

Data

- Some definitions
- Scales of Measurement
 - Nominal scale
 - Ordinal scale
 - Interval scale
 - Ratio scale
- Types of variables
 - Qualitative
 - Quantitative
 - Cross section data
 - Time series data
 - Pooled cross section data
 - Longitudinal data

Some Definitions

Element – the entities on which the data are collected.

Example: In this class we will use data that have persons or firms as elements.

Variable – a characteristic of interest for the elements

Observations – the set of measurements obtained for a particular element

Data set – all the data collected in a particular study.

Example: A partial version of the 2007 March Current Population Survey (CPS) is up on the Data page of the course website. There are three years of data available 1989, 2000, and 2007.

The **elements** of this data are persons, but via manipulation it the element could easily be changed to families.

The data set contains 25 **variables** and 206,552 **observations**. These variables are of two types.

1. Variables which are useful for determining each person's poverty status.
2. Demographic and background variables.

If you type **describe** in the STATA command prompt after opening the STATA data file you get a listing of the variables in the data set along with a short description of each variable

```

■ Contains data from C:\Documents and Settings\Geoffrey\My Documents\Work\Classes\PA818_2008\Data\poverty07.dta
■   obs:      206,552
■   vars:      21                8 Sep 2008 10:12
■   size:    10,947,256 (89.3% of memory free)
■ -----
■           storage  display  value
■ variable name  type    format   label      variable label
■ -----
■ region         byte    %8.0g    Region of the country;
■                1=Northeast, 2=Midwest,
■                3=South, 4=West
■ urbanindc     byte    %8.0g    Urban status; 1=Central City,
■                2=Other Metro Area, 3=Rural,
■                4=Unclassified
■ famsize       byte    %8.0g    Number of people in the family
■                unit
■ cutoff        long    %12.0g   Official poverty threshold
■ faminc1       long    %12.0g   Gross family cash income
■ age           byte    %8.0g    Age in years
■ ftfy          byte    %8.0g    Full-Time, Full-Year indicator;
■                1=FTFY, 0=not FTFT
■ ed            byte    %8.0g    Education level; 1=Less than
■                HS, 2=HS, 3=Some College, 4=BA+
■ famhead       byte    %8.0g    Family head indicator; 1=if
■                family head, 0=not family head
■ fsv           float   %9.0g    Value of family's foodstamps
■ edh1          byte    %8.0g    Education level of the family
■                head
■ wstath        byte    %8.0g    Employment Status of the family
■                head; 1=Not employed, 2=Employed
■                not FTFY, 3=FTFY
■ famtax1       double  %10.0g   Family tax liability (including
■                credits) as imputed by NBER's
■                TAXSIM
■ famtax2       long    %12.0g   Family tax liability (including
■                credits) as imputed by CPS
■ family        byte    %8.0g    1=if person is in a family,
■                0=if person is not in a family
■ famtype       byte    %8.0g    Type of family; 1=Married
■                couple, 2=Female Head, 3=Male
■                Head
■ nrace         byte    %8.0g    Race-Ethnicity; 1=white
■                (non-Hisp), 2=Black (non-Hisp),
■                3=Hispanic, 4=Other
■ nemployed     byte    %8.0g
■ fweight       float   %9.0g    CPS family weight
■ pweight       float   %9.0g    CPS person weight
■ year          float   %9.0g    Calendar year for income data
■ -----
```

Scales of Measurement

Data is measured in one of the following scales

- **Nominal scale** – data for a variable consist of labels or names used to identify an attribute of the element and there is no natural ordering of biggest to smallest or worst to best.

Example: The variables region, nrace, famtype, and urbanindc are all measured on a nominal scale.

Scales of Measurement (cont.)

- **Ordinal scale** – data for a variable consist of labels or names used to identify an attribute of the element and there is a natural ordering from biggest to smallest or worst to best.

Example: The variables ed, ftfy, and wstath are measured on an ordinal scale.

Scales of Measurement (cont.)

- **Interval scale** – displays the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measurement.

Example: The variables cutoff, faminc1, age, famsize, famtax1, and famtax2 are measured according to an interval scale.

Scales of Measurement (cont.)

- **Ratio scale** – a variable measured in a ratio scale has all the properties of interval data plus the additional requirement that the ratio between two values be meaningful. The key property of ratio scale data is that there is an absolute zero which implies the lack of the variable being measured.

Example: fsv and famsize are measured at the ratio level – a zero value for fsv means the family didn't get foodstamps – and by definition family units of size zero do not exist.

Example: If we are measuring temperature on the Calvin scale there is a zero temperature which will never be observed in the data ($0^{\circ}\text{K} = -273^{\circ}\text{C}$), thus temperature measured in degrees Calvin is at the ratio level. Temperature measured in the Celsius or Fahrenheit scales is measured at the interval level.

Qualitative versus Quantitative Data

- **Qualitative data (categorical data)** – data that include labels or names used to identify an attribute of each element – nominal and ordinal data.

Example: All of our nominal and ordinal data is qualitative

- **Quantitative data** – data that correspond to actual values that indicate how much or how many.

Example: faminc1 is quantitative.

Other Ways to Classify Data

Cross section data – consist of observations from a cross section of individuals or firms which is all collected at approximately the same time.

Example: Our CPS data is cross section data. In collecting this data the Census Bureau surveys a sample US non-institutional population in March-April of the of each year.

Time series data – data collected over several time periods. Time is a variable in time series data.

Longitudinal data – data on the same individuals or firms is collected over several time periods.

Descriptive Statistics

Descriptive Statistics – graphical, tabular, or numeric summaries of the data.

How data are best summarized depends on whether they are qualitative or quantitative (sort of)

Data

```
graph TD; Data[Data] --> Qual[Qualitative Data]; Data --> Quant[Quantitative Data]; Qual --> Qual_Graphical[Graphical Methods]; Qual --> Qual_Tabular[Tabular and Numeric Methods]; Quant --> Quant_Graphical[Graphical Methods]; Quant --> Quant_Tabular[Tabular and Numeric];
```

Qualitative Data

Graphical Methods

- Pie Chart
- Bar Graph

Tabular and Numeric Methods

- Frequency distribution
- Relative frequency distribution

Quantitative Data

Graphical Methods

- Histograms
- Other Graphs

Tabular and Numeric

- Relative frequency distributions
- Measures of location
- Measures of dispersion

Numeric and Tabular Methods for Describing Qualitative Data

- **Frequency distribution** -a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes.
- **Relative (or percent) frequency distribution**
 - similar to a frequency distribution except relative frequencies (or the percent in each category) are shown instead of counts.

Frequency Distributions w/ STATA

It is super easy to generate a frequency distribution using STATA.

Example: Generate a frequency distribution for the nrace variable.

To do this I simply open a STATA data file and type

```
tab nrace
```

in the STATA command prompt (or as a line in a *.do file) and the following appears in the STATA results window:

nrace	Freq.	Percent	Cum.
1	132,974	64.38	64.38
2	22,653	10.97	75.35
3	34,171	16.54	91.89
4	16,754	8.11	100.00
Total	206,552	100.00	

Frequency Distributions and STATA

(cont.)

If the data contains sampling weights (used to adjust for sampling procedures) and you want to use the summaries to make inferences about the underlying population parameters (i.e. the fraction of whites, blacks, Hispanics and others in the non-institutional US population), then the simple frequency calculations available with the STATA `tab` command are not sufficient.

In order to get proper relative frequencies from a data set that contains sampling weights we have to do a bit of data manipulation.

1. Create a new variable for each category of the original qualitative variable that will equal 1 if the element is in that category and 0 otherwise. This type of variable is often referred to as a binary variable, a dummy variable, or an indicator variable.
2. Summarize these new variables using the weight option.

Frequency Distributions w/ STATA

(cont.)

The following sequence of commands

```
gen white=cond(nrace==1,1,0)
gen black=cond(nrace==2,1,0)
gen hisp=cond(nrace==3,1,0)
gen other=cond(nrace==4,1,0)
sum white black hisp other [w=pweight]
```

produces the table below. Note that the weighted means in the table below can be interpreted as relative frequencies and are slightly different from the percentage frequencies produced by the **tab** command.

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
white	206552	296697352	.661124	.4733288	0	1
black	206552	296697352	.1212515	.3264201	0	1
hisp	206552	296697352	.1511127	.3581596	0	1
other	206552	296697352	.0665118	.2491752	0	1

Pie Charts

- Pie charts can be a useful way of summarizing qualitative data.

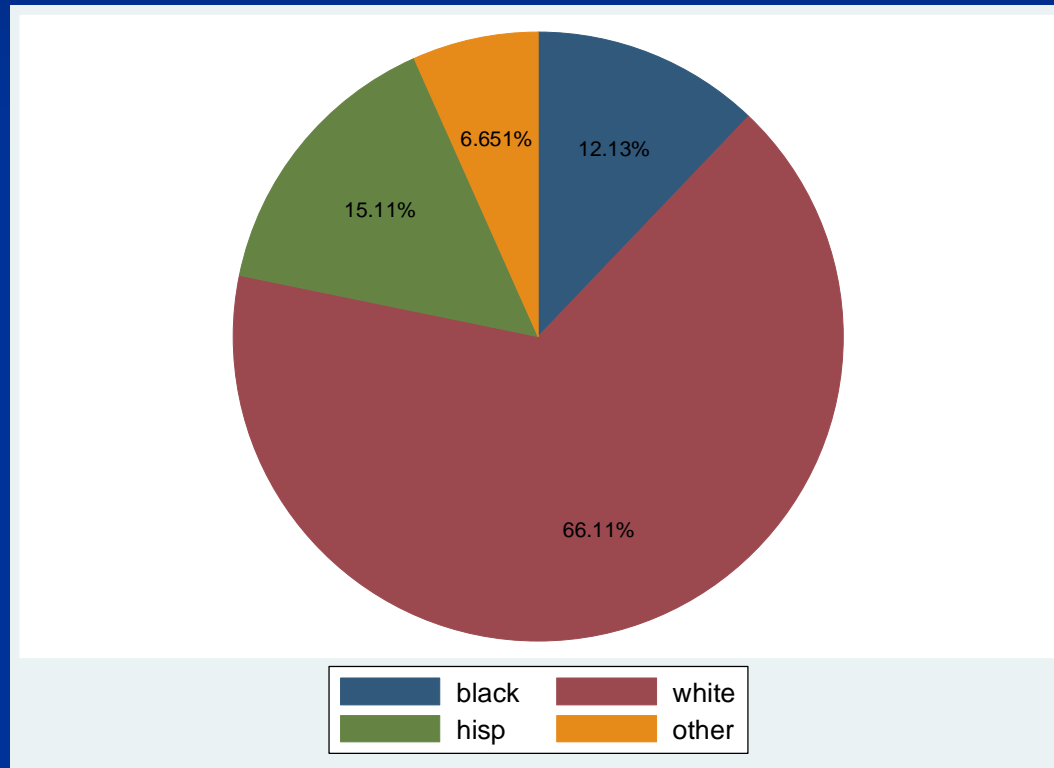
It is pretty easy to make a pie chart in STATA, but somewhat hard to make it look “presentation quality” and get all the labels and colors the way that you want them. Therefore, I’m going to recommend that you do your calculations to make a pie chart in STATA, but that you actually make the chart in MS-Excel.

- To make a pie chart the first step is to create a new variable for each category of the original qualitative variable that will equal 1 if the element is in that category and 0 otherwise. Since we have already done this for the race variable we are now ready to proceed with the command

```
graph pie black white hisp other, plabel(_all percent)
```

Which produces the following chart

Pie Chart of Racial-Ethnic Groupings in the 2007 CPS (created with STATA)



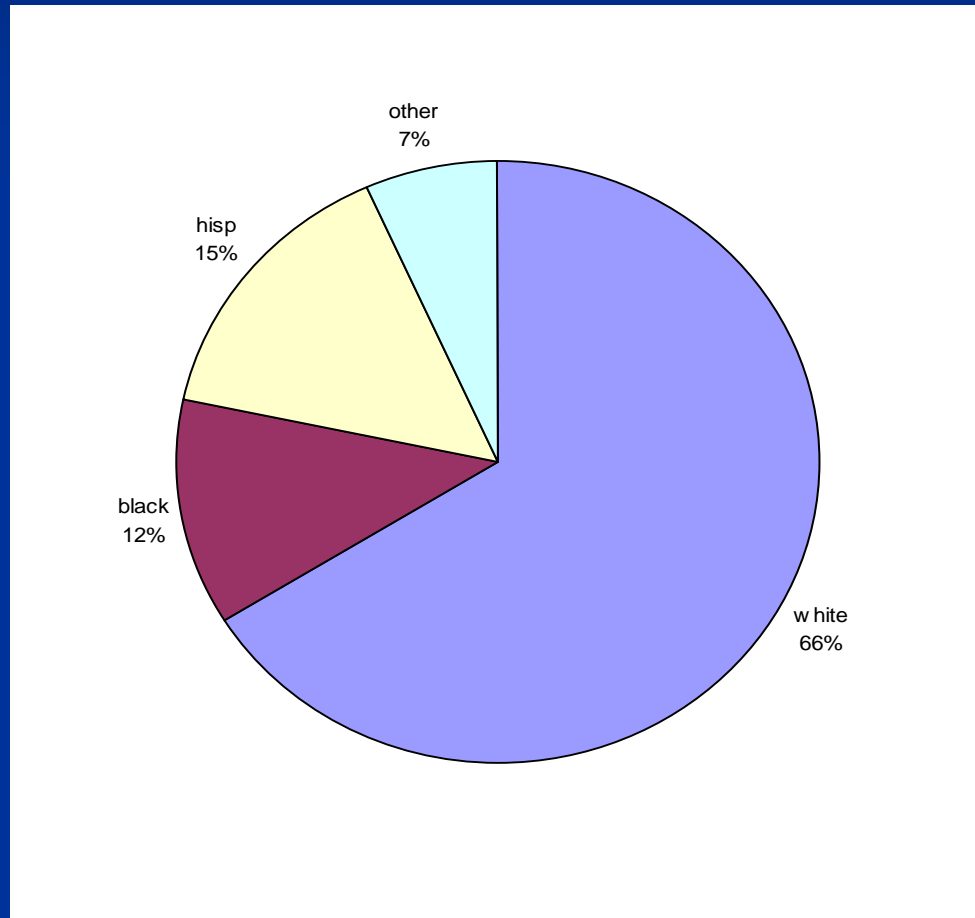
Pie Charts w/ Excel

If we wanted to do the same thing in excel simply cut and past the table of means from above into WordPad and eliminate the column headers. When you are done you should have something that looks like this

white		206552	296697352	.661124	.4733288	0	1
black		206552	296697352	.1212515	.3264201	0	1
hisp		206552	296697352	.1511127	.3581596	0	1
other		206552	296697352	.0665118	.2491752	0	1

The next step is to save the WordPad file as simple *.txt document somewhere on your computer. Then you want to open this file with MS-Excel, making sure to select *.txt in the open file dialog box) and highlight the Variable and Mean columns. Once you have done this go to the **Insert** menu in Excel and select insert chart. You will be given a menu of charts from which you can select pie charts. Follow the instructions in the wizard and you will end up with something like the following that can be cut and pasted into documents

Pie Chart of Racial-Ethnic Groupings in the 2007 CPS (created with Excel)



Bar Graphs w/ STATA

- Bar graphs can also be a useful way of summarizing qualitative data

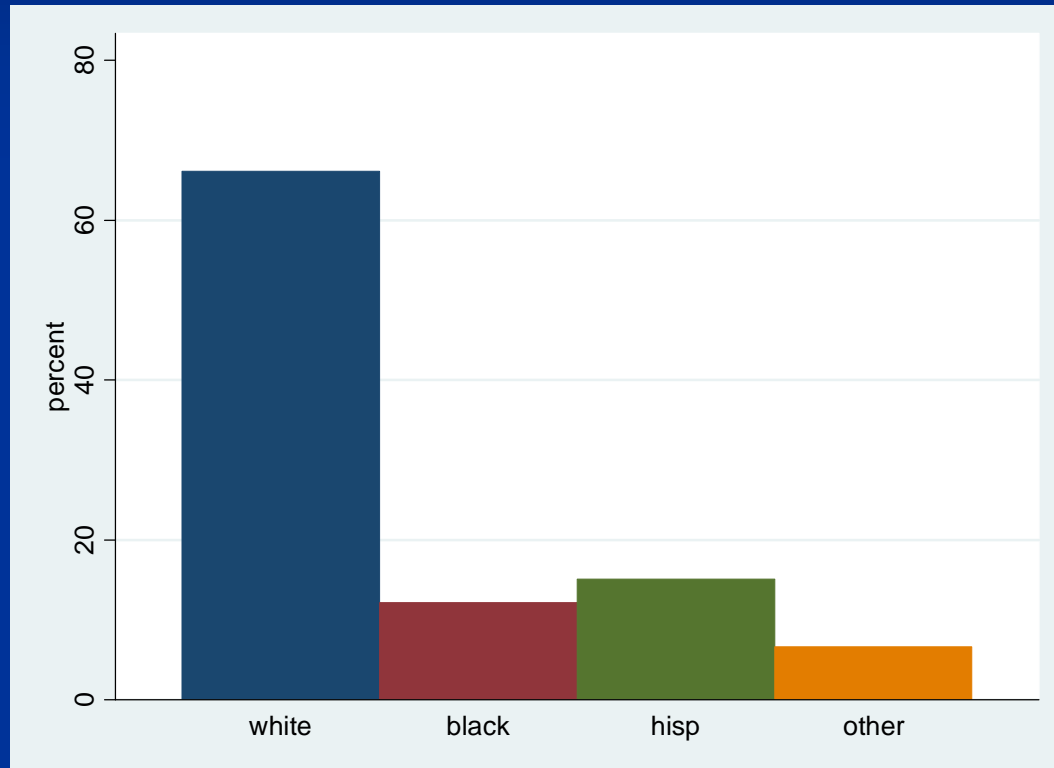
Like pie charts, bar graphs can be made in either STATA or Excel. I find it somewhat easier to use excel, but I'm going to show you how to do both.

- To make a bar of our racial-ethnic groupings in STATA simply type

```
graph bar (sum) white black hisp other [w=pweight], percent nolabel showyvars legend(off)
```

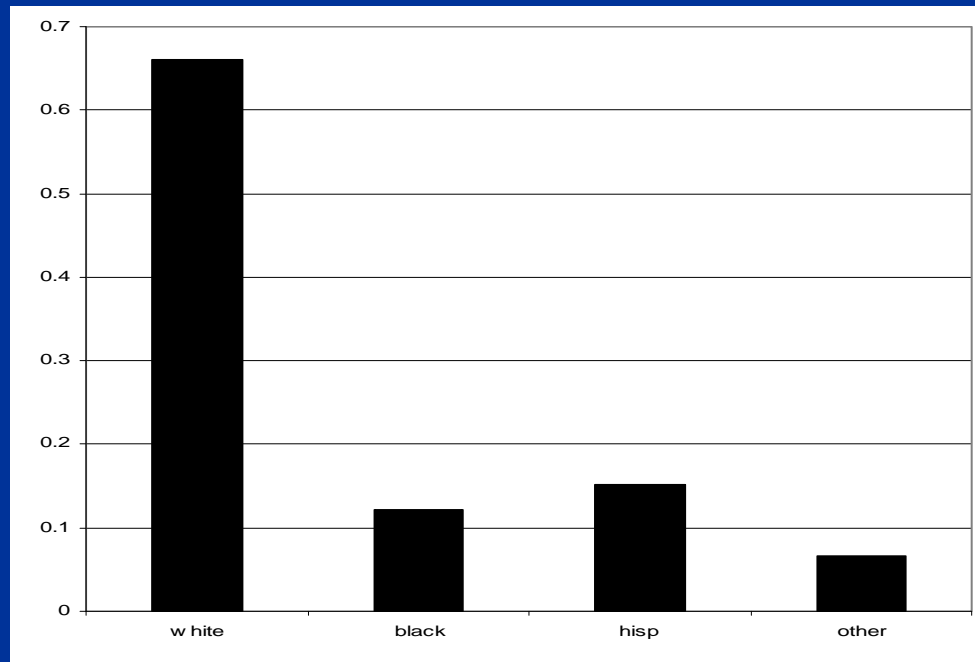
in the STATA command prompt and the following table will appear in another window

Bar Graph of Racial-Ethnic Groupings in the 2007 CPS



Bar Graphs w/ Excel

Bar graphs in excel are essentially the same as pie charts except within the chart wizard you select bar graph as apposed to pie chart. What you will end up with is something like the following which can be formatted in different ways and cut and pasted into documents



Numeric and Tabular Methods for Describing Quantitative Data

There are a number of ways to summarize quantitative data

- Frequency distribution
- Histogram
- Sample Statistics
 - Measures of Location
 - Mean
 - Median
 - Percentiles
 - Mode
 - Measures of dispersion and shape
 - Variance-standard deviation
 - Interquartile range
 - Skewness

Frequency Distributions and Histograms w/ STATA

Constructing a frequency distribution or histogram for quantitative data is more difficult than with qualitative data because the researcher has to specify

- The number of non-overlapping classes
- The width of each class
- The limits of each class

Thankfully there are some guidelines for choosing these parameters

Frequency Distributions and Histograms w/ STATA

- Determining the number of classes
 - The number of classes should be 5 and 20 depending on how much data you have – more observations should mean more classes.
 - Because we have a lot of data and a variable with a pretty large range a larger number of categories are needed. We will go with 20 to start

Frequency Distributions and Histograms w/ STATA

- Determining the width of each class

$$\text{approximate class width} = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$$

So for our data

$$\begin{aligned} \text{approximate class width} &= \frac{1,202,802 - (-16,993)}{20} \\ &= \frac{1,212,795}{20} = 60,990 \end{aligned}$$

Frequency Distributions and Histograms w/ STATA

- Determining the class limits

Class limits should be determined so that each observation belongs to one and only one class. We do this by determining the lower class limit - or the upper class limit. Once we do this the class width imposes the rest of the limits.

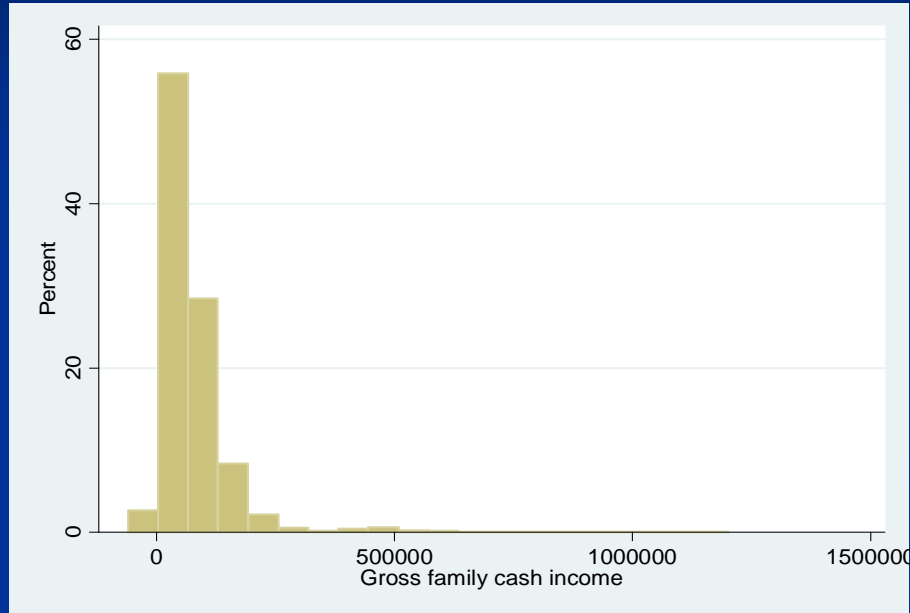
Example: Construct a histogram for our income data

We probably want 0 to be a divider between the first and second class so we will make -60,000 the lower limit to the first class. We want 20 bins or classes.

The following command creates a histogram of our family income variable in STATA.

```
hist faminc1, bin(20) start(-60000)  
percent
```

Histogram of Family Income Attempt #1



* This histogram isn't really that interesting or descriptive because many of the bins (classes) don't have very many observations, but at the bottom of the distribution we are grouping people together with very different incomes.

The solution to this problem is to eliminate some of the outliers from the data.

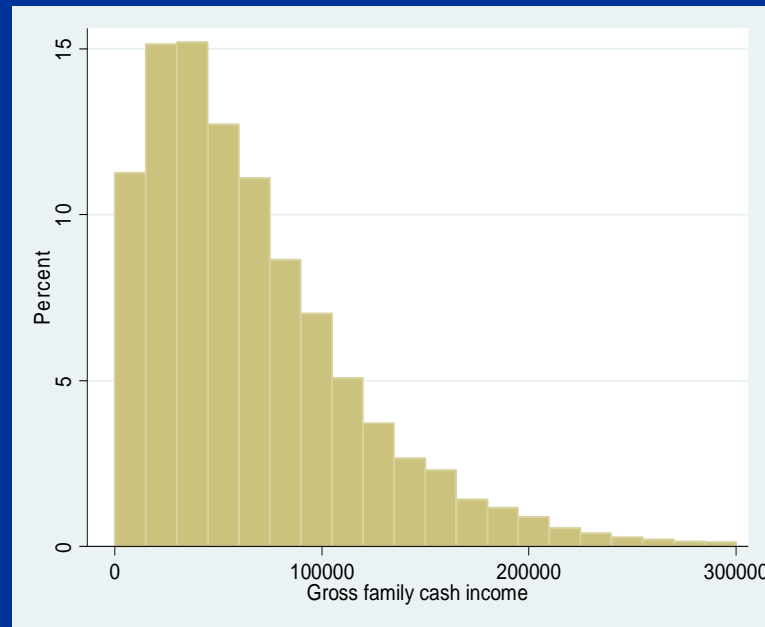
Less than 2% of the observations have levels of faminc1 less than \$0 or greater than \$300,000.

If we eliminate these observations and recalculate the class width we can make a new histogram

$$\begin{aligned}\text{approximate class width} &= \frac{300,000 - (0)}{20} \\ &= \frac{300,000}{20} = 15,000\end{aligned}$$

The following sequence of commands eliminate the outliers and produces a histogram that gives us a much better idea of the shape of the empirical distribution of family income.

```
gen out=cond(faminc1<0 | faminc1>300000,1,0)
drop if out==1
hist faminc1, bin(20) start(0) percent
```



Example: Construct a relative frequency distribution for our family income data

This is a little more difficult and tedious than making a histogram as we have to code the group that each element belongs to. We could do this via 20 commands (one for each group) or we could save some time and write a loop

The Loop

```
gen group=0
local i=0
while `i'<20 {
  local j=`i'+1
  replace group=cond(faminc1>=`i'*15000 & faminc1<15000*`j',`j',group)
  local i=`i'+1
}
tab group, gen(g)
```

These commands produce the following table

Frequency Distribution of Faminc1

group	Freq.	Percent	Cum.
1	13,764	16.84	16.84
2	15,577	19.06	35.91
3	13,184	16.13	52.04
4	9,700	11.87	63.91
5	7,730	9.46	73.37
6	5,708	6.98	80.35
7	4,557	5.58	85.93
8	3,115	3.81	89.74
9	2,321	2.84	92.58
10	1,607	1.97	94.55
11	1,371	1.68	96.23
12	850	1.04	97.27
13	679	0.83	98.10
14	525	0.64	98.74
15	331	0.41	99.14
16	244	0.30	99.44
17	158	0.19	99.64
18	121	0.15	99.78
19	93	0.11	99.90
20	82	0.10	100.00
Total	81,718	100.00	

Other Types of Graphs

At some point you might be asked to make a figure that shows a particular variable over time.

Example: Produce a graph that plots the poverty rate against the state monthly unemployment rate.

Sample Statistics

- Measures of Location
 - Sample mean
 - Median
 - Mode
 - Percentiles
- Measures of Variability
 - Range
 - Interquartile range
 - Sample variance

Measures of Location

- **Sample Mean**

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

- **Median** - the median is the middle observation.
 - If there are an odd number of data points the median is the true middle value
 - If there are an even number of data points the median is the average of the two middle values
- **Mode** – the most frequently occurring observation
- **Percentiles** – the p'th percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100-p) percent of the observations are greater than or equal to this value.
- **Quartiles** – the percentiles corresponding to 25, 50, and 75.
- **Quintiles** – the percentiles corresponding to 20, 40, 60, and 80.

Measures of Variability

- **Range**=largest value – Smallest value
- **Interquartile range**=the 75th percentile less the 25th percentile
- **Sample variance**

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

Measures of Variability (cont.)

- Sample standard deviation

$$S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2}$$

The standard deviation allows us to bound the limits of the distribution.

Chebyshev's Theorem - At least $\left(1 - \frac{1}{z^2}\right)$ of the data values must be within z standard deviations of the mean where z is any value greater than 1.

Example: For any distribution at least 75% of the data values lie within 2 standard deviations of the mean.

Sample Statistics and STATA

The STATA **sum** command can be used to generate most of these descriptive statistics. This command provides us with, or the ability to calculate, the sample mean, the variance, the sample standard deviation, percentiles, the and the interquartile range.

Example: Generate descriptive stats on faminc1

```
sum faminc1 [w=pweight], detail
```

```
-----  
                          Gross family cash income  
-----  
Percentiles      Smallest  
1%                0          -16993  
5%               7203       -16993  
10%              12792      -13058      Obs                206552  
25%              27000      -13058      Sum of Wgt.       296697352  
50%              52707  
75%              93135      Largest  
90%              144425     1145689  
95%              186144     1202802      Variance           6.06e+09  
99%              454528     1202802      Skewness           3.796878  
Kurtosis           25.72733  
-----
```

Sample Statistics and STATA

The **egen** command can also be very useful, particularly when you want to do calculations that involve summary statistics. The **egen** command allows you to generate a new variable that has a value equal to a specified summary statistic.

Example: Calculate the interquartile range.

The following sequence of commands produces the interquartile range.

```
egen famincp75=pctile(faminc1), p(75)
egen famincp25=pctile(faminc1), p(25)
gen intqrange=famincp75-famincp25
sum intqrange
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
intqrange	54366	65825	0	65825	65825