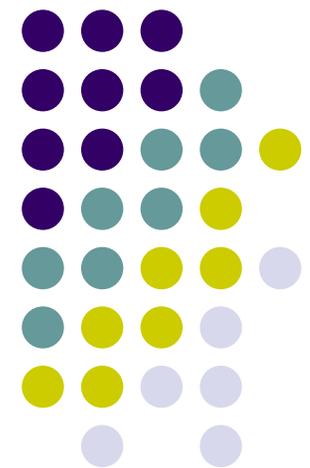


# Basic Analysis of BRFSS Data Using SPSS Complex Samples

Jennifer Haussler MS  
Center for Health Statistics  
October 14, 2009





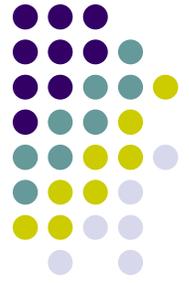
# Objectives

- Light intro to SPSS
- Recode a variable into a new variable
- Create a new variable
- Create a csaplan
- Analyze frequency using complex samples
- Analyze crosstabs using complex samples
- Understanding the outcome of crosstabs
- Use of windows menus
- Use of syntax

# Opening an SPSS Database



- Variable View
- Data View
- Output Window
- Syntax Window



# Creating agegrp50

- Recode age to create agegrp50
- Two values: “under age 50” & “age 50 +”
- Use menus to recode
- Transform – Recode into different variable
- Select “What is your age?” or c12q01 as input variable.
- Enter agegrp50 as output variable name.
- Label: Two category age group [50]



# Creating agegrp50

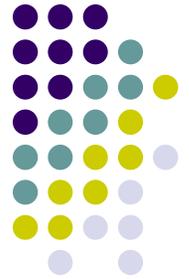
- Click “change” and then “old and new values”
- Old value “18 through 49” – New value 1
- Old value “value (50) through Highest” – New value 2
- Old value “System- or user-missing” – New value “system-missing”
- Click okay then continue



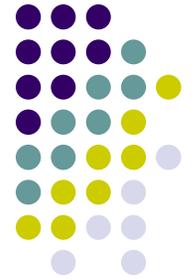
# Creating agegrp50

- Create labels for your values
- Go to variable view – your new variable is at the very bottom
- Double click on the value cell
- Fill in values and labels
- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- C12q01 in rows, agegrp50 in columns

# Creating agegrp50 Using Syntax



```
RECODE c12q01 (MISSING=SYSMIS) (Lowest thru 49=1)
(50 thru Highest=2) INTO agegrp50.
VARIABLE LABELS agegrp50 'Two category age group [50] '.
VALUE LABELS agegrp50
1 'under age 50'
2 'age 50 +'.
FORMATS agegrp50 (F1.0).
```



# Creating “raceeth”

- Variable for race/ethnicity
- Values of White, Black, Hispanic, and Other
- Created from multiple questions in the BRFSS

c12q02

Are you Hispanic or Latino?

c12q04

Which one of these groups would you say best represents your race?

c12q03a

Which one or more of the following would you say is your race?

# Syntax for creating raceeth



```
COMPUTE race = c12q04.  
IF (missing(c12q04)) race = c12q03a.
```

```
FORMATS race (f1.0).  
VARIABLE LABEL race 'Race'.  
VALUE LABEL race  
  1 'White'  
  2 'Black or African American'  
  3 'Asian'  
  4 'Native Hawaiian or Other Pacific Islander'  
  5 'American Indian or Alaska Native'.
```

```
RECODE race (1=1) (2=2) (3 thru 6 = 4) (7 thru 9 = SYSMIS) INTO raceeth.
```

```
IF (c12q02 = 1) raceeth = 3.  
FORMATS raceeth (f1.0).  
VARIABLE LABEL raceeth 'Race/Ethnicity'.  
VALUE LABEL raceeth  
  1 'White'  
  2 'Black'  
  3 'Hispanic'  
  4 'Other'.  
FORMATS raceeth (F1.0).
```



# Check your work again

- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- c12q02 in rows, raceeth in columns
- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- race in rows, raceeth in columns



# Creating “undinsur”

- Variable for three levels of insurance coverage
- Insured, underinsured & uninsured
- Uses two questions from the survey

c03q01

Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?

c03q03

Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?

# Syntax for “undinsur”



```
DO IF ((c03q01=1) and (c03q03=2)).  
  COMPUTE undinsur = 1.  
ELSE IF ((c03q01=1) and (c03q03=1)).  
  COMPUTE undinsur = 2.  
ELSE IF (c03q01=2).  
  COMPUTE undinsur =3.  
END IF.  
EXECUTE .  
VARIABLE LABELS undinsur 'Levels of Insured'.  
VALUE LABELS undinsur  
  1 'Insured'  
  2 'Underinsured'  
  3 'Uninsured'.  
FORMATS undinsur (F1.0).
```



# Check your work again

- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- c03q01 in rows, undinsur in columns
- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- c03q02 in rows, undinsur in columns



# Creating “disabil”

- Variable for disability
- Creates dichotomist variable
- Combines two questions from the survey

c10q01

Are you limited in any way in any activities because of physical, mental, or emotional problems?

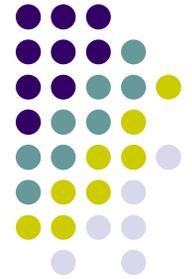
c10q02

Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone?



## Define “disabil”

- Disabled is when c10q01 is yes or c10q2 is yes
- Not disable is when c10q01 is no AND c10q02 is no



# Syntax for “disabil”

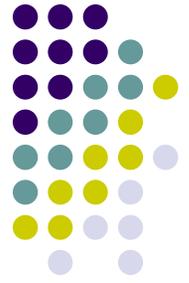
```
DO IF ((c10q01=1) or (c10q02=1)).  
  COMPUTE disabil = 1.  
ELSE IF ((c10q01=2) and (c10q02=2)).  
  COMPUTE disabil = 2.  
END IF.  
EXECUTE .  
VARIABLE LABELS disabil 'Disability'.  
VALUE LABELS disabil  
  1 'yes'  
  2 'no'.  
FORMATS disabil (F1.0).
```



# Check your work again

- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- c10q01 in rows, disabil in columns
- Check your work with crosstabs
- Analyze – Descriptive statistics – crosstabs
- c10q02 in rows, disabil in columns

# Analysis with Complex Samples



- Always use a csaplan file
- Make sure you're using the correct file
- There are separate files for children, split surveys, etc
- It's **ALWAYS** a good idea to have a copy of the questionnaire with you when you are doing analysis.



# Creating a csaplan

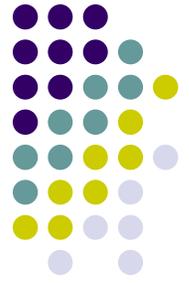
- Click “analyze” – “complex samples” – “prepare for analysis” – “create a plan”
- Name and locate your csaplan
- Strata is always “ststr”
- Cluster is always “seqno”
- Weight changes but for our purposes will be “cdcfwt”
- Click “next” – select “WR” – click “next” – verify selections and click “next” – “save you specifications to a plan file” – click “finish”

# Weighted Frequency of “undinsur”



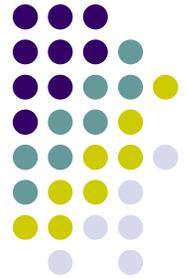
- Analyze – Complex Samples – Frequencies
- Browse for correct csaplan
- Highlight “Levels of insured” and move it to frequency table box using arrow button
- Click on Statistics button
- Uncheck “Population size” and “Standard error”
- Check “Table percent”, “Confidence Interval” and “Unweighted count.”
- Click “continue”, click “okay”

# Why do we want the unweighted count?



- Make sure there is a large enough sample size for a stable estimate
- Valuable tool for replication of analysis
- If you are getting a different unweighted count than someone else doing the same analysis, then you can pinpoint the problem lying within the data or the use of unknowns and refused.
- If you are getting the same unweighted count then you're probably using different weighting plans for the analysis.

# Weighted Frequency of “undinsur”



## Levels of Insured

		Estimate	95% Confidence Interval		Unweighted Count
			Lower	Upper	
% of Total	Insured	65.9%	64.2%	67.5%	8012
	Underinsured	8.5%	7.6%	9.6%	814
	Uninsured	25.6%	24.0%	27.2%	1838
	Total	100.0%	100.0%	100.0%	10664

# Syntax for weighted frequency for “undinsur”



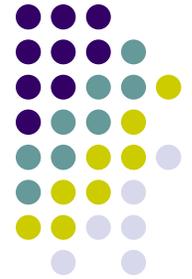
```
CSTABULATE  
  /PLAN FILE='I:\brfss_data\brfss08\State  
\cdcfintwt08.csaplan'  
  /TABLES VARIABLES=undinsur  
  /CELLS TABLEPCT  
  /STATISTICS CIN(95) COUNT  
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

# Crosstabs “undinsur” by “cvd”



- Need to know which way you want to look at it
- People who are underinsured that have CVD
- People who have CVD that are underinsured
- Follow same steps for “undinsur” frequency, just add CVD to “subpopulations”
- Now switch the two
- Which one gives you the table you need?

# Prevalence of Levels of Insured by CVD



**Levels of Insured**

Cardiovascular Disease			Estimate	95% Confidence Interval		Unweighted Count
				Lower	Upper	
Yes	% of Total	Insured	71.3%	66.7%	75.5%	1090
		Underinsured	13.1%	10.4%	16.4%	132
		Uninsured	15.6%	12.0%	20.0%	102
		Total	100.0%	100.0%	100.0%	1324
No	% of Total	Insured	65.4%	63.6%	67.2%	6843
		Underinsured	8.1%	7.1%	9.2%	660
		Uninsured	26.5%	24.8%	28.2%	1722
		Total	100.0%	100.0%	100.0%	9225

# Prevalence of CVD by Levels of Insured



## Cardiovascular Disease

Levels of Insured			Estimate	95% Confidence Interval		Unweighted Count
				Lower	Upper	
Insured	% of Total	Yes	8.5%	7.8%	9.3%	1090
		No	91.5%	90.7%	92.2%	6843
		Total	100.0%	100.0%	100.0%	7933
Underinsured	% of Total	Yes	12.2%	9.5%	15.6%	132
		No	87.8%	84.4%	90.5%	660
		Total	100.0%	100.0%	100.0%	792
Uninsured	% of Total	Yes	4.8%	3.6%	6.4%	102
		No	95.2%	93.6%	96.4%	1722
		Total	100.0%	100.0%	100.0%	1824

# The correct syntax



```
CSTABULATE
```

```
/PLAN FILE='I:\brfss_data\brfss08\State\cdcfinwt08.csaplan'
```

```
/TABLES VARIABLES=undinsur
```

```
/SUBPOP TABLE=cvd DISPLAY=LAYERED
```

```
/CELLS TABLEPCT
```

```
/STATISTICS CIN(95) COUNT
```

```
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

## Why Frequency with subpopulation instead of crosstabs?



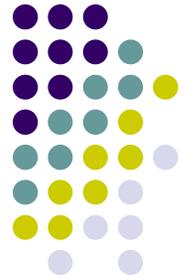
- It makes it easier to see what you're looking at
- Use of multiple subpopulations

# Add Race/Ethnicity



Levels of Insured							
Cardiovascular Disease	Race/Ethnicity			Estimate	95% Confidence Interval		Unweighted Count
					Lower	Upper	
Yes	White	% of Total	Insured	81.5%	76.7%	85.5%	859
			Underinsured	10.1%	7.2%	14.0%	69
			Uninsured	8.4%	5.6%	12.4%	49
			Total	100.0%	100.0%	100.0%	977
	Black	% of Total	Insured	74.5%	61.5%	84.2%	67
			Underinsured	19.1%	10.6%	31.9%	19
			Uninsured	6.4%	2.7%	14.4%	7
			Total	100.0%	100.0%	100.0%	93
	Hispanic	% of Total	Insured	48.3%	38.3%	58.5%	125
			Underinsured	18.6%	12.4%	27.1%	41
			Uninsured	33.0%	23.5%	44.2%	42
			Total	100.0%	100.0%	100.0%	208
	Other	% of Total	Insured	73.4%	47.2%	89.5%	30
			Underinsured	7.1%	1.9%	23.1%	3
			Uninsured	19.5%	5.8%	48.9%	4
			Total	100.0%	100.0%	100.0%	37

# Syntax for CVD by insurance and race/ethnicity



```
CSTABULATE
```

```
/PLAN FILE='I:\brfss_data\brfss08\State\cdcfinwt08.csaplan'
```

```
/TABLES VARIABLES=undinsur
```

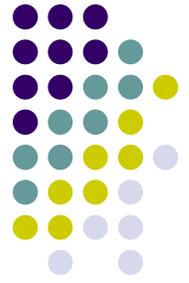
```
/SUBPOP TABLE=cvd BY raceeth DISPLAY=LAYERED
```

```
/CELLS TABLEPCT
```

```
/STATISTICS CIN(95) COUNT
```

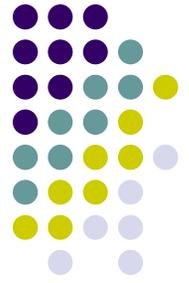
```
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

# Syntax for more complex analysis



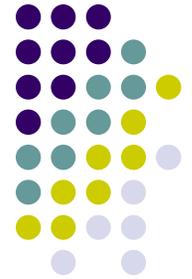
```
CSTABULATE  
  /PLAN FILE='I:\brfss_data\brfss08\State  
\cdcfinwt08.csaplan'  
  /TABLES VARIABLES=undinsur  
  /SUBPOP TABLE=cvd BY raceeth BY agegrp50  
DISPLAY=LAYERED  
  /CELLS TABLEPCT  
  /STATISTICS CIN(95) COUNT  
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

# Another weighted frequency ~ “disabil”



- Analyze – Complex Samples – Frequencies
- Check to make sure correct csaplan is still selected
- Highlight “Disability” and move it to frequency table box using arrow button
- Click on Statistics button
- Check to make sure “Table percent”, “Confidence Interval” and “Unweighted count” are still selected
- Click “continue”, click “okay”

# Frequency of Disability



## Disability

		Estimate	95% Confidence Interval		Unweighted Count
			Lower	Upper	
% of Total	yes	21.2%	20.0%	22.5%	2985
	no	78.8%	77.5%	80.0%	7701
	Total	100.0%	100.0%	100.0%	10686

# Syntax for frequency of disability



```
CSTABULATE
```

```
/PLAN FILE='I:\brfss_data\brfss08\State\cdcfinwt08.csaplan'
```

```
/TABLES VARIABLES=disabil
```

```
/CELLS TABLEPCT
```

```
/STATISTICS CIN(95) COUNT
```

```
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

# Insurance coverage for persons with disabilities



- Which is the population and which is the subpopulation?

# Insurance coverage for persons with disabilities syntax



```
CSTABULATE
```

```
/PLAN FILE='I:\brfss_data\brfss08\State\cdcfinwt08.csaplan'
```

```
/TABLES VARIABLES=undinsur
```

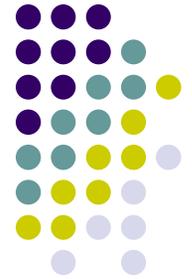
```
/SUBPOP TABLE=disabil DISPLAY=LAYERED
```

```
/CELLS TABLEPCT
```

```
/STATISTICS CIN(95) COUNT
```

```
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

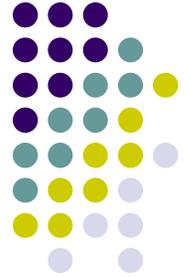
# Insurance coverage for persons with disabilities



**Levels of Insured**

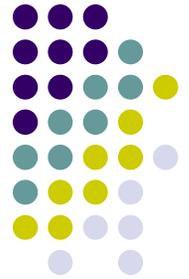
Disability			Estimate	95% Confidence Interval		Unweighted Count
				Lower	Upper	
yes	% of Total	Insured	62.8%	59.4%	66.1%	2191
		Underinsured	16.2%	14.0%	18.8%	408
		Uninsured	20.9%	17.9%	24.3%	368
		Total	100.0%	100.0%	100.0%	2967
no	% of Total	Insured	66.7%	64.7%	68.6%	5799
		Underinsured	6.5%	5.5%	7.6%	403
		Uninsured	26.9%	25.1%	28.7%	1467
		Total	100.0%	100.0%	100.0%	7669

# Insurance coverage for persons with disabilities by race/ethnicity



Levels of Insured							
Disability	Race/Ethnicity			Estimate	95% Confidence Interval		Unweighted Count
					Lower	Upper	
yes	White	% of Total	Insured	70.0%	66.2%	73.7%	1678
			Underinsured	14.2%	11.7%	17.3%	248
			Uninsured	15.7%	12.7%	19.3%	193
			Total	100.0%	100.0%	100.0%	2119
	Black	% of Total	Insured	65.1%	56.0%	73.2%	150
			Underinsured	16.9%	11.5%	24.1%	46
			Uninsured	18.0%	11.7%	26.7%	31
			Total	100.0%	100.0%	100.0%	227
	Hispanic	% of Total	Insured	45.1%	37.6%	52.7%	296
			Underinsured	17.7%	12.9%	23.8%	97
			Uninsured	37.2%	29.2%	46.0%	135
			Total	100.0%	100.0%	100.0%	528
	Other	% of Total	Insured	55.9%	38.2%	72.2%	49
			Underinsured	38.0%	22.0%	57.1%	15
			Uninsured	6.1%	2.4%	14.7%	8
			Total	100.0%	100.0%	100.0%	72

# Insurance coverage for persons with disabilities by race/ethnicity syntax



```
CSTABULATE
```

```
/PLAN FILE='I:\brfss_data\brfss08\State\cdcfinwt08.csaplan'
```

```
/TABLES VARIABLES=undinsur
```

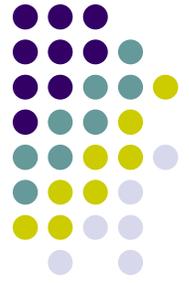
```
/SUBPOP TABLE=disabil BY raceeth DISPLAY=LAYERED
```

```
/CELLS TABLEPCT
```

```
/STATISTICS CIN(95) COUNT
```

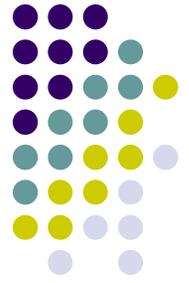
```
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

# Say we were interested in people under 65 only



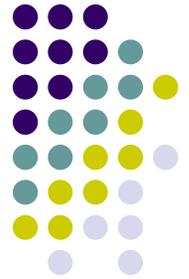
- Question is: what is the insurance status of persons with disabilities under the age of 65?
- Individuals who definitely do not qualify for Medicare although they may qualify for Medicaid or some Medicare programs.
- Sort the data by age and restrict the dataset to those under age 65

# Say we were interested in people under 65 only



- Verify that 7 & 9 are set to missing
- Select “Data” – “Sort cases” – select “What is your age?” – click “okay”
- Select “Data” – “Select cases” – select “if condition satisfied” – click “if...” – highlight “What is your age” and click the arrow to move it to the box – click the “<” and “65” – click “continue” then “okay”

## What is the insurance status of persons with disabilities under the age of 65?



- Sort by “Geographic and Density combined” (“ststr”) and “Annual Sequence number” (“seqno”)
- Rerun frequency of insurance coverage of people with disabilities
- Remember to return to “Data” and select “all cases”
- Sort by ststr and seqno again

# What is the insurance status of persons with disabilities under the age of 65?



Levels of Insured

Disability			Estimate	95% Confidence Interval		Unweighted Count
				Lower	Upper	
yes	% of Total	Insured	54.0%	49.8%	58.1%	988
		Underinsured	18.2%	15.3%	21.6%	295
		Uninsured	27.8%	23.8%	32.1%	341
		Total	100.0%	100.0%	100.0%	1624
no	% of Total	Insured	63.2%	61.0%	65.3%	3839
		Underinsured	6.8%	5.7%	8.1%	335
		Uninsured	30.0%	28.0%	32.1%	1394
		Total	100.0%	100.0%	100.0%	5568

# Insurance coverage for persons with disabilities of all ages



**Levels of Insured**

Disability			Estimate	95% Confidence Interval		Unweighted Count
				Lower	Upper	
yes	% of Total	Insured	62.8%	59.4%	66.1%	2191
		Underinsured	16.2%	14.0%	18.8%	408
		Uninsured	20.9%	17.9%	24.3%	368
		Total	100.0%	100.0%	100.0%	2967
no	% of Total	Insured	66.7%	64.7%	68.6%	5799
		Underinsured	6.5%	5.5%	7.6%	403
		Uninsured	26.9%	25.1%	28.7%	1467
		Total	100.0%	100.0%	100.0%	7669

# Why do we sort by ststr & seqno?



- “ststr” combines state, stratum & density variables
- “seqno” is the annual unique identifier for each record
- Data should be sorted by design variables according to stages of sample selection
- An improperly sorted file affects the estimates of variance and will produce unpredictable results

# Why.....



- ...do I need to use appropriate software?
- ...do I need a sample size of at least 50?
- ...do you suppress to 50?
- ... can't I give out preliminary estimates?
- ...do I assume WR replacement when it's really not?
- ...is there only one record/PSU per cluster?

# Questions?



# Technical Assistance is available



- [BRFSS@dshs.state.tx.us](mailto:BRFSS@dshs.state.tx.us)
- Jennifer Haussler – 512-458-7111 ext.2564
- Anna Vincent – ext. 2520
- Rebecca Wood – ext. 6579
- Michelle Cook – ext. 6593