control variable* control categories and values*
DIM2 ... DIM10
FINALWEIGHT=varname POPTOTAL=value
```

```
/DS1 DS=dataset name* CATVAR=category variable* TOTVAR=marginal totals variable*
/DS2 ... /DS10
```

```
/OPTIONS DELTA=value ITERATIONS=value CONVERGENCE=value CHECKEMPTY=YES** or NO
SHOW=YES or NO**
SHOWWEIGHTS=YES** or NO
```

```
PLOT HISTOGRAM=YES** or NO
YVAR=varname XVAR=varname PANELDOWNVAR=varname PANELACROSS=varname
or AUTOHEATMAP = number
```

```
/HELP
```

* Required

** Default

SPSSINC RAKE /HELP. prints this information and does nothing else.

Example:

```
SPSSINC RAKE
DIM1=jobcat 1 500 2 300 3 200
DIM2=minority 1 800 2 200
FINALWEIGHT=weight.
/PLOT YVAR=jobcat XVAR=minority.
```

Example:

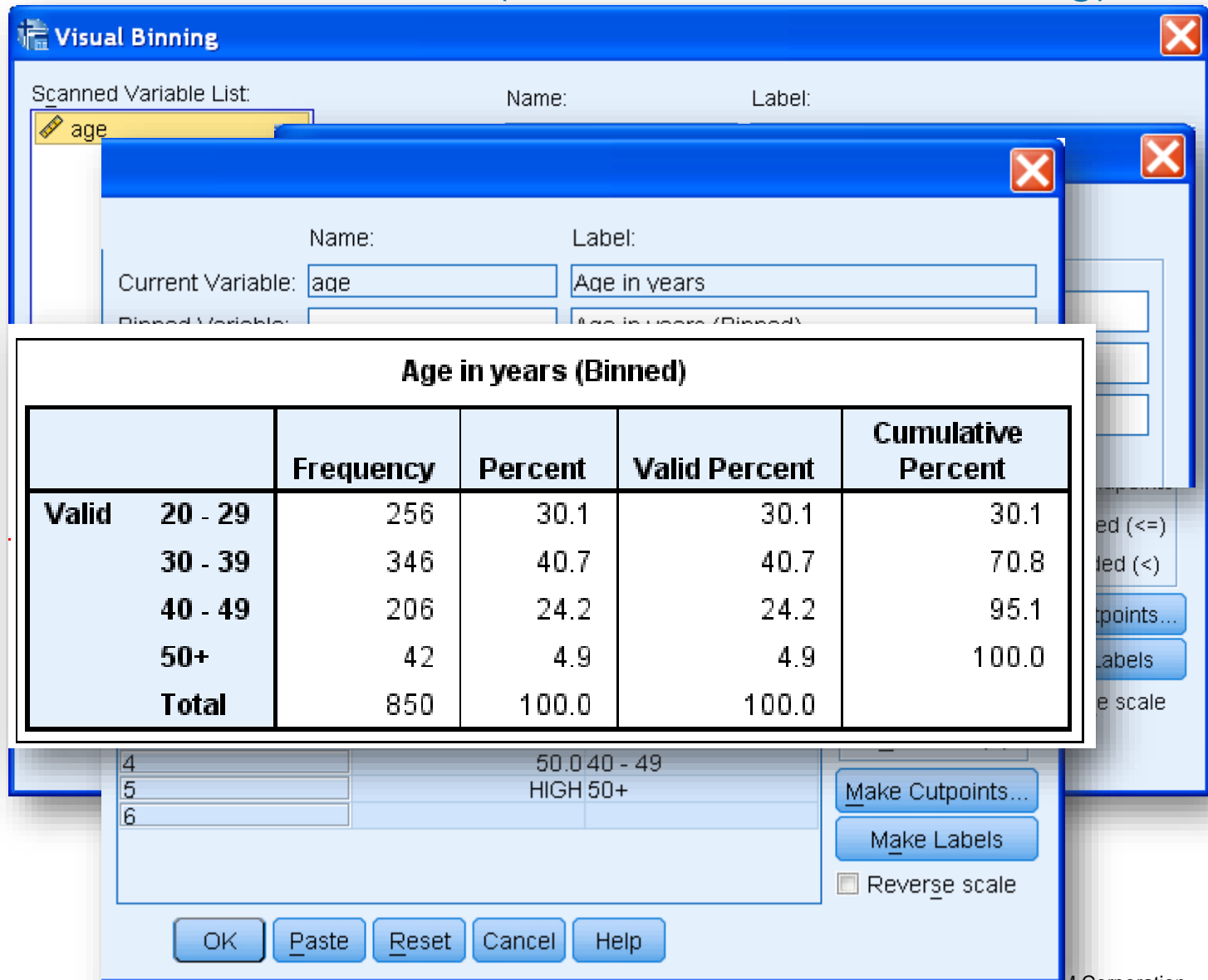
Raking example: one dimension: *bankloan.sav* (in .../samples/english)

- Adjust by age
- Dataset has 850 cases
- Age reported by year
- Suppose we have known percentages by decade of age
- Recode age using Visual Binner
 - *Transform > Visual Binning*

Age in years					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	2	.2	.2	.2
	21	12	1.4	1.4	1.6
	22	14	1.6	1.6	3.3
	23	21	2.5	2.5	5.8
	24	30	3.5	3.5	9.3
	25	25	2.9	2.9	12.2
	26	30	3.5	3.5	15.8
	27	33	3.9	3.9	19.6
	28	38	4.5	4.5	24.1
	29	51	6.0	6.0	30.1
	30	30	3.5	3.5	33.6
	31	42	4.9	4.9	38.6
	32	30	3.5	3.5	42.1
	33	31	3.6	3.6	45.8
	34	38	4.5	4.5	50.2
	35	40	4.7	4.7	54.9
	36	33	3.9	3.9	58.8
	37	31	3.6	3.6	62.5
	38	30	3.5	3.5	66.0
	39	41	4.8	4.8	70.8
	40	32	3.8	3.8	74.6
	41	36	4.2	4.2	78.8
	42	16	1.9	1.9	80.7
	43	22	2.6	2.6	83.3
	44	15	1.8	1.8	85.1
	45	21	2.5	2.5	87.5

Recoding age with the Visual Binner (*Transform > Visual Binning*)

- *Make Cutpoints* can find the breaks
- Binner can do automatic labels
- Generates RECODE and metadata syntax.



Visual Binning

Scanned Variable List: **age** Name: Label:

Current Variable: **age** Label: **Age in years**

Age in years (Binned)

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 20 - 29	256	30.1	30.1	30.1
30 - 39	346	40.7	40.7	70.8
40 - 49	206	24.2	24.2	95.1
50+	42	4.9	4.9	100.0
Total	850	100.0	100.0	

4 50.0 40 - 49
5 HIGH 50+
6

Make Cutpoints...
Make Labels
☐ Reverse scale

OK Paste Reset Cancel Help

Rake to adjust proportions

unweighted

Age Group	Age Code	Control Percent age
< 20	1	0
20-29	2	40
30-39	3	40
40-49	4	20
50+	5	10

Age in years (Binned)					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20 - 29	309	36.4	36.4	36.4
	30 - 39	309	36.4	36.4	72.7
	40 - 49	155	18.2	18.2	90.9
	50+	77	9.1	9.1	100.0
	Total	850	100.0	100.0	

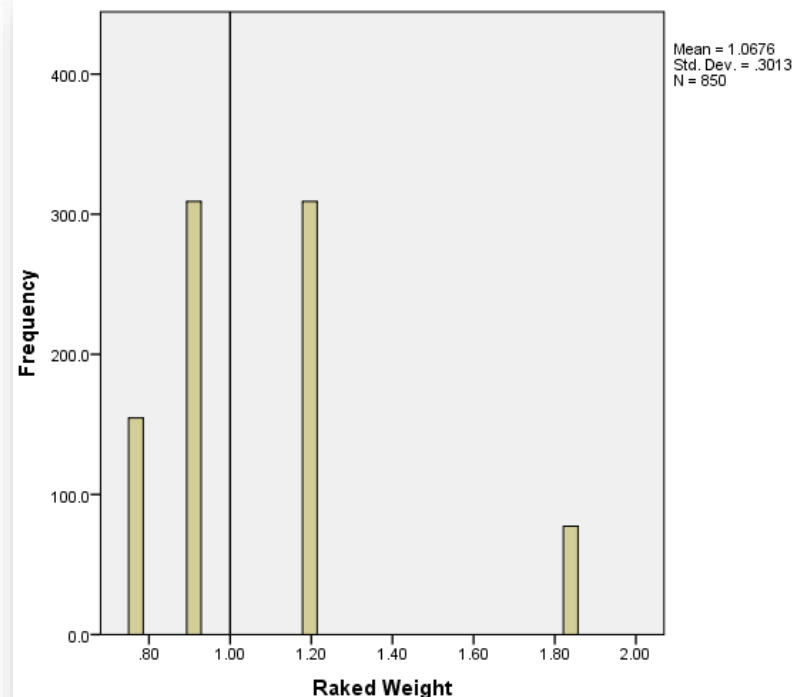
Age in years (Binned)	Percent
< 20	30.1
20-29	40.7
30-39	24.2
40-49	4.9
50+	100.0

Percentages will be normalized
(these could be counts)

Case count unchanged

You could do this without fancy
technology!

5 equations in 5 unknowns



Raking example: two dimensions (*schoolsurvey.sav*)

- High school survey
 - simulated but realistic data
 - 4 equal grade sizes
 - N of males = N of females
 - 2000 respondents
- Marginal control totals are known but *not* the full joint distribution
- Response rates vary by gender and grade
 - Gender
 - 60 % female
 - 50 % males
 - Grade
 - 90 % freshman
 - 80 % soph
 - 70 % junior
 - 60 % senior
- No response rate interaction between gender and grade

gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1 Female	1199	60.0	60.0	60.0
2 Male	801	40.1	40.1	100.0
Total	2000	100.0	100.0	

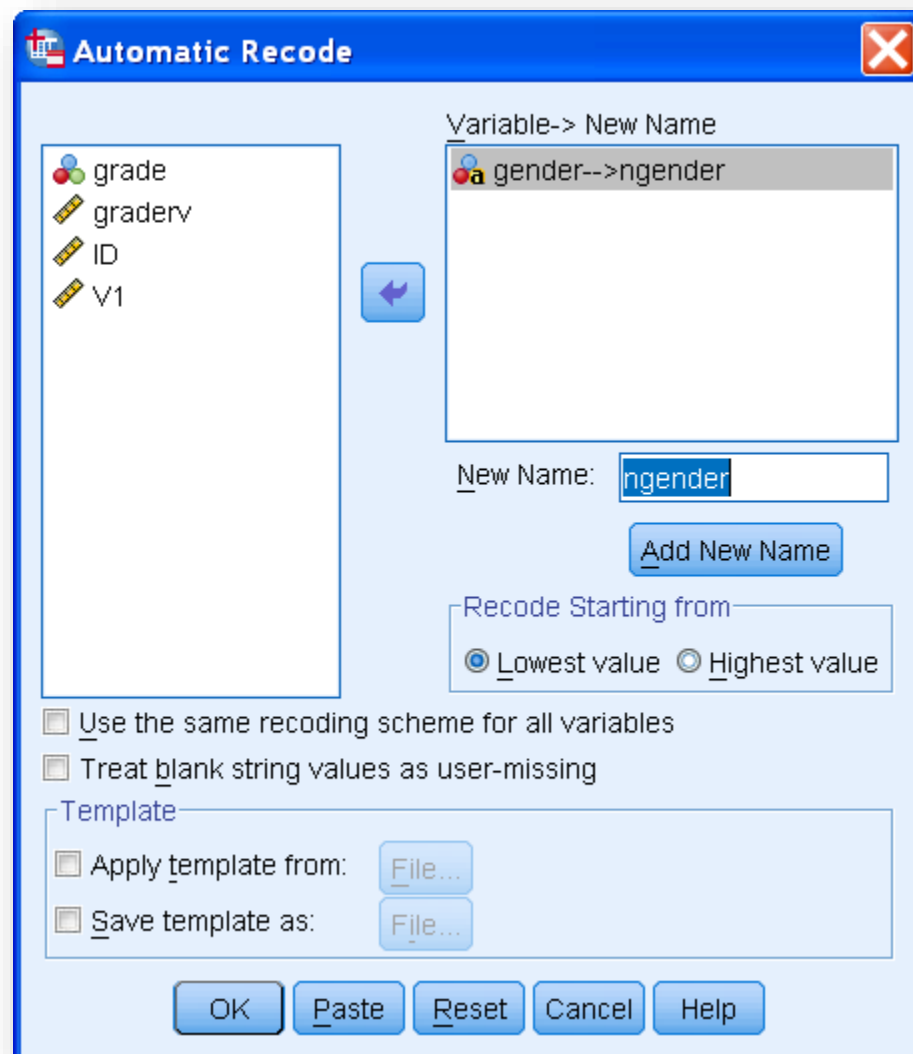
grade

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 9	603	30.2	30.2	30.2
10	518	25.9	25.9	56.1
11	468	23.4	23.4	79.5
12	411	20.6	20.6	100.0
Total	2000	100.0	100.0	

school size = 3636

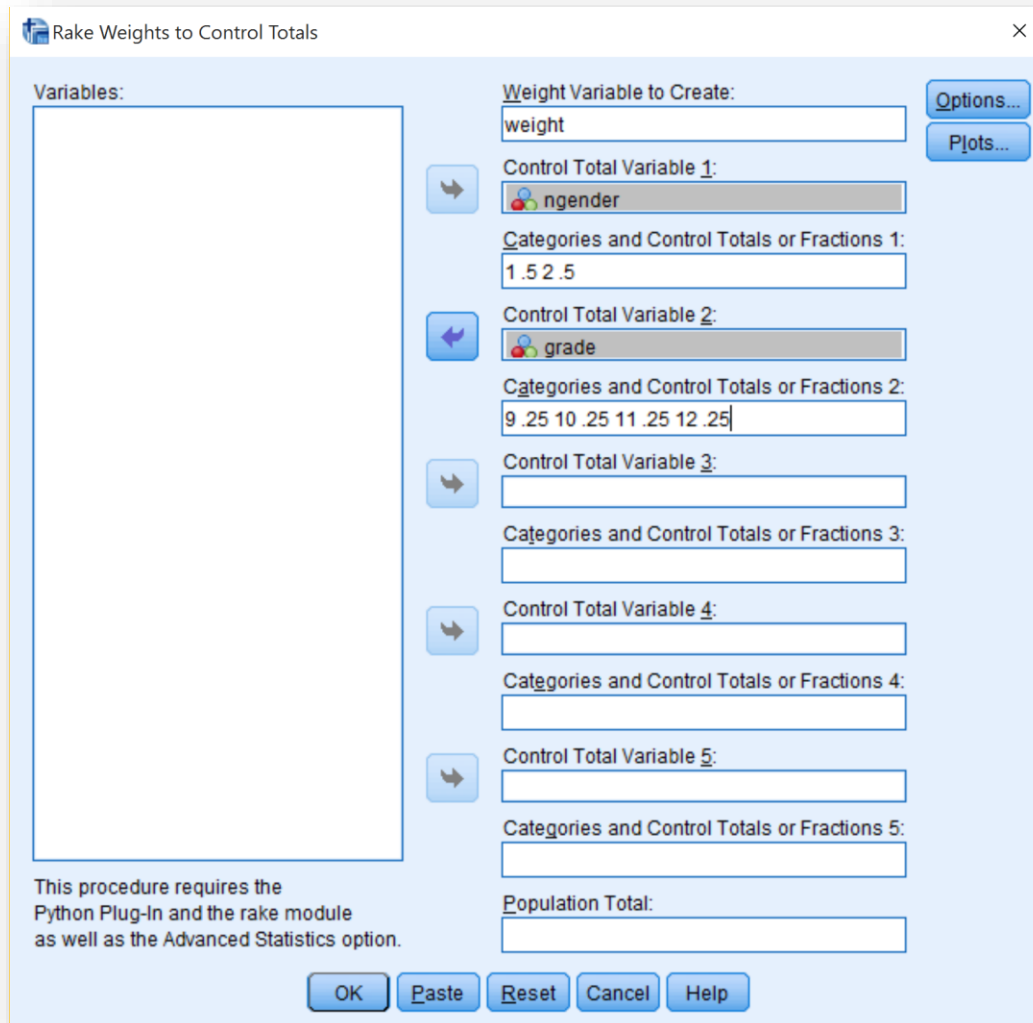
AUTORECODE converts gender to a numeric variable

- RAKE does not accept string variables
- *Transform > Automatic Recode...*
- Generates appropriate value labels



Raking adjusts for the unit nonresponse

- *Data > Rake Weights...*
- Name a weight variable (must not already exist)
- For each control variable, enter pairs of category value and proportion or percentage
- Need not add to 100
- Omitted categories get weight of SYSMIS
- Only numeric variables can be used
- Optionally scale to a fixed population total
- Generates SPSSINC RAKE syntax
 - Up to 10 dimensions in syntax
- SPSSINC RAKE /HELP shows syntax chart



Rake Weights to Control Totals

Variables:

Weight Variable to Create: weight

Control Total Variable 1: ngender

Categories and Control Totals or Fractions 1: 1 .5 2 .5

Control Total Variable 2: grade

Categories and Control Totals or Fractions 2: 9 .25 10 .25 11 .25 12 .25

Control Total Variable 3:

Categories and Control Totals or Fractions 3:

Control Total Variable 4:

Categories and Control Totals or Fractions 4:

Control Total Variable 5:

Categories and Control Totals or Fractions 5:

Population Total:

This procedure requires the Python Plug-In and the rake module as well as the Advanced Statistics option.

OK Paste Reset Cancel Help

```
SPSSINC RAKE DIM1 = gender 1 .5 2 .5
DIM2=grade 9 .25 10 .25 11 .25 12 .25
FINALWEIGHT=weight.
```

Raking results in exact adjustment

- RAKE automatically turns on the calculated weight
- Six equations (constraints) in 8 unknowns
- In general:
 - $N1 + N2$ equations in $N1 * N2$ unknowns

		gender			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Female	1000	50.0	50.0	50.0
	2 Male	1000	50.0	50.0	100.0
	Total	2000	100.0	100.0	

		grade			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	9	500	25.0	25.0	25.0
	10	500	25.0	25.0	50.0
	11	500	25.0	25.0	75.0
	12	500	25.0	25.0	100.0
	Total	2000	100.0	100.0	

Histogram shows the weight distribution

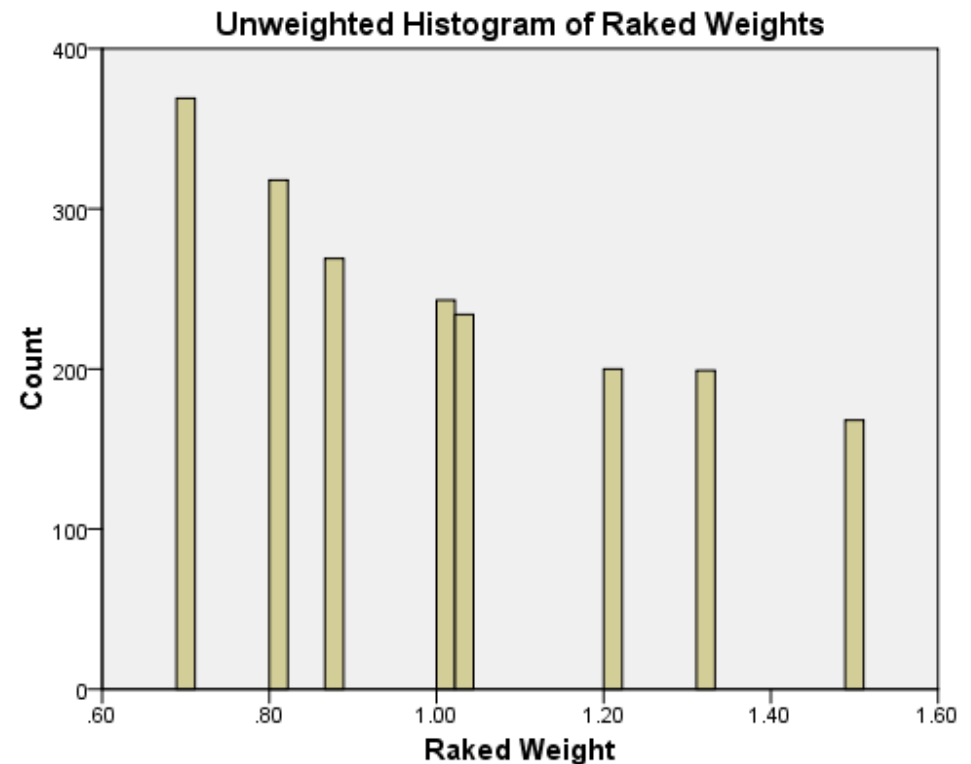
- There are only 8 distinct weights
- Sample balance is good

Sample Balance Based on Variables: ngender, grade

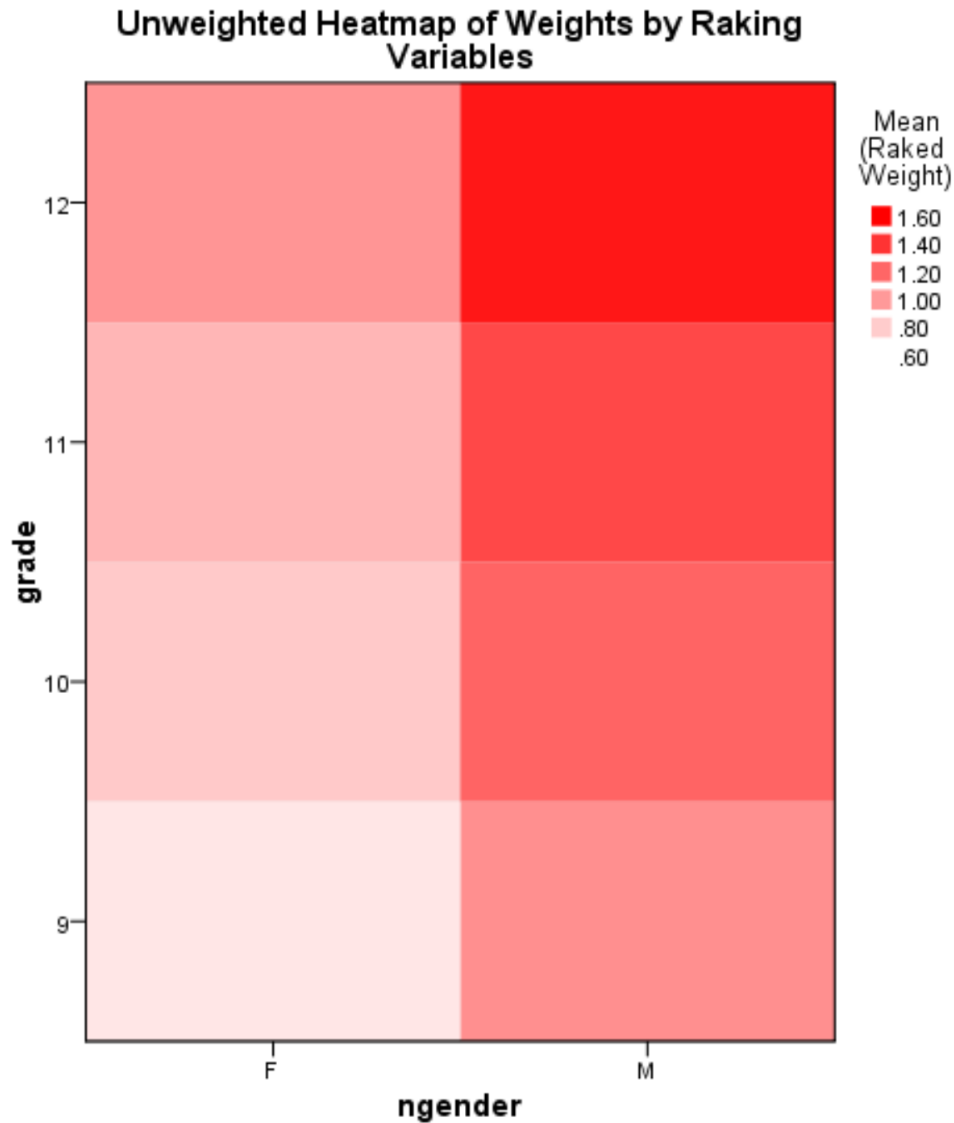
	Balance
Sample Balance	94.286

Raked Weights

ngender, grade	Category Rake Weight	Unweighted Case Count
1.0, 9.0	.697	369.000
1.0, 10.0	.812	318.000
1.0, 11.0	.885	269.000
1.0, 12.0	1.014	243.000
2.0, 9.0	1.037	234.000
2.0, 10.0	1.209	200.000
2.0, 11.0	1.317	199.000
2.0, 12.0	1.509	168.000



It's easy to see the pattern of the weights in the heatmap



Exercise 1: Let's do one

- Using the schools dataset, *schoolsurvey.sav*, or, better, a dataset of your own, run a raking adjustment in two or more dimensions
- Make up your own control totals
- If your dataset already has a weight variable, set it on before raking
- Note that SPSSINC RAKE automatically turns on the output weight variable
- Run RAKE a second time with the weight on naming a new output weight variable but same control totals
- Compare the two weight variables

Unweighted and weighted crosstabs – its easy to confirm the weights

grade * gender Crosstabulation

			gender		Total
			F	M	
grade	9	Count	369	234	603
		% within grade	61.2%	38.8%	100.0%
		% within gender	30.8%	29.2%	30.2%
	10	Count	318	200	518
		% within grade	61.4%	38.6%	100.0%
		% within gender	26.5%	25.0%	25.9%
	11	Count	269	199	468
		% within grade	57.5%	42.5%	100.0%
		% within gender	22.4%	24.8%	23.4%
	12	Count	243	168	411
		% within grade	59.1%	40.9%	100.0%
		% within gender	20.3%	21.0%	20.5%
Total	Count	1,199	801	2,000	
	% within grade	60.0%	40.1%	100.0%	
	% within gender	100.0%	100.0%	100.0%	

Unweighted

grade * ngender Crosstabulation

			ngender		Total
			F	M	
grade	9	Count	257	243	500
		% within grade	51.4%	48.6%	100.0%
		% within ngender	25.7%	24.3%	25.0%
	10	Count	258	242	500
		% within grade	51.6%	48.4%	100.0%
		% within ngender	25.8%	24.2%	25.0%
	11	Count	238	262	500
		% within grade	47.6%	52.4%	100.0%
		% within ngender	23.8%	26.2%	25.0%
	12	Count	246	254	500
		% within grade	49.2%	50.8%	100.0%
		% within ngender	24.6%	25.4%	25.0%
Total	Count	999	1,001	2,000	
	% within grade	50.0%	50.0%	100.0%	
	% within ngender	100.0%	100.0%	100.0%	

Weighted

- By default, CROSSTABS counts are rounded

Joint distribution is different after weighting (Spearman correlations)

Correlations				
			ngender	grade
Spearman's rho	ngender	Correlation Coefficient	1.000	.024
		Sig. (2-tailed)	.	.285
		N	2,000	2,000
	grade	Correlation Coefficient	.024	1.000
		Sig. (2-tailed)	.285	.
		N	2,000	2,000

Unweighted

			ngender	grade
Correlation Coefficient	Sig. (2-tailed)	N	1.000	.143**
			.	.000
			2,168	2,168
Correlation Coefficient	Sig. (2-tailed)	N	.143**	1.000
			.000	.
			2,168	2,168

**. Correlation is significant at the 0.01 level (2-tailed).

Weighted

- The total sample size is different in the two tables
- **NONPAR CORR in Help > Algorithms**
- If a WEIGHT variable is specified, it is used to replicate a case as many times as indicated by the weight value rounded to the nearest integer. If the workspace requirements are exceeded and sampling has been selected, a random sample of cases is chosen for analysis using the algorithm described in SAMPLE.

The weight distribution shows only eight distinct values

- Weighted case count = unweighted count
- This is important when degrees of freedom matter
- $IQR = 1.2086 - .8122 = .396$
- Max weight = median wt + $1.25 * IQR$

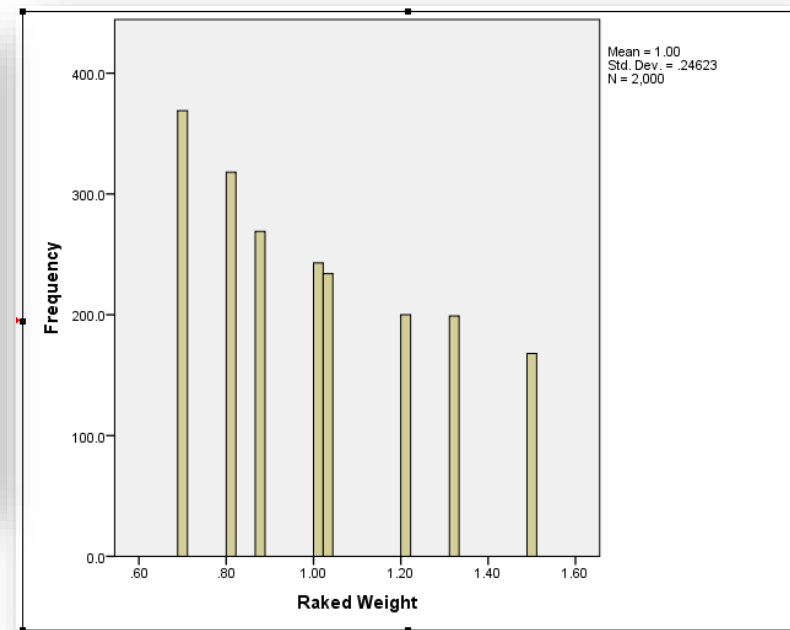
Statistics		
Raked Weight		
N	Valid	2000
	Missing	0
Mean		1.0000
Median		1.0142
Minimum		.70
Maximum		1.51
Percentiles	25	.8122
	50	1.0142
	75	1.2086

Raked Weight				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
.70	369	18.5	18.5	18.5
.81	318	15.9	15.9	34.4
.88	269	13.5	13.5	47.8
1.01	243	12.2	12.2	60.0
1.04	234	11.7	11.7	71.7
1.21	200	10.0	10.0	81.7
1.32	199	10.0	10.0	91.6
1.51	168	8.4	8.4	100.0
Total	2000	100.0	100.0	

Custom Tables

[schoolSurvey] C:\aarp\schoolsurvey.sav

			gender			
			1 Female		2 Male	
			Mean	Standard Deviation	Mean	Standard Deviation
grade	9	Raked Weight	.70	.00	1.04	.00
	10	Raked Weight	.81	.00	1.21	.00
	11	Raked Weight	.88	.00	1.32	.00
	12	Raked Weight	1.01	.00	1.51	.00



RAKE can take control totals from datasets

- With many categories, listing categories and values becomes impractical
- If syntax is to be reused, using dataset inputs generalizes syntax and reduces work
- Dataset and within-command specifications can be mixed
- Dataset input is not supported in the dialog box
- With many weights, turn off output of table of weights
 - /OPTIONS SHOWWEIGHTS=NO

Syntax:

```
SPSSINC RAKE DIM1=control variable* control categories and values*  
DIM2 ... DIM10  
FINALWEIGHT=varname POPTOTAL=value
```



```
/DS1 DS=dataset name* CATVAR=category variable* TOTVAR=marginal totals variable*  
/DS2 ... /DS10
```

Example of using datasets with *emp* as datasets to be weighted

jobcatds

 jobcat	 value
1.00	.50
2.00	.30
3.00	.20

minorityds

 minority	 value
.00	.80
1.00	.20

```
dataset activate emp.  
spssinc rake finalweight = weight2  
/ds1 ds=jobcatds catvar=jobcat totvar=value  
/ds2 ds=minorityds catvar=minority totvar=value.
```

Result is the same as if control totals were specified inline

**Sample Balance
Based on Variables:
jobcat, minority**

Balance	
Sample Balance	44.994

Raked Weights

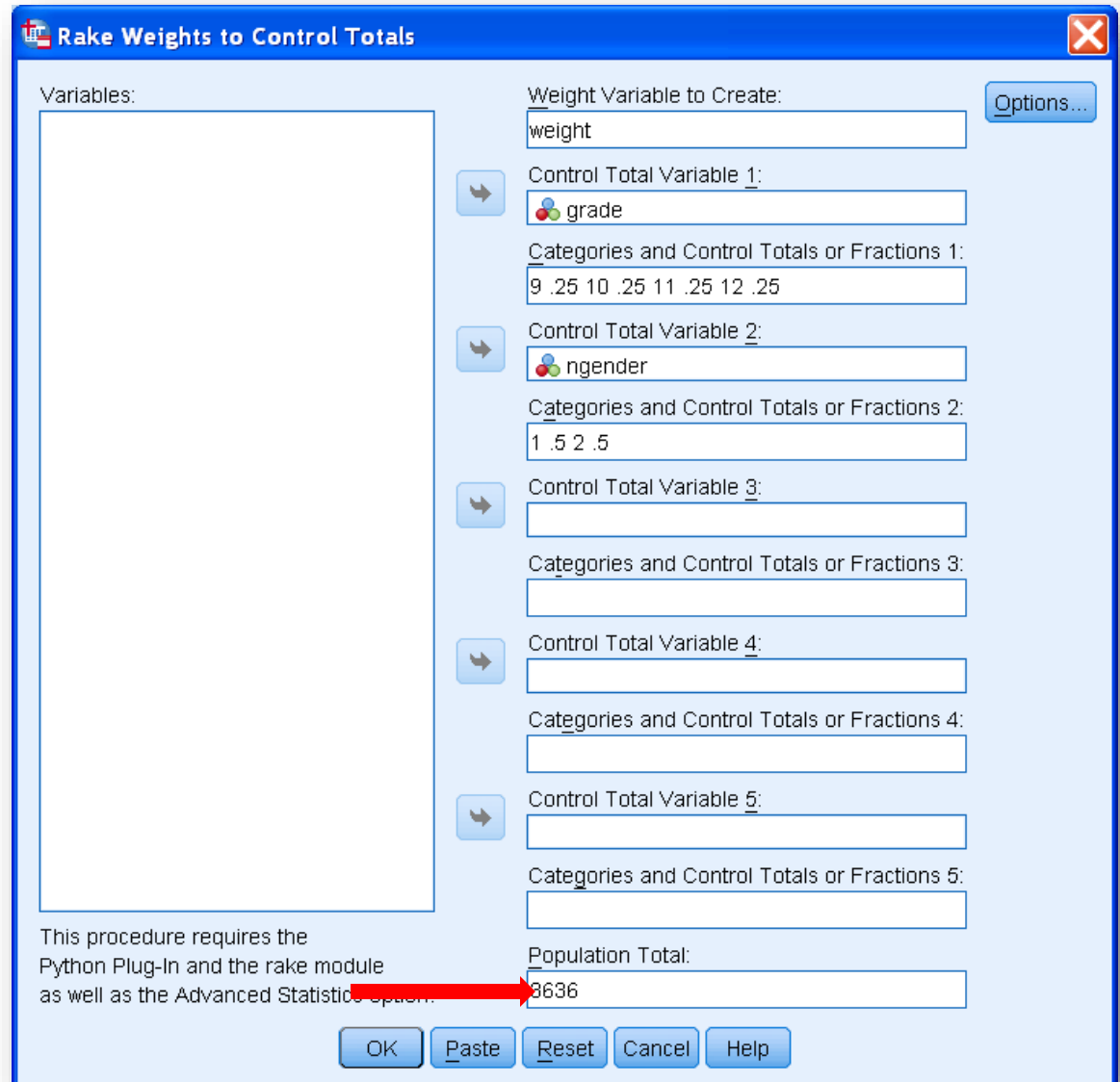
jobcat, minority	Category Rake Weight	Unweighted Case Count
1.0, 0.0	.716	276.000
1.0, 1.0	.452	87.000
2.0, 0.0	6.403	14.000
2.0, 1.0	4.043	13.000
3.0, 0.0	1.149	80.000
3.0, 1.0	.725	4.000

Exercise 2: Study the weight distribution of your output weights

- Turn off weights first
- Frequencies on the weight
- Histogram
- Custom table of mean weight using raking variables in rows, columns, layer

Raking can scale weights up to population totals


- Convenient for scaling tabulations
- DO NOT USE for inference





Rake Weights to Control Totals


Variables:


Weight Variable to Create: [Options...](#)

 Control Total Variable 1:
Categories and Control Totals or Fractions 1:

 Control Total Variable 2:
Categories and Control Totals or Fractions 2:

 Control Total Variable 3:
Categories and Control Totals or Fractions 3:

 Control Total Variable 4:
Categories and Control Totals or Fractions 4:

 Control Total Variable 5:
Categories and Control Totals or Fractions 5:

Population Total:

This procedure requires the Python Plug-In and the rake module as well as the Advanced Statistics option.

[OK](#) [Paste](#) [Reset](#) [Cancel](#) [Help](#)

ISSUES WITH RAKING

Empty categories stay empty

- Using
... \samples\english\employee data.sav
- There are no female custodians in this dataset.
- Rake proportions (hypothetical)
 - jobcat .67 .10 .23
 - gender: .45 .55
- Weight distribution still achieves marginals
- Completely empty categories are ignored

Employment Category * Gender Crosstabulation				
Count				
Before		Gender		Total
		f Female	m Male	
Employment Category	1 Clerical	206	157	363
	2 Custodial	0	27	27
	3 Manager	10	74	84
Total		216	258	474

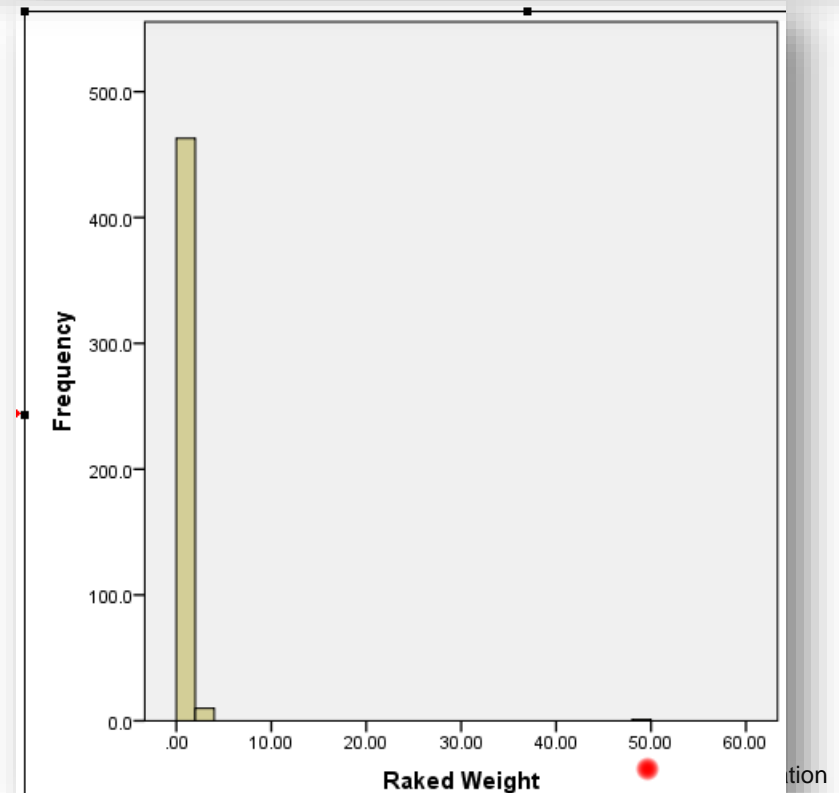
After weighting			Gender		
			Female	Male	Total
Employment Category	Clerical	Count	198	120	318
		Row N %	62.2%	37.8%	100.0%
		Column N %	92.6%	46.1%	67.0%
	Custodial	Count	0	47	47
		Row N %	0.0%	100.0%	100.0%
		Column N %	0.0%	18.2%	10.0%
	Manager	Count	16	93	109
		Row N %	14.5%	85.5%	100.0%
		Column N %	7.4%	35.8%	23.0%
Total		Count	213	261	474
		Row N %	45.0%	55.0%	100.0%
		Column N %	100.0%	100.0%	100.0%

			Gender		
			Female	Male	Total
			Mean	Mean	Mean
Employment Category	Clerical	Raked Weight	.96	.76	.87
	Custodial	Raked Weight	.	1.76	1.76
	Manager	Raked Weight	1.58	1.26	1.30
	Total	Raked Weight	.99	1.01	1.00

Consider what happens with a very small count

After		Gender					
		Female		Male		Total	
		Count	Column N %	Count	Column N %	Count	Column N %
Employment Category	Clerical	202	89.7%	83	30.3%	285	57.0%
	Custodial	0	0.0%	50	18.2%	50	10.0%
	Manager	23	10.3%	92	33.4%	115	23.0%
	4	0	0.0%	50	18.2%	50	10.0%
	Total	225	100.0%	275	100.0%	500	100.0%

- Change one case to jobcat=4
- Rake proportions
 - jobcat .57 .10 .23 .10
 - gender: .45 .55
- The raker meets the goal, but the weight is extreme
 - Median: .9795
 - IQR: .4494
 - Max : median + 109 * IQR !
- This increases variability of results
- Solutions
 - Bound the weights by iteratively reraking
 - Collapse rare categories
 - Clustering?



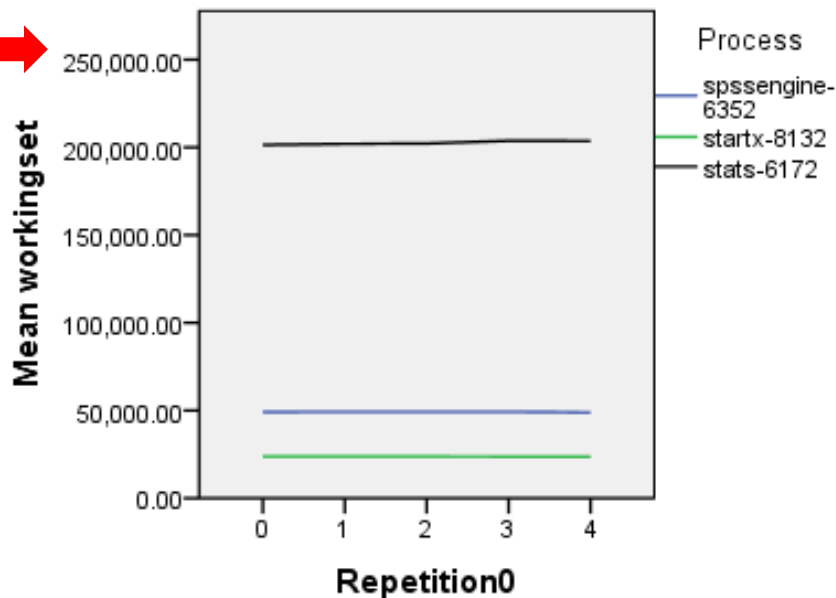
SPSSINC RAKE allows up to ten dimensions

- The higher the number of dimensions the sparser and larger is the table
 - May produce more extreme weights
 - May not converge
 - May take a long time or run out of memory on 32-bit systems
- Ten dimensions with 4 categories each implies a weight table of size $4^{10} = 1,048,576$ cells
 - Likely to exceed your sample size resulting in many empty cells
- Run time will depend on
 - Number of dimensions
 - Number of categories
 - Number of cases
 - How far the input data marginals are from the control totals
- A problem with 10 dimensions, each with 4 categories, total of 2000 cases, simulated data
 - GENLOG ran out of memory on a 32-bit system
 - Completed on a Win7 64-bit system in 1 hour
 - SHOW=YES would display the GENLOG output, but in this case that includes a table with more than 2,000,000 cells, which triggers an error and stops the procedure unless you up the limit using SHOW
 - But don't!

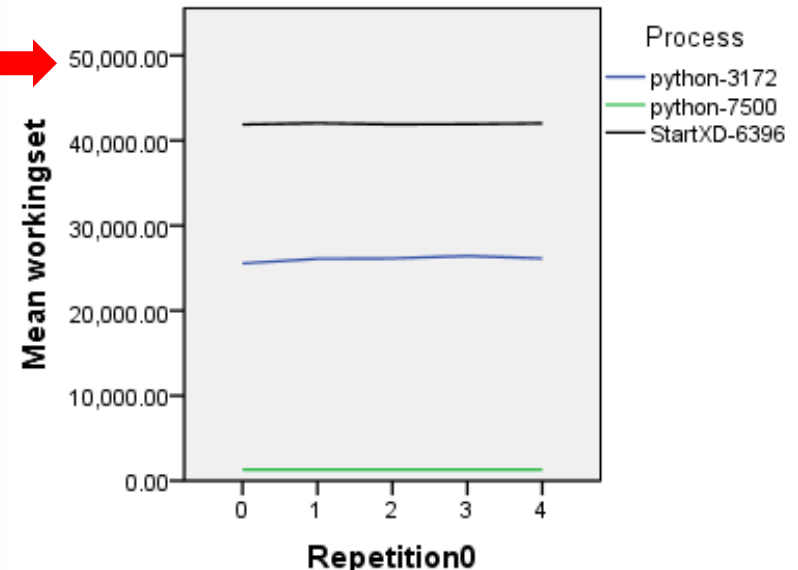
Performance statistics show the benefit of *external mode*

- External mode runs from a command prompt with no user interface like Statisticsb via Python
 - Statisticsb is part of SPSS Statistics Server
- External mode or Statisticsb can use much less memory and time
- A job can be converted to external mode with just a few keystrokes and run from Python
 - Cannot use scripting apis
 - Can capture Viewer output with OMS

Memory Usage by Process - Syntax Window
Five Dimensions with Four Categories



Memory Usage by Process - External Mode
Five Dimensions with Four Categories



The rake job in traditional and external mode forms

Internal mode – syntax window

```
get file="c:/aarp/tendimensions.sav".
* five rake vars each with 4 categories.
) SPSSINC RAKE DIM1 = V1 0 15 1 35 2 45 3 20
DIM2=V2 0 15 1 35 2 45 3 20
DIM3=V3 0 15 1 35 2 45 3 20
DIM4=V4 0 15 1 35 2 45 3 20
DIM5=V5 0 15 1 35 2 45 3 20
) FINALWEIGHT=weight.
```

External mode – run from Python

```
import spss
spss.Submit("""insert file="c:/aarp/rake5.sps"."""))
```

or

```
import spss
spss.Submit("""get file="c:/aarp/tendimensions.sav".
* five rake vars each with 4 categories.
SPSSINC RAKE DIM1 = V1 0 15 1 35 2 45 3 20
DIM2=V2 0 15 1 35 2 45 3 20
DIM3=V3 0 15 1 35 2 45 3 20
DIM4=V4 0 15 1 35 2 45 3 20
DIM5=V5 0 15 1 35 2 45 3 20
FINALWEIGHT=weight."""))
```

- In external mode there is no user interface, Viewer, or Data Editor present, but Viewer (spv) files can be produced via OMS
- Run the Python job as
 - *python myjob.py*
 - Plain text output can be captured or suppressed
- Statisticsb operates similarly but without the need for Python

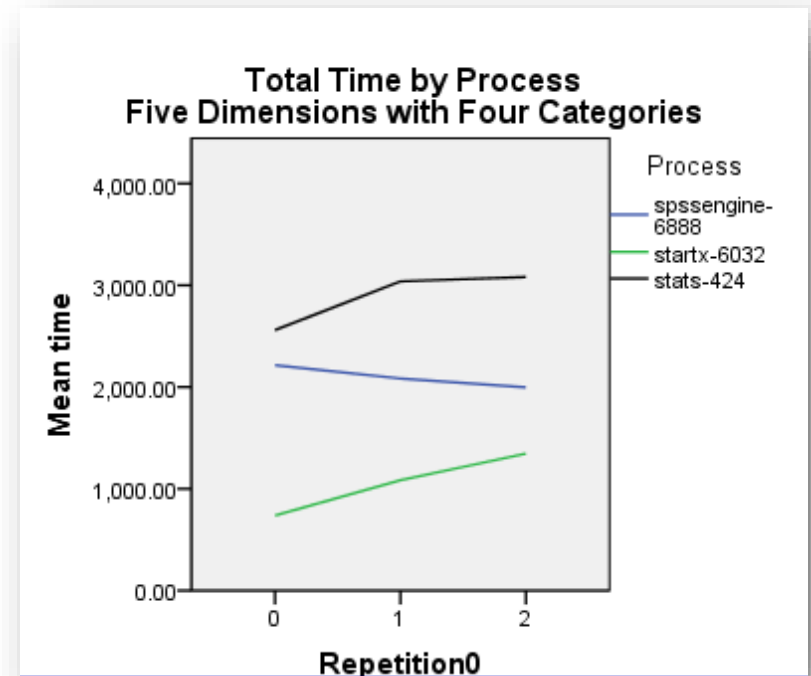
Exercise 3: Run rake in external mode

- Create *myjob.py* in any text editor such as Notepad or using the syntax window
 - Use quotes around the name when saving to force the extension to be .py
- Add SAVE command, e.g.,
 - *SAVE OUTFILE="c:\aarp\tendimensionsraked.sav"*.
- Run the Python job as
 - *python myjob.py*
 - May need path: *c:\python27\python myjob.py* or *c:\python26\python myjob.py*
- Open Statistics and examine the weight
- Set up OMS output capture and run again

```
oms /destination outfile="c:/temp/results.spv" format=spv.  
<your syntax>  
omsend.
```

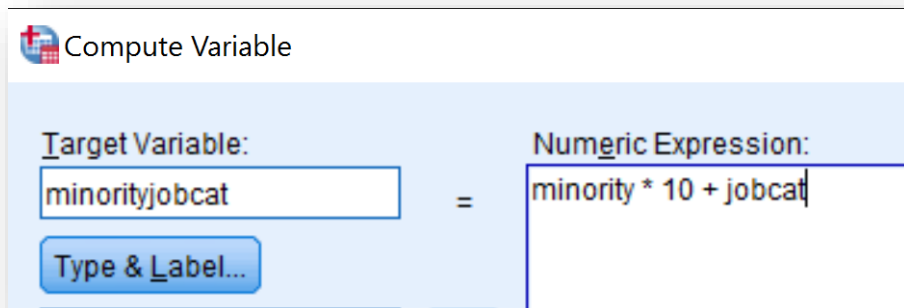
External mode can be much faster and use much less memory

- Capture Viewer output in spv format with OMS
- Viewer output by default appears as text in a console window, but that can be suppressed
- External mode requires Python Essentials
 - So does SPSSINC RAKE
- External mode applications can also be built with .NET plugin and provide a user interface



Control total proportions can be nested to allow for interactions

- You may have nested proportions, e.g., minority proportions within job categories
- As jobcat and minority dimensions raking would adjust only overall proportions
- If there are interactions
 - Combine the interacting dimensions into a single variable
 - Express the control totals as multiplicative terms
- Requires more knowledge of the joint distribution
- In the limit this amounts to specifying the entire joint distribution



The image shows the 'Compute Variable' dialog box in SPSS. The 'Target Variable' is 'minorityjobcat' and the 'Numeric Expression' is 'minority * 10 + jobcat'. There is a 'Type & Label...' button below the target variable field.

Target Variable:		Numeric Expression:
minorityjobcat	=	minority * 10 + jobcat

Type & Label...

Distribution before weighting shows an interaction by the IOT

Employment Category * Minority Classification Crosstabulation

			Minority Classification		Total
			0 No	1 Yes	
Employment Category	1 Clerical	Count	276	87	363
		% within Employment Category	76.0%	24.0%	100.0%
	2 Custodial	Count	14	13	27
		% within Employment Category	51.9%	48.1%	100.0%
	3 Manager	Count	80	4	84
		% within Employment Category	95.2%	4.8%	100.0%
Total	Count		370	104	474
	% within Employment Category		78.1%	21.9%	100.0%

Raked proportions take nested structure into account

Weight Variable to Create:

w

Control Total Variable 1:

minorityjobcat

Categories and Control Totals or Fractions 1:

1 ".75*.8" 2 ".05*.1" 3 ".20*.9" 11 ".75*.2" 12 ".05*.9" 13 ".20*.1"

Employment Category

	Frequency	Percent	Valid Percent	Cumulative Percent
1 Clerical	356	75.0	75.0	75.0
2 Custodial	24	5.0	5.0	80.0
3 Manager	95	20.0	20.0	100.0
Total	474	100.0	100.0	

Minority Classification

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 No	372	78.5	78.5	78.5
	1 Yes	102	21.5	21.5	100.0
	Total	474	100.0	100.0	

minorityjobcat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	284	60.0	60.0	60.0
	2.00	2	.5	.5	60.5
	3.00	85	18.0	18.0	78.5
	11.00	71	15.0	15.0	93.5
	12.00	21	4.5	4.5	98.0
	13.00	9	2.0	2.0	100.0
	Total	474	100.0	100.0	

Partial knowledge of the joint distribution can be used along with the marginal distributions

- School example revisited
- Suppose known population information includes
 - gender totals
 - grade totals
 - gender by grade totals
 - totals of students from single parent households by grade
 - But not known by gender grade
- Solution
 - Compute a variable representing both gender and grade

COMPUTE gendergrade=10 * ngender + grade.

19 = ngender = 1 and grade = 9, i.e.
female frosh
 - Rake by gendergrade and singleparent using joint distribution of gender and grade and marginal distribution of singleparent
- With self-reported data, possible underreporting bias of single parent status (or gender or grade) is not fully resolved
 - differential nonresponse is corrected assuming accurate status

gendergrade			
		Frequency	Percent
Valid	19.00	369	18.5
	20.00	318	15.9
	21.00	269	13.5
	22.00	243	12.2
	29.00	234	11.7
	30.00	200	10.0
	31.00	199	10.0
	32.00	168	8.4
	Total	2,000	100.0

Issues: control totals

- Where do these come from?
- Algorithm treats them as exact
 - Use a source with little or no sampling variability such as census data
 - Could use a prior identical survey to control period to period variability
 - Alternative is to use the raking dimensions as controls
- Case-control matching might be useful if there is a treatment
 - FUZZY extension command can select these samples
- What if some categories of a raking variable are not available for other variables?
 - state by ethnicity
 - American Indian by state
 - No data on Indians except in a few states
 - Combined with white elsewhere
 - solution: create a nested rake dimension combining State and Ethnicity in one dimension using the partial knowledge approach
 - Control totals can be expressed as formulas
- A nested rake dimension removes the independence assumption otherwise used

stateIndian	control	state	nonindian
...	-	-	-
Nebraska	.10	.10	1.00
NewMexicoNonIndian	.04	.05	.80
NewMexicoIndian	.01	.05	.20
NewYork	.20	.20	1.00

Sources of control totals (Michael P. Battaglia, David Izrael, David C. Hoaglin, and Martin R. Frankel, AAPOR)

- Large samples or actual censuses minimize sampling error
 - Census short-form data.
 - Census long-form variables,
 - Census PUMS,
 - Current Population Survey,
 - Census Bureau population projections,
 - National Health Interview Survey
 - American Community Survey
- Variable definitions not always consistent or implemented in the same way
- Sources may have only partial coverage
 - Rake the relevant data and use the raked weights (1 outside the initially raked cases) as input into the larger rakings
 - Or rake each subgroup separately, aggregate and rake to known national totals

Issue: missing data

- Values of a raking variable recorded as missing for some cases (item nonresponse)
- Example: age group sometimes not reported
- Affects matching national totals
- Solution: Calculate missing percentage and subtract from total age counts

Imputation of missing data in control variables

▪ Approaches

- Omit cases and adjust control totals
- *Data > Replace Missing Values (RMV)*
 - Meant for time series data
- *Analyze > Missing Value Analysis (MVA)*
 - Missing Values option
- *Analyze > Multiple Imputation (MULTIPLE IMPUTATION)*
 - Missing Values option
- Hot deck – replace missing values with values from a similar case
 - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/>
 - No built-in command
 - Use FUZZY extension command with additional processing
 - FUZZY does case-control matching
 - Treat pool of complete cases as the controls
 - Match on selected non-missing values
 - Merge values from matched cases into the target dataset
- All missing value imputation methods have assumptions that are often unverifiable
- Do sensitivity analysis to see how imputations affect results

MVA – first understand the missing data

Missing Value Analysis: Patterns

Display

☒ Tabulated cases, grouped by missing value patterns
 Omit patterns with less than % of cases

☒ Sort variables by missing value pattern

☐ Cases with missing values, sorted by missing value patterns
☒ Sort variables by missing value pattern

☐ All cases, optionally sorted by selected variable

Variables

Missing Patterns for:

Additional Information for:

jobtime
prevexp
salary
salbegin
educ
gender
jobcat

jobtime
prevexp
salary
salbegin

Sort by:

Sort Order

☒ Ascending
☐ Descending

	N	Mean	Std. Deviation	Missing		No. of Extremes ^a	
				Count	Percent	Low	High
jobtime	465	80.89	10.032	9	1.9	0	0
prevexp	474	95.86	104.586	0	.0	0	26
salary	466	\$34,552.41	\$17,178.874	8	1.7	0	51
salbegin	468	\$16,949.48	\$7,881.182	6	1.3	0	60
educ	471			3	.6		
gender	468			6	1.3		
jobcat	471			3	.6		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Tabulated Patterns

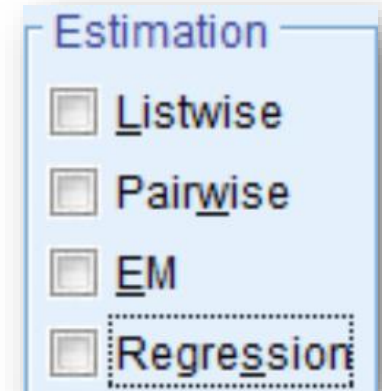
Number of Cases	jobtime	gender	salary	Complete if ... ^b	jobtime ^c	prevexp ^c	salary ^c	salbegin ^c
443				443	80.18	96.98	\$34,390.35	\$17,131.48
8			X	451	95.25	120.13	.	\$13,537.50
6		X		449	94.00	114.33	\$27,875.00	\$14,250.00
5	X			448	.	16.00	\$24,630.00	\$10,740.00

Patterns with less than 1% cases (5 or fewer) are not displayed.

- a. Variables are sorted on missing patterns.
- b. Number of complete cases if variables missing in that pattern (marked with X) are not used.
- c. Means at each unique pattern
- d. Frequency distribution at each unique pattern

MVA: Use imputation methods to fill in values – fixed values

- Listwise and pairwise report estimated statistics
- Regression and EM methods can add random term for dispersion
- Assumptions
 - MCAR – missing completely at random
 - Missing pattern does not depend on the data at all – completely random
 - Assumed for listwise, pairwise, and regression
 - Gives unbiased estimates of means, covariances if true
 - MAR – missing at random
 - Data with the same observed values have the same conditional distribution as the other variables
 - Probability that values are missing and the values depend on observed values but not unobserved
 - Assumed for EM method
- Difficult to validate these assumptions but Little's MCAR tests MCAR null
- Imputation does not guarantee that realized values are actually legal or in range
 - Data Validation procedure can be very helpful here
 - Checks valid values and cross-variable relationships
- If missing data is substantial, check sensitivity of results to different imputation methods



MULTIPLE IMPUTATION preserves the distribution

- Builds a model of missingness
 - MCAR and MAR are handled
 - Automatic method or customized
 - Important that measurement levels are set correctly
 - Uses frequency and sampling weights
 - Support for range constraints and custom imputation models
- Generates multiple samples drawing randomly from the estimated missing distributions
- Uses SPLIT FILES mechanism to manage these
 - Special split variable is created: *Imputation_*
 - Don't turn this off!
 - You do not suddenly have five times as much data
- Procedures that support multiple imputation show result for each split and then merge results appropriately
- Procedures supporting MI including Frequencies, Descriptives, Crosstab, Means, Correlations, Generalized Linear Models, Regression, Binary Logistic, and others but not Custom Tables
- Rake first and then impute and analyze

Issues: unrakeable problems

- Can't make bricks without straw – if a category doesn't occur in a sample, raking can't match its total

Before

x * y Crosstabulation					
			y		Total
			1.00	2.00	
x	1.00	Count	10	0	10
		% within x	100.0%	0.0%	100.0%
		% within y	100.0%	0.0%	33.3%
	2.00	Count	0	20	20
		% within x	0.0%	100.0%	100.0%
		% within y	0.0%	100.0%	66.7%
Total	Count	10	20	30	
	% within x	33.3%	66.7%	100.0%	
	% within y	100.0%	100.0%	100.0%	

After

			y		Total
			1.00	2.00	
x	1.00	Count	15	0	15
		% within x	100.0%	0.0%	100.0%
		% within y	100.0%	0.0%	50.0%
	2.00	Count	0	15	15
		% within x	0.0%	100.0%	100.0%
		% within y	0.0%	100.0%	50.0%
Total	Count	15	15	30	
	% within x	50.0%	50.0%	100.0%	
	% within y	100.0%	100.0%	100.0%	

```
SPSSINC RAKE DIM1 = x 1 50 2 50
DIM2=y 1 75 2 25
FINALWEIGHT=wt.
```

Impossible problems are - impossible

- Structurally empty cells will stay that way
- Seeding empty cells when data are sparse may result in large weights and, hence larger variance
- Probably better to collapse categories to avoid empty cells except for structurally empty (e.g. pregnant males)

Issues: Bounding the weights

- Sometimes raking can produce a few very large weights
- Examine histogram and quartiles of weight
 - May increase sampling variability
 - Bounding the weight introduces some bias but may reduce MSE
- Ad hoc rules for bounding
 - $5 * \text{mean weight}$
 - $(\text{median weight}) + 6 * \text{IQR of weight}$
- Algorithm
 - Rake
 - Truncate output weights according to bounding rule
 - Rake starting with truncated weights
 - Repeat until all weights within bounds or boredom sets in
 - In truncation after first step use 1 + criterion
 - This can take many iterations

Python program run in Statistics or in external mode can bound weights

Part 1 - run a RAKE command in a program from the syntax window

- rakeTrimmer.sps
- Python code placed between *begin program* and *end program*

```
begin program.  
# rake iteration example - straw man  
import spss, spssaux  
  
cmd = r"""SPSSINC RAKE DIM1 = jobcat 1 .57 2 .10 3 .23 4 .10  
DIM2=ngender 1 .45 2 .55 FINALWEIGHT=weight."""  
spss.Submit(cmd)
```

Part 2 - Check the weight distribution and, if necessary, trim large weights and run RAKE again

```

for i in range(10):
    spss.Submit("WEIGHT OFF")
    tag, err = spssaux.createXmlOutput(r""""FREQUENCIES weight /format notable/ntiles=4/statistics=max""",
        visible=True)
    values = spssaux.getValuesFromXmlWorkspace(tag, "Statistics", cellAttrib="number")
    median = float(values[-2])
    iqr = float(values[-1]) - float(values[-3])
    themax = values[-4]
    criterion = median + 6 * iqr
    if i > 0:
        criterion += 1
    if themax <= criterion:
        break
    spss.Submit("""compute weight2 = min(weight, %s).
execute.
delete variable weight.
weight by weight2."" % criterion)
    spss.Submit(cmd)

```

Statistics		
Raked Weight		
N	Valid	474
	Missing	0
Maximum		50.00
Percentiles	25	.5301
	50	.9795
	75	.9795

Part 3 – Report the outcome

```
print ""Trimming Iterations: %s\n
Final Criterion: %s\n
Final Maximum Weight: %s"" % (i+1, criterion, themax)
end program.
```

Trimming Iterations: 10

Final Criterion: 2.49540909795

Final Maximum Weight: 20.342139354167

- A real implementation would parameterize the rake specification and maximum iteration limit, and stabilize the count

Employment Category				
		Frequency	Percent	Valid
Valid	Clerical	106	57.0	
	Custodial	19	10.0	
	Manager	43	23.0	
	4	19	10.0	
	Total	186	100.0	

Gender				
		Frequency	Percent	Valid I
Valid	Female	84	45.0	
	Male	102	55.0	
	Total	186	100.0	

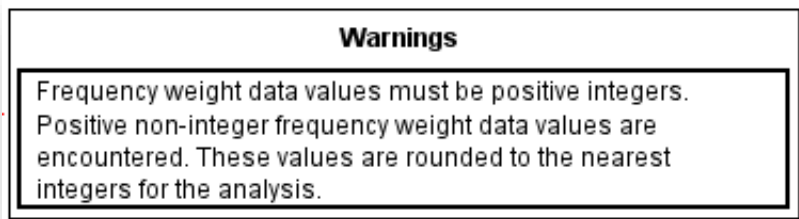
USING RAKE-WEIGHTED DATA IN STATISTICS PROCEDURES

There are several types of weights

- Types of weights
 - Frequency or replication weights
 - Probability of selection weights
 - Sample adjustment weights
 - Importance weights
 - Modeling weights
- All of these apply to some procedures in Statistics
- It is important to understand how a particular procedure interprets weights
- But the implications are different depending on the statistic
- Generally counts, means, percentages can be handled, but inferential statistics are messier

SPSS procedures vary in how weights are treated

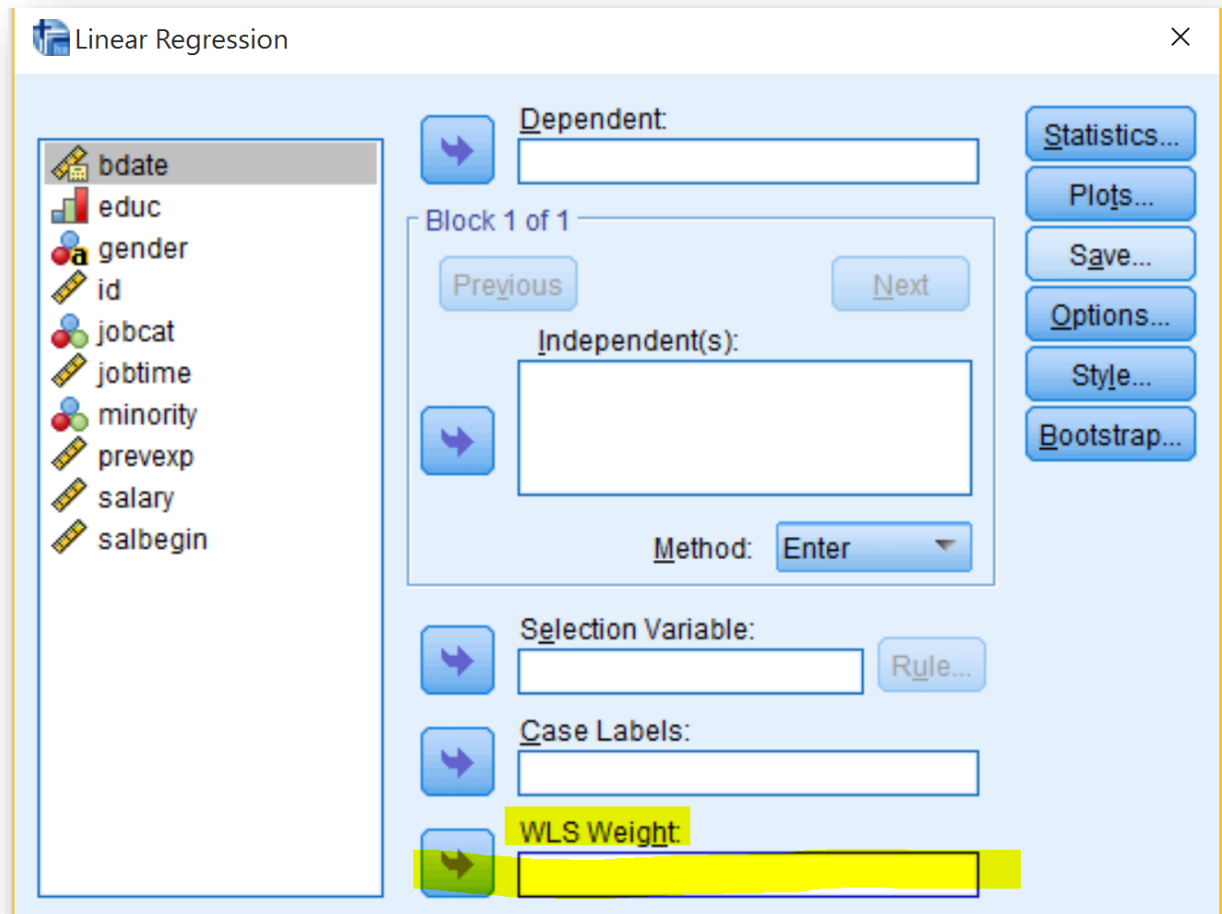
- Procedures such as Frequencies, Crosstabs, and Ctables will use fractional weights
- Most other procedures treat weight as a replication weight and round
 - NPTESTS



- Some procedures ignore weights
- Degrees of freedom will be wrong in procedures such as regression if sum of weights does not match number of cases and weights do not represent replication
- Complex Samples procedures treat weights as representing probability of selection
 - Computations reflect sampling design
 - CSCOXREG vs COXREG for survival problems
 - Gentle raking could be used for post stratification
- Weight status reported in Notes table and on Data Editor status bar

Regression procedure weights are different

- REGRESSION weights are replication weights represent actual case and are not rounded
- Upscaled weights overstate degrees of freedom
- Use the raked weight as the WLS weight to preserve proper degrees of freedom
- Estimates will match CSGLM but standard errors are not the same



The image shows the SPSS Linear Regression dialog box. On the left, a list of variables includes bdate, educ, gender, id, jobcat, jobtime, minority, prevexp, salary, and salbegin. The main area contains fields for Dependent, Independent(s), Selection Variable, Case Labels, and WLS Weight. The WLS Weight field is highlighted in yellow. The Method dropdown is set to Enter. On the right, there are buttons for Statistics..., Plots..., Save..., Options..., Style..., and Bootstrap....

Linear Regression

Dependent:

Block 1 of 1

Previous Next

Independent(s):

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Statistics... Plots... Save... Options... Style... Bootstrap...

Output reports more degrees of freedom than there are cases if standard weight setting is used and weights are not normalized

- t and F d.f are wrong
- Coefficients are okay

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.052 ^a	.003	.002	2.95237

a. Predictors: (Constant), grade

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	85.887	1	85.887	9.853	.002 ^b
	Residual	31,675.656	3,634	8.716		
	Total	31,761.543	3,635			

a. Dependent Variable: y

b. Predictors: (Constant), grade

There are only 2000 cases

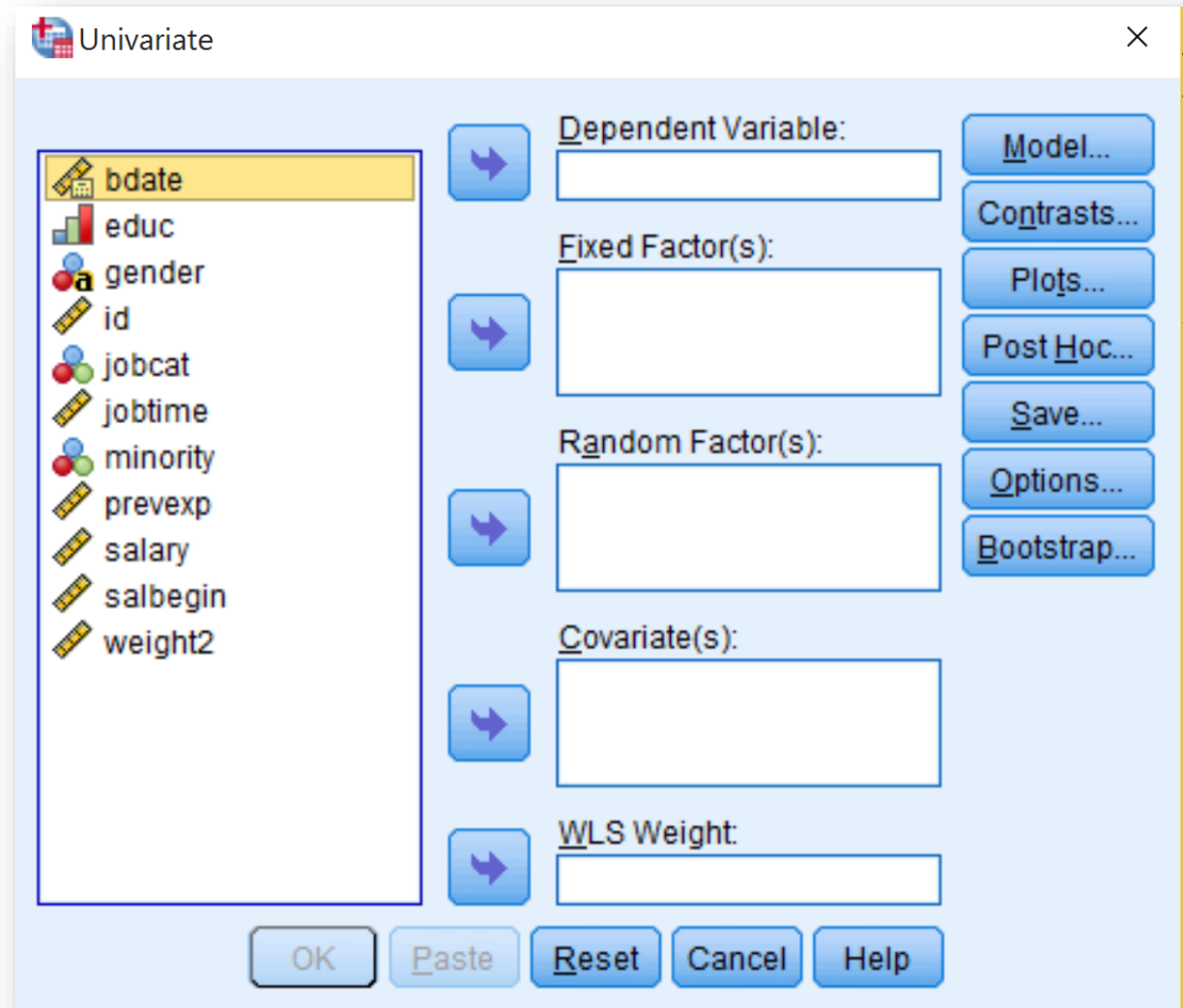
Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.730	.462		23.203	.000
	grade	.137	.044	.052	3.139	.002

a. Dependent Variable: y

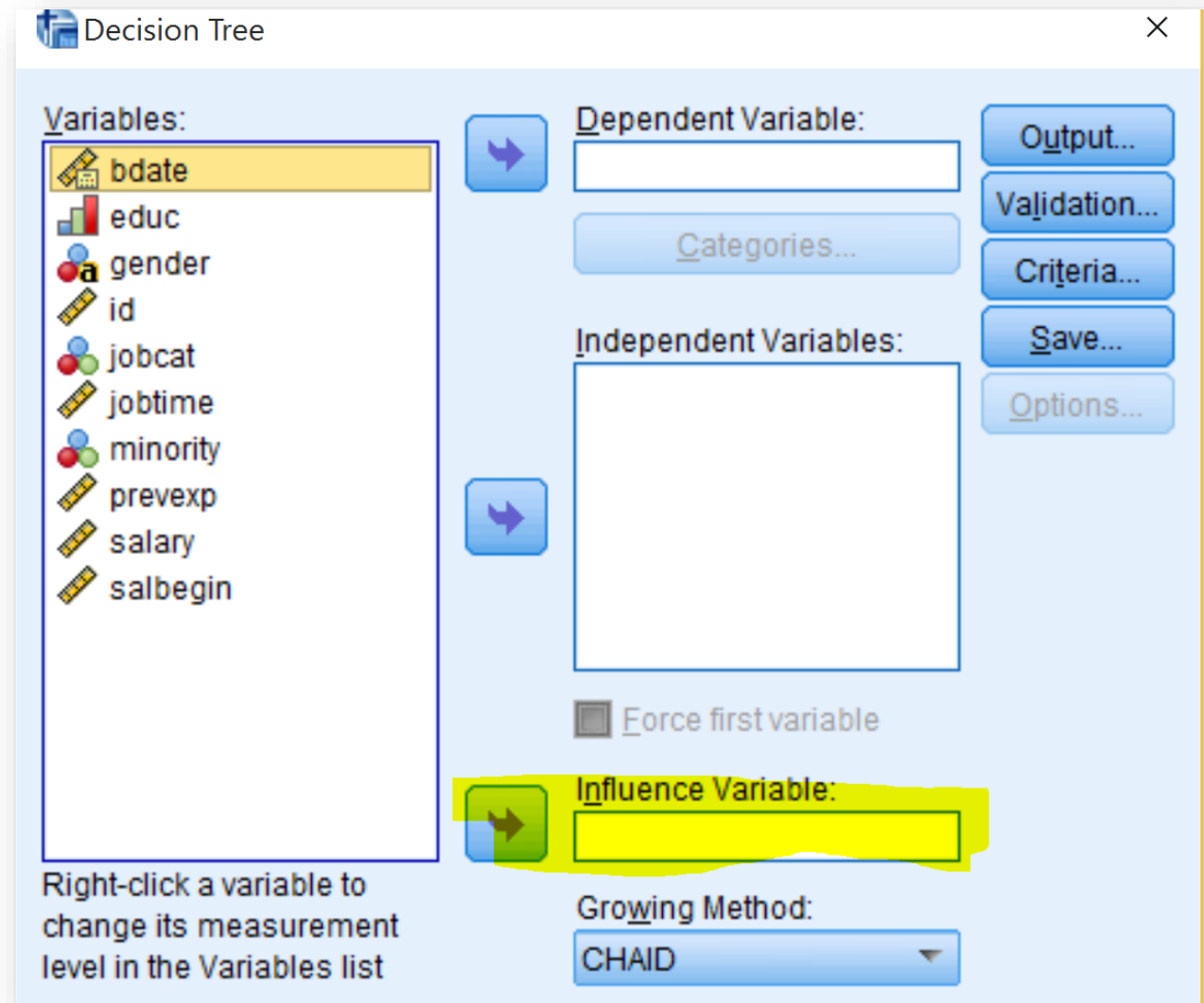
GLM Univariate is partially consistent

- Regular weight is rounded in this procedure
- Use raked weight as WLS weight
- Differ from CSGLM
- **CSGLM** estimators are approximately heteroscedastic consistent



Trees influence weight is not a sampling or replication weight

- Influence weight reflects importance of cases
 - E.g., most profitable customers in a churn model



CROSSTABS weight options determine how fractional weights are treated

- " if the data file is currently weighted by a weight variable with fractional values (for example, 1.25), cell counts can also be fractional values. You can truncate or round either before or after calculating the cell counts or use fractional cell counts for both table display and statistical calculations."
- CROSSTABS options
 - use weights as is
 - round
 - truncate
 - round or truncate accumulated cell counts
- Statistics such as Chi-square are calculated based on the table
 - Set to No adjustments to handle non-integer weights
- Exact Tests require integer weights

Noninteger Weights

- ☒ Round cell counts
- ☐ Round case weights
- ☐ Truncate cell counts
- ☐ Truncate case weights
- ☐ No adjustments

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.038 ^a	3	.257
Likelihood Ratio	4.039	3	.257
Linear-by-Linear Association	2.016	1	.156
N of Valid Cases	3,636		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 454.50.

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Interval by Interval	Pearson's R	.024	.017	1.420	.156 ^c
Ordinal by Ordinal	Spearman Correlation	.024	.017	1.420	.156 ^c
N of Valid Cases		3,636			

Custom tables procedure offers choices for weight treatment

- Table without weights

Custom Tables

Table Titles Test Statistics Options

Variables: bdate educ gender id jobcat jobtime minority prevexp salary salbegin

Categories: Female Male

Columns

		Employment Category							
		Clerical				Custodial			
		Gender				Gender			
		Female		Male		Female		Male	
		Count	Row N %	Count	Row N %	Count	Row N %	Count	Row N %
Educational Level (years)	8	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	12	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	14	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	15	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	16	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	17	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	18	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
	19	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%
20	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	
21	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	
Total	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	nnnn	nnnn.n%	

Define

N% Summary Statistics...

Categories and Totals...

Summary Statistics

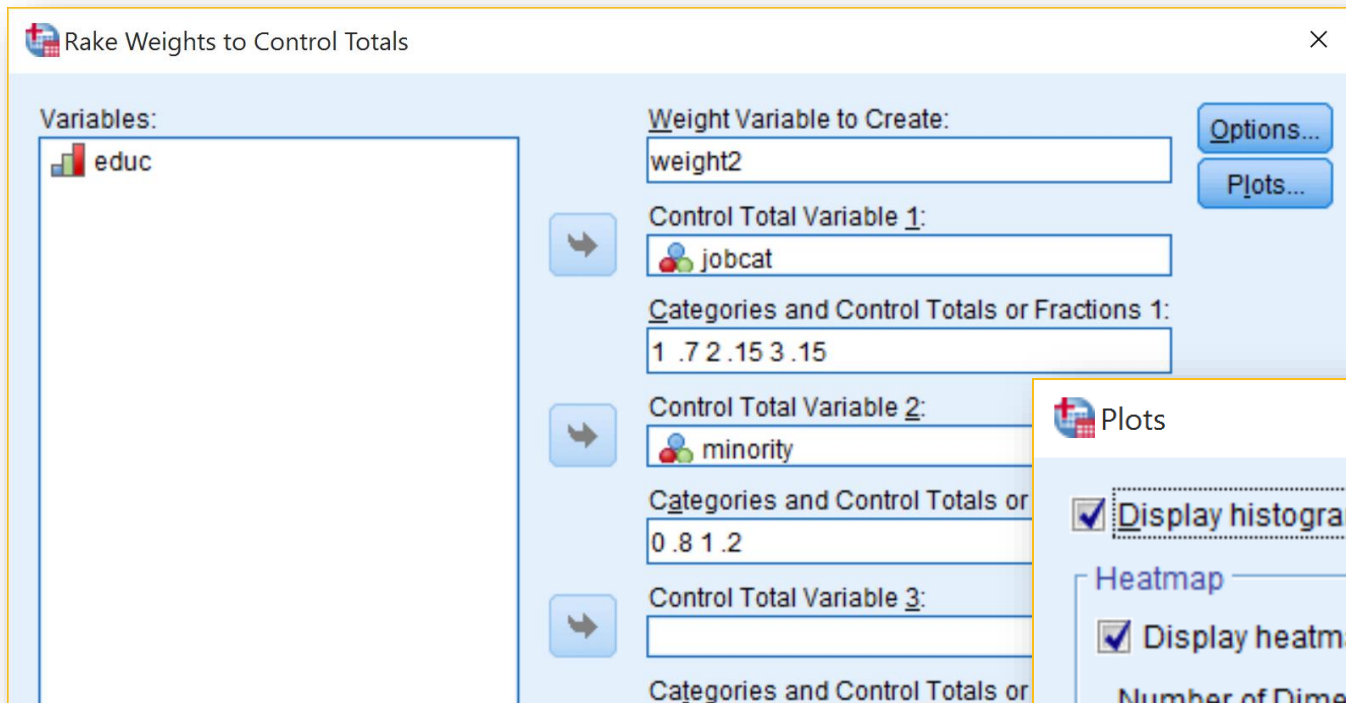
Position: Columns

Source: Row Variables

Category Position: Default

OK Paste Reset Cancel Help

Rake to new control totals



Rake Weights to Control Totals

Variables:
educ

Weight Variable to Create:
weight2

Control Total Variable 1:
jobcat

Categories and Control Totals or Fractions 1:
1 .7 2 .15 3 .15

Control Total Variable 2:
minority

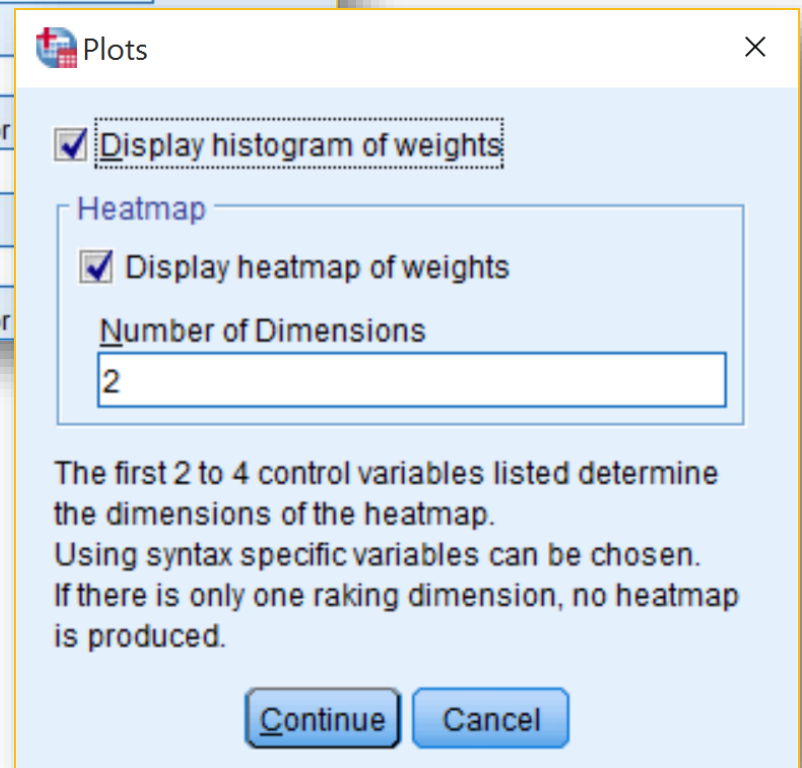
Categories and Control Totals or Fractions 2:
0 .8 1 .2

Control Total Variable 3:

Categories and Control Totals or Fractions 3:

Options...

Plots...



Plots

☒ Display histogram of weights

Heatmap

☒ Display heatmap of weights

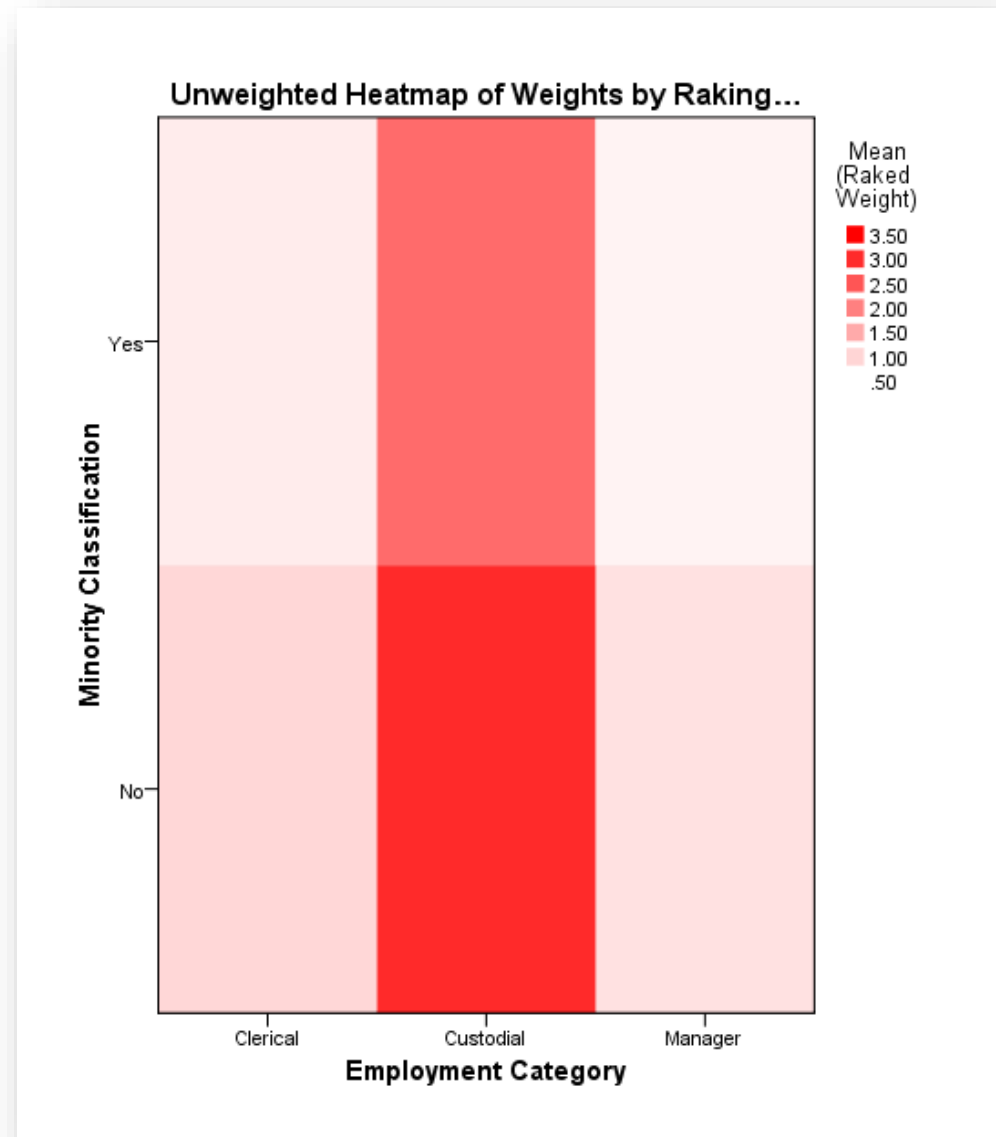
Number of Dimensions
2

The first 2 to 4 control variables listed determine the dimensions of the heatmap.
Using syntax specific variables can be chosen.
If there is only one raking dimension, no heatmap is produced.

Continue Cancel

Weighting the sample

- Heatmap shows the distribution of the weights by the raking variables



Use weight2 as effective base

- More accurate calculation of standard errors, confidence intervals, and test statistics
- Effective base is available as a cell statistic
- Replaces standard weight variable if specified

Effective Base

☒ Use effective base weight variable

Variables

educ
id
jobcat
jobtime

Weight Variable

weight2

		Employment					
		Clerical			Gender		
		Female			Male		
		Count	Unweighted Count	Adjusted Count	Count	Unweighted Count	Adjusted Count
Educational Level (years)	8	28	30	30	9	10	10
	12	118	128	126	42	48	47
	14	0	0	0	6	6	6
	15	30	33	33	72	78	77
	16	13	14	14	9	10	10

Exploring Effective Weight

- Effective weight typically gives more accurate test statistics
- Still an approximation

Table Weighted by weight2

		Employment Category					
		Clerical (A)		Custodial (B)		Manager (C)	
		Count	Unweighted Count	Count	Unweighted Count	Count	Unweighted Count
Educational Level (years)	8	37 C	40	34 A C	13	0	0
	12	160 C	176	35 C	13	1	1
	14	6	6	0	0	0	0
	15	102 B C	111	2	1	3	4
	16	22	24	0	0	30 A B	35
	17	2	3	0	0	7 A	8
	18	2	2	0	0	6 A	7
	19	1	1	0	0	22 A B	26
	20	0	0	0	0	2 A	2
	21	0	0	0	0	1	1
	Total	332	363	71	27	71	84

Results are based on two-sided tests. For each significant pair, the key of the category with the smaller column proportion appears in the category with the larger column proportion.

Significance level for upper case letters (A, B, C): .05¹

1. Tests are adjusted for all pairwise comparisons within a row of each innermost subtable using the Bonferroni correction.

Use Utilities > Modify Table Appearance to highlight significant cells

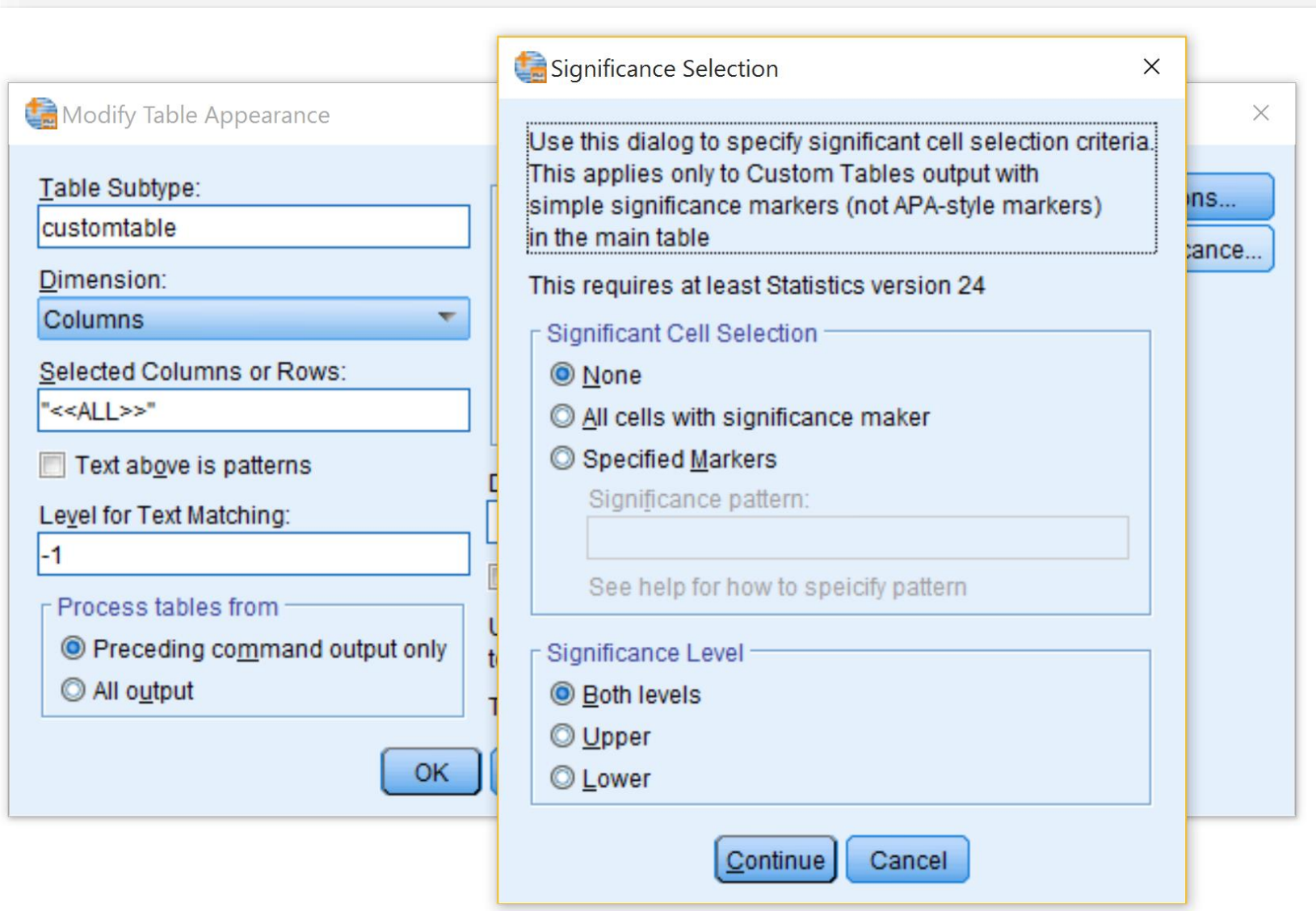
Table Weighted by weight2

		Employment Category					
		Clerical (A)		Custodial (B)		Manager (C)	
		Count	Unweighted Count	Count	Unweighted Count	Count	Unweighted Count
Educational Level (years)	8	37 C	40	34 A C	13	0	0
	12	180 C	176	35 C	13	1	1
	14	6	6	0	0	0	0
	15	102 B C	111	2	1	3	4
	16	22	24	0	0	30 A B	35
	17	2	3	0	0	7 A	8
	18	2	2	0	0	6 A	7
	19	1	1	0	0	22 A B	28
	20	0	0	0	0	2 A	2
	21	0	0	0	0	1	1
	Total	332	363	71	27	71	84

Results are based on two-sided tests. For each significant pair, the key of the category with the smaller column proportion appears in the category with the larger column proportion.

Significance level for upper case letters (A, B, C): .05¹

Requires new version of SPSSINC MODIFY TABLES extension



Effective weight with gender example

numeric Gender
2
1
1
1
1
1
1
1
1
1
1

		Count	Unweig...	Adjuste...
numeric Gender	f	nnnn.nn	nnnn.nn	nnnn.nn
	m	nnnn.nn	nnnn.nn	nnnn.nn
	Total	nnnn.nn	nnnn.nn	nnnn.nn

- Data were raked to give 50/50 split
- Adjusted count is the effective base for a cell
- Cells are homogeneous here
- $3.64 = 11 * .331$

Using Effective Base

		Count	Unweighted Count	Adjusted Count
numericGender	f	5.50	10.00	10.00
	m	5.50	1.00	1.00
	Total	11.00	11.00	3.64

IN CONCLUSION

Summary

- Raking can adjust for nonrepresentativeness of a sample based on knowledge of only the marginal distributions of one or more relevant classifying variables
- This may increase variability of results, reducing bias while increasing variance
- The SPSSINC RAKE extension command constructs the weight
- SPSSINC RAKE is one of many extension commands available free from the SPSS Community website (<http://www.ibm.com/developerworks/spssdevcentral>) or new IBM Predictive Analytics Community (<https://developer.ibm.com/predictiveanalytics/>)
- Good practice is to examine the distribution of the constructed weights and consider adjustment of extremely large values by truncation or iteration
- External mode can conserve computing resources
- Implications of using fractional weights should be considered for each statistical procedure used
- Weighted/reweighted datasets can give a better picture of the population when used in moderation

Questions



My contact information

- Jon Peck
- jkpeck@gmail.com