Minnesota Population Center
Training and Development

NAPP Extraction and Analysis

Exercise 1

OBJECTIVE: Gain an understanding of how the NAPP dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the NAPP dataset to explore historical demographic characteristics of Iceland.

1/8/2013

NAPP Training and Development

Research Questions

What were the most common occupations in Iceland in 1801 and 1901? Were farm households more likely to have more generations living together? How frequent was immigration into Iceland between 1801 and 1901?

Objectives

- Create and download a NAPP data extract
- Decompress data file and read data into SAS
- Analyze the data using sample code
- Validate data analysis work using answer key

NAPP Variables

- OCCHISCO: HISCO occupation classification
- FARMIPUM: Farm household by 19th century definition
- NUMGEN: The number of generations in the household
- NAPPSTER: NAPP country of birth
- YRIMMIG: Year of immigration to Iceland
- COUNTYIS: Iceland county

SAS Code to Review

Code	Purpose
proc freq;	Begins a frequency procedure
proc means;	Begins a means procedure, returns the mean value of a variable
tables	Required syntax to display frequencies
where	Selects only specified cases to include in a procedure

Review Answer Key (page 7)

Common Mistakes to Avoid

1 Not fully decompressing the data

2 Giving the wrong filepath to indicate the dataset

3 Forget to close a procedure with "run;"

4 Forget to terminate a command with a semicolon ";"

Registering with NAPP

Go to http://www.nappdata.org/napp/, click on User Registration & Login, and apply for access. On login screen, enter email address and password and submit it!

Step 1

Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box. Check the boxes for the 1801 and 1901 Iceland samples. Click the Submit sample selections box
- Using the drop down menu or search feature, select the following variables:

  OCCHISCO: HISCO occupation classification

  FARMIPUM: Farm household by 19th century definition

  NUMGEN: The number of generations in the household

  NAPPSTER: NAPP country of birth

  YRIMMIG: Year of immigration to Iceland

  COUNTYIS: Iceland county

- Click the green VIEW CART button under your data cart
- Review variable selection. Click the green Create Data Extract button
- Review the 'Extract Request Summary' screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download.
- To get to the page to download the data, follow the link in the email, or follow the Download and Revise Extracts link on the homepage.

•••

Step 2

Request the Data

Getting the data into your statistics software

The following instructions are for SAS.

Step 1

Download the Data

Go to http://www.nappdata.org/napp/ and click on Download or Revise Extracts

- Right-click on the data link next to extract you created
- Choose "Save Target As..." (or "Save Link As...")
- Save into "Documents" (that should pop up as the default location)
- Do the same thing for the SAS link next to the extract

•••

Step 2

Decompress the Data

- Find the "Documents" folder under the Start menu
- Right click on the ".dat.gz" file
- Use your decompression software to extract here
- Double-check that the Documents folder contains three files starting "napp_000..."
- Free decompression software is available at http://www.irnis.net/soft/wingzip/

•••

Step 3

Read in the Data

- Open the "napp_000##.sas" file
- In the do file window, change the first line from "libname IPUMS '.'" to "libname IPUMS \\Documents...;" using the file directory where you saved your data files
- After "filename ASCIIDAT", enter the full file location, ending with "napp_000##.dat";
- Choose Submit under the Run file menu

Analyze the Sample – Part I Frequencies of OCCHISCO

Section 1

Analyze the Data

A) On the website, find the codes page for the OCCHISCO variable. Go to the Comparability tab and find how individuals were coded who were too young to work in Iceland 1801 and 1901.

_____

B) What were the 3 most common occupations in Iceland in 1801?_____

C) What about 1901?_____

```
proc freq order = freq;
            tables occhisco* year;
run;
```

• • •

Note: The "order = freq" option orders the results by descending frequency.

Note on Weights

Using weights (PERWT)

In other data projects, you might be familiar with using weights to make your analysis more representative of the entire population. However, because the Iceland samples are already a 100 percent sample, the weight for each person is always one. If you compare multiple countries, however, you'll need to use the PERWT weight. To learn more about using weights, see the NAPP data exercise 2.

Analyze the Sample – Part II Relationships in the Data

Section 1

Analyze the Data

A) Go to the codes page for the variable FARMIPUM. What are the codes for this variable? _____

B) Which two counties in Iceland have the lowest proportion of farm households in 1901? _____

```
proc freq;
        tables countyis*farmipum;
        by year;
run;
```

C) What is the average number of generations in an Icelandic household in 1901? _____

```
proc means;
            where year = 1901.
            var numgen;
run;
```

Now we'll use two different ways of testing whether farm households tend to have more generations

D) Is the mean of NUMGEN different between farms and non-farms in 1901? _____

```
proc means;
            where year = 1901.
            var numgen;
            class farmipum;
run;
```

E) Does being a farm household make a family more likely to live with more generations in 1901? Is this significantly significant?

```
proc reg;
          model numgen = farmipum
run;
```

Analyze the Sample – Part III Frequencies in the Data

Section 1

Analyze the Data

A) Go to the Universe tab for YRIMMIG. What is the universe for YRIMMIG in Iceland 1901? _____

B) What are the codes for "Unknown" and "Not in Universe"? To whom does "Not in Universe" apply?

_____

C) How many people were immigrants from Norway and Denmark living in Iceland in 1901?

```
proc freq;
        tables nappster;
        where year = 1901;
run;
```

D) What years did the majority of these immigrants move to Iceland? _____

...

Complete! Check your Answers!

```
proc freq;
        tables yrimmig*nappster;
        where year = 1901;
        where nappster = 4 | nappster = 7;
run;
```

ANWERS: Analyze the Sample – Part I Frequencies of OCCHISCO

Section 1

Analyze the Data

A) On the website, find the codes page for the OCCHISCO variable. Go to the Comparability tab and find how individuals were coded who were too young to work in Iceland 1801 and 1901.
Unknown/No occupation

B) What were the 3 most common occupations in Iceland in 1801
General Farmers, Servants nfs, Farmer and Fisherman

C) What about 1901? **Servants nfs, Farm workers, Fishermen**

```
proc freq order = freq;
            tables occhisco* year;
run;
```

• • •

Note: The "order = freq" option orders the results by descending frequency.

Note on Weights

Using weights (PERWT)

In other data projects, you might be familiar with using weights to make your analysis more representative of the entire population. However, because the Iceland samples are already a 100 percent sample, the weight for each person is always one. If you compare multiple countries, however, you'll need to use the PERWT weight. To learn more about using weights, see the NAPP data exercise 2.

ANSWERS: Analyze the Sample – Part II Relationships in the Data

Section 1

Analyze the Data

A) Go to the codes page for the variable FARMIPUM. What are the codes for this variable? **Non-farm: 1; Farm: 2**

B) Which two counties in Iceland have the lowest proportion of farm households in 1901? **Gullbringusýsla 20.42%, Reykjavíkurkaupstaður 3.65%**

```
proc freq;
        tables countyis*farmipum;
        by year;
run;
```

C) What is the average number of generations in an Icelandic household in 1901? **2.06 generations**

```
proc means;
            where year = 1901.
            var numgen;
run;
```

D) Is the mean of NUMGEN different between farms and non-farms in 1901? **Yes, the difference is 0.149 generations.**

```
proc means;
            where year = 1901.
            var numgen;
            class farmipum;
run;
```

E) Does being a farm household make a family more likely to live with more generations in 1901? Is this significantly significant?

Using to the model, the difference is significant at the 0.001 level.

```
proc reg;
            model numgen = farmipum
run;
```

ANSWERS: Analyze the Sample – Part III Frequencies in the Data

Section 1

Analyze the Data

A) Go to the Universe tab for YRIMMIG. What is the universe for YRIMMIG in Iceland 1901? **All foreign-born persons.**

B) What are the codes for "Unknown" and "Not in Universe"? To whom does "Not in Universe" apply? **Unknown: 0000; NIU: 9999. NIU applies to anyone born in Iceland, or not-foreign born.**

C) How many people were immigrants from Norway and Denmark living in Iceland in 1901? **207 from Norway, 110 from Denmark**

```
proc freq;
        tables nappster;
        where year = 1901;
run;
```

D) What years did the majority of these immigrants move to Iceland? **1901 for Norway (95 people); 1892/1901 for Denmark (5)**

```
proc freq;
        tables yrimmig*nappster;
        where year = 1901;
        where nappster = 4 | nappster = 7;
run;
```