A Survey on Survey Statistics: What is done, can be done in Stata, and what's missing?

Frauke Kreuter & Richard Valliant

Joint Program in Survey Methodology
University of Maryland, College Park
fkreuter@survey.umd.edu

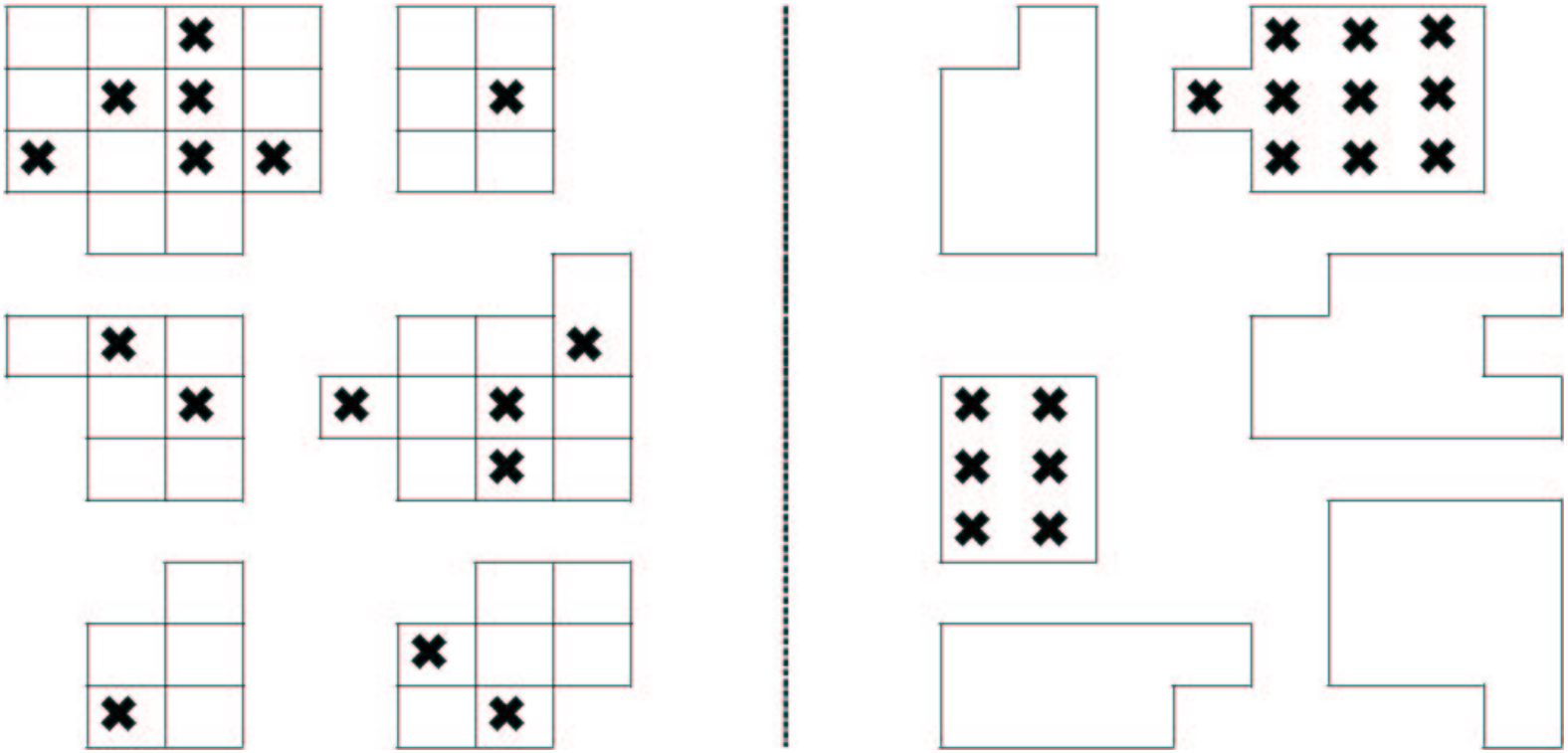
Outline - Questions

- What are the survey design features that I need to take into account?
- How does the survey design effect bias and variance?
- How do I account for complex design in practice?
- How do I analyze subgroups?
- What are my PSUs/clusters?
- What does Stata do compared to other Software?
- A question *not* answered here: How do I treat missing data?

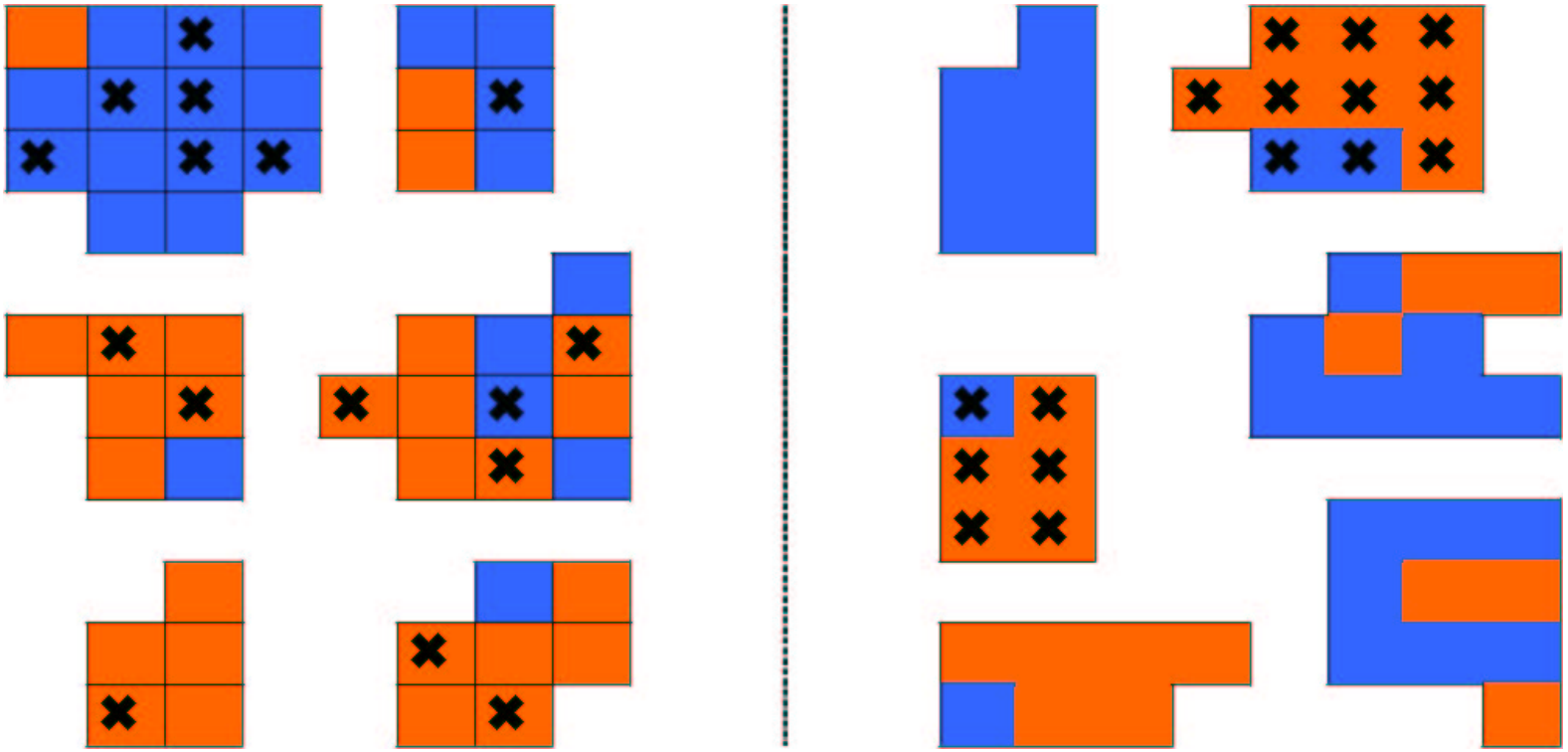
Complex Designs Features that Affect Analysis

- Stratification
  - Units are put into similar groups for sampling
  - Strata are nonoverlapping and cover whole population
  - Example: States within Germany, countries within Europe, types of schools (e.g. PISA)
- Clustering
  - Groups of units that are selected as a group
  - Example: Election districts within States
- Weights account for selection probabilities, nonresponse, adjustment to external control counts (poststratification)

Stratified sample – and – Cluster sample



Effects on standard error



- Stratification
  - may reduce standard errors
- Clustering
  - usually increases standard errors
  - different units within a cluster may tend to be similar in the education and services they receive
  - repeated measures on the same student are correlated; the student can be treated as a cluster for some analyses
- Weights-used to account for unequal selection probabilities, nonresponse adjustments, poststratification
  - to bring sample to level of population when estimating totals
  - when used in models, estimates are of "model that would be fitted if you had entire population in sample"

Design Effects

deff A measure of how much different your sample is from a simple random sample (or sample where data can be treated as independent and identically distributed)

– Definition

$$\text{deff}(\hat{\theta}) = \frac{\hat{\text{var}}(\hat{\theta})}{\hat{\text{var}}_{\text{SRS}}(\tilde{\theta})} = \frac{\text{variance accounting for complexity}}{\text{variance assuming SRS}}$$

another way to think about

$$\text{deff} = 1 + \rho(n-1)$$

This applies to any estimate: mean, total, model parameter. Effects on standard errors are reported as deft $= \sqrt{\text{deff}}$ (if no fpc specified).

In clustered samples the deft's are usually > 1. **Stata** reports deff's, deft's, and meff's.

Examples

- National Health and Nutrition Examination Survey III: 23 strata, 2 PSUs per strata

Hypertension	Sample size	Unweighted	Weights	deff
Yes	449	5.4 %	3.9%	4.19

- Social Science Survey 1997 (Sozialwissenschaftenbus - SowiBus): 603 PSUs

Fear of crime	Subpop.	n	Estimate	SE	deff
	West	2,168	22.5%	0.013	2.00
	East	1,100	30.2%	0.020	2.08

Accounting for Complex Design in Practice

- Outdated: Include terms in the model to implicitly incorporate design features, (e.g., include stratification variables as x variables)
- Use weights but adjust independence-based standard errors using estimated design effects
- Estimate standard errors with methods that account for complex design
- Estimate standard error based on underlying model

Methods for estimating standard errors

- Exact formulas
- Linearization or Taylor series estimation:
  - Approximate an estimator with a linear function, then compute variance of approximation using formula appropriate to sample design
- Replication:
  - Divide sample into subsamples, compute estimate from each subsample and variance among subsample estimates
  - Jackknife, balanced repeated replication (BRR, balanced half-sampling), bootstrap

Pros

Linearization	Replication
good large sample properties	good large sample properties
applies to complex forms of estimates	applies to complex forms of estimates
can be computationally faster	sample adjustments easy to reflect
maximizes degrees of freedom (stability)	no knowledge of design needed
sandwich version is model-robust	avoids disclosure of PSU and Strata

... and Cons

Linearization	Replication
separate formula for each estimate	computationally intensive
special purpose programming	may be unclear how best to form replicates
hard to account for adjustments	increased file sizes
	sometimes applied in ways that loose dfs

Example Implementation in Stata 8

```
. svyset [pweight=examwgt], psu(psu) strata(stratum)

. svydes

pweight:  examwgt
Strata:   stratum
PSU:      psu
```

Strata			#Obs per PSU		
stratum	#PSUs	#Obs	min	mean	max
1	2	370	160	185.0	210
2	2	339	149	169.5	190
3	2	285	129	142.5	156
...					
23	46	8360	70	181.7	246

Stata 8

```
. svyset [pweight=examwgt], psu(psu) strata(stratum)

. svymean poverty food_bev weight , deft

pweight:  examwgt                     Number of obs(*) =      8360
Strata:   stratum                     Number of strata =        23
PSU:      psu                         Number of PSUs   =        46
-------------------------------------------------------------------
    Mean |   Estimate    Std. Err.       Deft
---------+---------------------------------------------------------
 poverty |    3.21537    .1163011    4.822822
food_bev |   2551.368    38.30955    2.675161
  weight |   167.3652    .9343583    2.098078
---------+---------------------------------------------------------
```

Stata 9

```
. svyset psu [pweight=examwgt], strata(stratum)

. svy: mean poverty food_bev weight

. estat effects, deft
```

New Implementation in Stata 9

- Jackknife variance estimators
- BRR variance estimators
- Poststratification

Example code Stata 9:

```
. svy jackknife slope= _b[weight]: reg weight height

. svyset psu [pweight=examwgt], strata(stratum)
brrweight(rpl01-rpl24) vce(brr)
```

Some practical issues

- Strata with one PSU
  - Common error message

    ```
    stratum with only one PSU detected
    r(460);
    ```

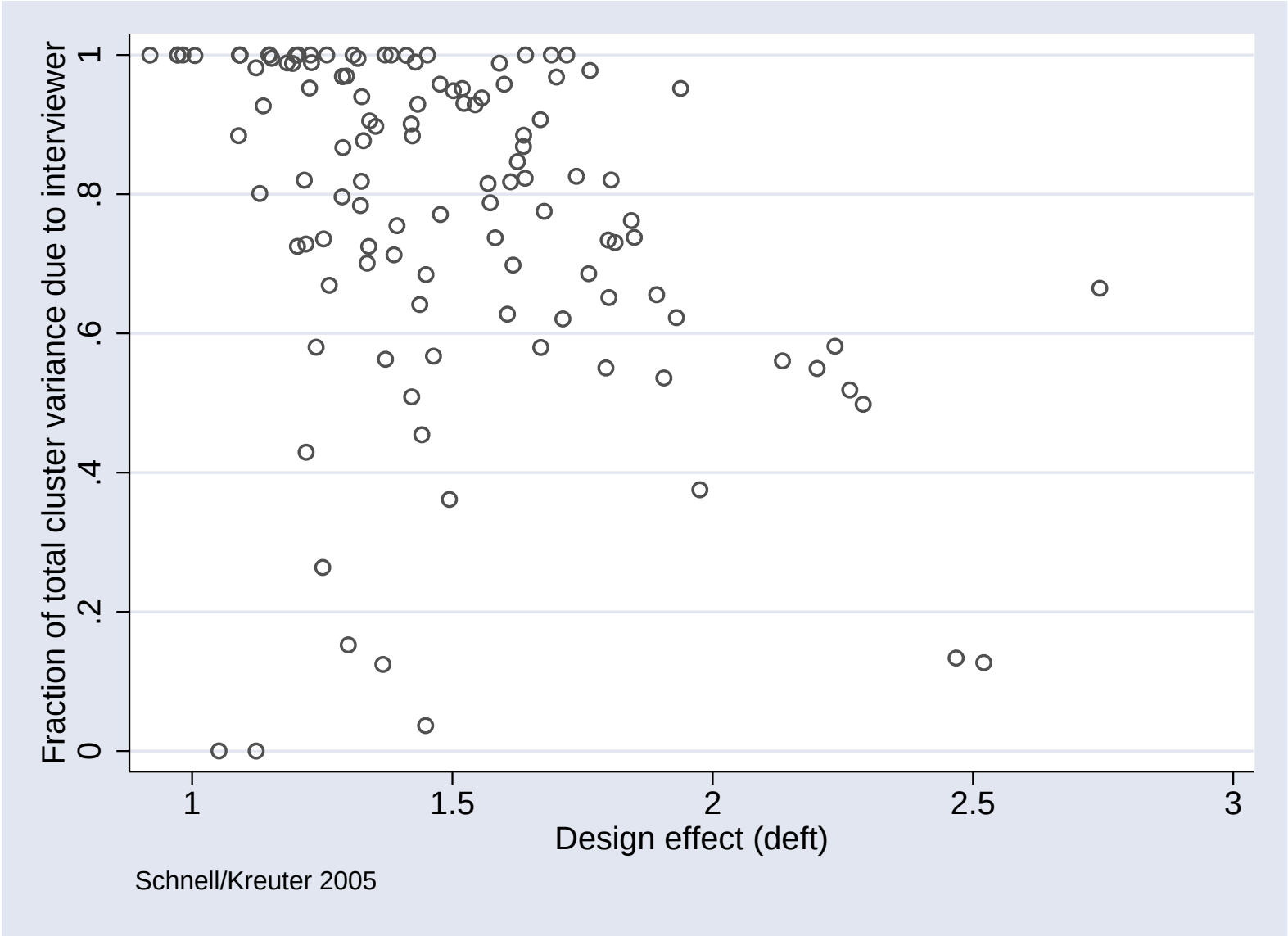
  - Locating singleton PSU

    ```
    . svydes
    ```

- Comparing subgroups
  - Can often lead to singleton PSU
  - Recommended procedure

    ```
    svytab agcat gadlt1, subpop(rhsp)
    ```

Interviewer as part of design effect



Schnell/Kreuter 2005

Comparison to other programs

Method	SPSS CS	STATA	SUDAAN	WesVar	SAS
Taylor Linearization	?	x	x		x
Replicate Weights		x	x	x	
Descriptives	**SPSS CS**	**STATA**	**SUDAAN**	**WesVar**	**SAS**
Means	x	x	x	x	x
Totals	x	x	x	x	x
Ratios	x	x	x	x	
Proportions	x	x	x	x	x
Geometric Means		x	x	x	
Quantiles		?	x	x	

Note: This is a moving target.

Analysis Features	SPSS CS	STATA	SUDAAN	WesVar	SAS
Linear Regression	x	x	x	x	x
Instrumental variables		x			
Interval and censored regression		x			
Logistic Regression	x	x	x	x	x
Multinomial LR	?	x	x	x	
Ordered LR	?	x	x		
Probit Models		x			
Loglinear Models			x		
Tests of Independence in Tables	x	x	x	x	x
Linear Contrasts, Differences		x	x	x	
Poisson regression		x	x		
Survival Analysis		?	x		

Outlook

Is there still something missing?

- Random groups
- Bootstrap estimation - Statistics Canada (used like BRR)
- Sample selection routines
- Weight calculation for nonresponse or unknown eligibility
- Weight calculation for general regression estimator (GREG)

Some questions regarding the current implementation:

- How is Poststratification done?
- Are BRR and jackknife replicate created within Stata? How?
- Any plans to allow the svy prefix for survival models?
- How would you suggest to handle cross-classification?