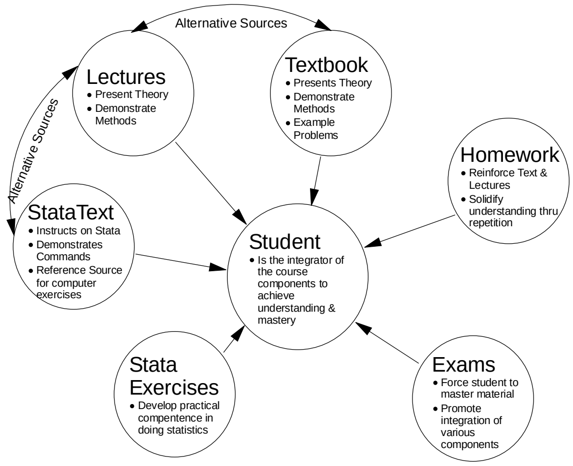
Economics 400

Professor Boone A. Turchi (Tur'shee)

"There are three kinds of lies--lies, damned lies and statistics"

Mark Twain, Autobiography

Components of the Economics 400 Course



Tips for Doing Well in Econ 400:

- Keep up with the readings & lectures.
- Do the assigned problems religiously -- hey, do even more! I generally assign odd-numbered book problems (*i.e.,* those without answers in book). Use even-numbered problems (answers at end of book) to get clues.
- Recopy your lecture notes, integrating text & other material
- Use my and the TA's office hours
- Take the computer exercises seriously and don't wait until the last minute to do them

Basic Concepts:

Population: is the set of *all* relevant objects for the problem under analysis.

Sample: is some part of the population which we use to make *inferences* about the population as a whole

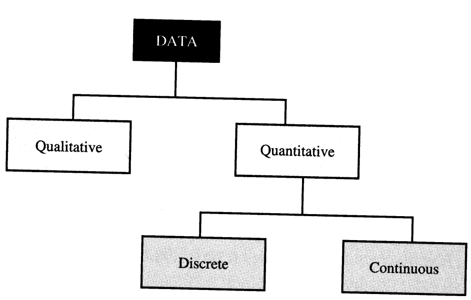
3 elements in the study of statistics:

- ✓ Describe available data using **descriptive techniques**
- ✓ Use **probability theory** as a framework for data analysis
- ✓ Use **statistical inference** to make inferences about unknown populations

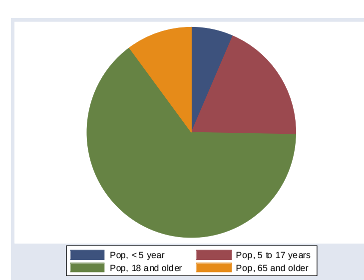
Why sample?

- ✓ Limited resources
- ✓ Limited data availability
- ✓ Destructive testing

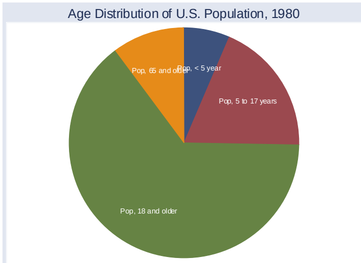
FIGURE 2.1
Types of data



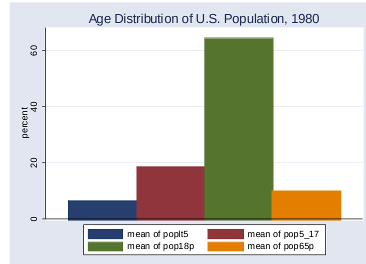
graph pie poplt5 pop5_17 pop18p pop65p



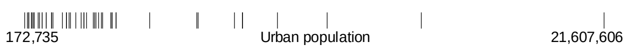
. graph pie poplt5 pop5_17 pop18p pop65, plabel(_all name, size(*1.0) color(white)) legend(off) title("Age Distribution of U.S. Population, 1980")



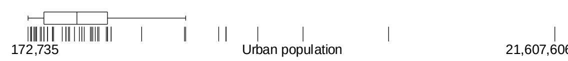
graph bar poplt5 pop5_17 pop18p pop65p, title("Age Distribution of U.S. Population, 1980") percent



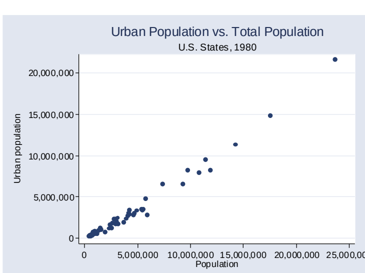
gr7 popurban, oneway



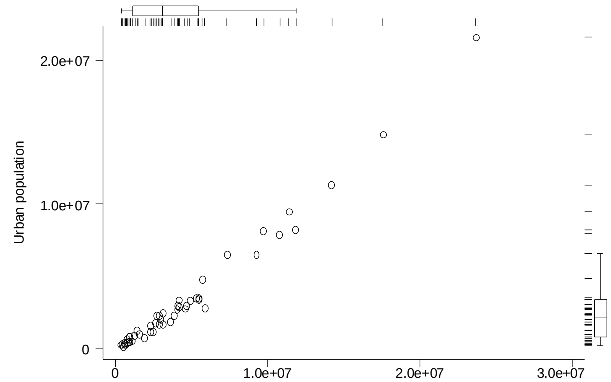
gr7 popurban, oneway box



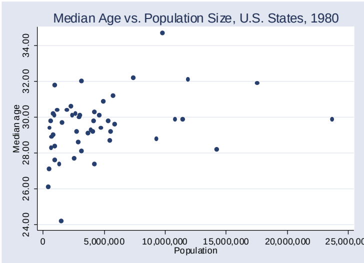
twoway (scatter popurban pop), ylabel(, angle(horizontal)) title(Urban Population vs. Total Population) subtitle("U.S. States, 1980")



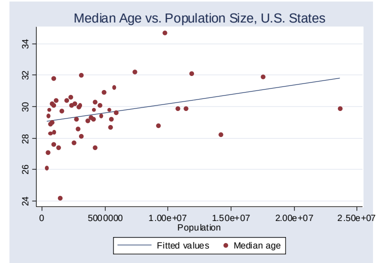
gr7 popurban pop, twoway oneway box xlabel ylabel



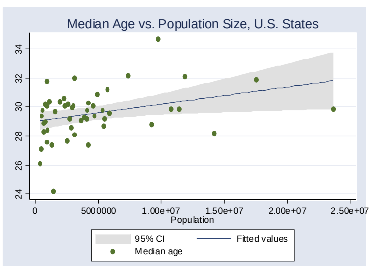
graph twoway scatter medage pop, title("Median Age vs. Population Size, U.S. States, 1980")



graph twoway (lfit medage pop) (scatter medage pop), title("Median Age vs. Population Size, U.S. States")

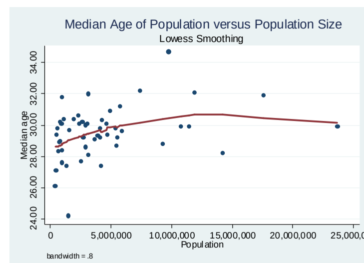


graph twoway (lfitci medage pop) (scatter medage pop), title("Median Age vs. Population Size, U.S. States")

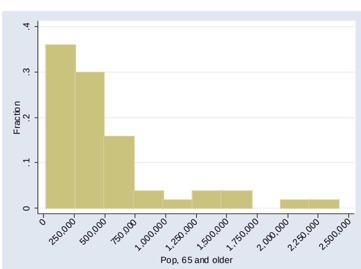



```
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    5.84
       Model |  15.2464213     1  15.2464213           Prob > F      =  0.0195
    Residual |  125.273564    48  2.60986591           R-squared     =  0.1085
-------------+------------------------------           Adj R-squared =  0.0899
       Total |  140.519985    49   2.8677548           Root MSE      =  1.6155

------------------------------------------------------------------------------
      medage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pop |   1.18e-07   4.89e-08     2.42   0.019     1.99e-08    2.17e-07
       _cons |   29.00548   .3179694    91.22   0.000     28.36616     29.6448
------------------------------------------------------------------------------

.
```

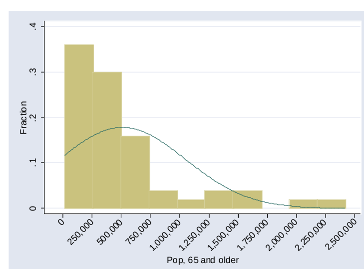
lowess medage pop, lineopts(lwidth(thick)) title(Median Age of Population versus Population Size) subtitle(Lowess Smoothing)



graph twoway histogram pop65p, fraction bin(10) xlabel(0 (250000) 2500000,angle(45))



histogram pop65p, fraction bin(10) xlabel(0 (250000) 2500000,angle(45)) norm



NB: Superimposing a normal curve over a histogram must be done with the "histogram" command, not the "graph twoway histogram" command, which cannot superimpose normal or kernel densities

```
stem  marriage

Stem-and-leaf plot for marriage (Number of marriages)

   0**** | 4437,5226,5361,6094,6868,7490,8336,8800,9251
   1**** | 1856,2040,3428,4239,6641,6958,7391
   2**** | 3004,4847,6048,6513,7474,7908
   3**** | 0223,2727,4917,7641
   4**** | 1111,3460,6273,6278,6509,6718,7728,9018
   5**** | 3915,4625,5794,7853,9175
   6**** | 0210
   7**** | 0638
   8**** | 6898
   9**** | 3673,9832
  10**** | 8344,9823
  11**** | 4333
  12**** |
  13**** |
  14**** | 4518
  15**** |
  16**** |
  17**** |
  18**** | 1762
  19**** |
  20**** |
  21**** | 0864
```

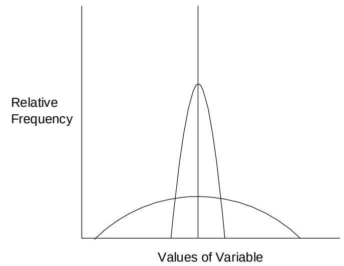
6,094

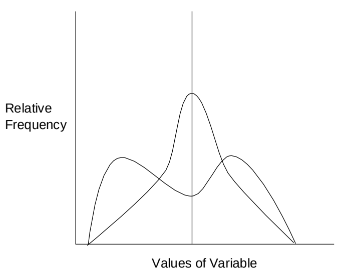
17,391

59,175

109,823

210,864





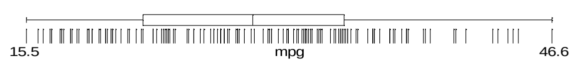

```
. summarize mpg,detail

                            mpg
-------------------------------------------------------------
-
      Percentiles      Smallest
 1%       16.2           15.5
 5%       17.6           16.2
10%       19.1           16.5       Obs                 154
25%       22.4           16.9       Sum of Wgt.         154

50%       28.9                      Mean           28.79351
                        Largest     Std. Dev.       7.37721
75%       34.3             44
90%         38           44.3       Variance       54.42323
95%       40.9           44.6       Skewness       .1115568
99%       44.6           46.6       Kurtosis       2.155094
```

n = 154
mean = 28.8
s = 7.4

```
      -2s
mpg             23            29.8        34.7
15.5            23            29.8        35
16.2            23.2          29.9        35.1
16.5            23.5          30          35.7
16.9            23.6          30          36
17              23.7          30.4        36
17              23.8          30.7        36
17.5            23.9          30.9        36
17.6            23.9          31          36
17.6  -1.5s     24            31          36.1
17.7            24.2          31          36.1
18.1            24.3          31.3        36.4  +1s
18.1            25            31.5        37
18.2            25.1          31.6        37
18.5            25.4          31.8        37
18.6            25.4          31.9        37.2
19.1            25.8          32          37.3
19.2            26            32          37.7
19.2            26.4          32          38
19.2            26.6          32.1        38
19.4            26.6          32.2        38
19.4            26.8          32.3        38
19.8            27            32.4        38.1
19.9            27            32.4        39
20.2            27            32.7        39.1
20.2            27            32.8        39.4  +1.5s
20.2            27.2          32.9        40.8
20.2            27.2          33          40.9
20.3            27.2          33.5        41.5
20.5            27.4          33.7        43.1
20.5            27.5          33.8        43.4  +2s
20.6            27.9          34          44
20.6            28            34          44.3
20.8            28            34.1        44.6
21.1  -1s       28            34.1        46.6
21.5            28.1          34.2
21.6            28.4          34.3
22              28.8  Mean    34.4
22.3            29            34.5
22.4            29.5          34.5
```

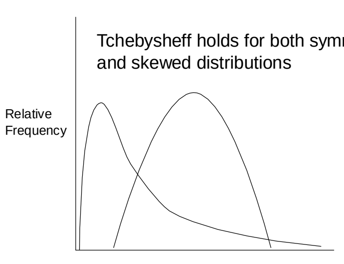
k	Tchebysheff Predicts	Actual
1.5	55.6%	89.6%
2	75%	97.4%

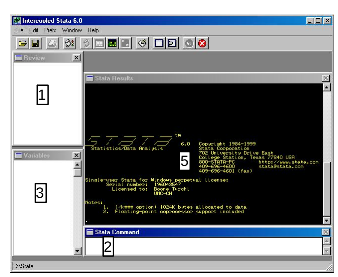
Miles per Gallon for 154 Cars: cardata.dta



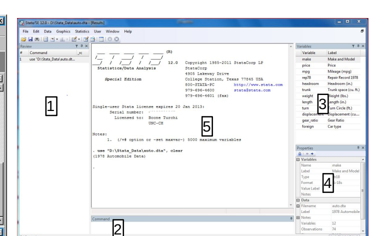
Knowing only the sample size, the mean and the standard deviation, one can **conservatively** predict the percentage of observations within *k* standard deviations of the mean using **Tchebysheff's Theorem**

Tchebysheff holds for both symmetrical and skewed distributions





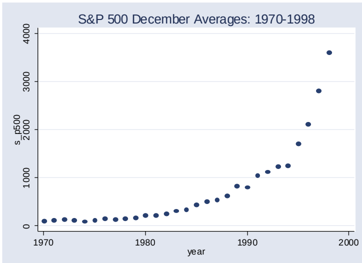
Traditional Stata Window



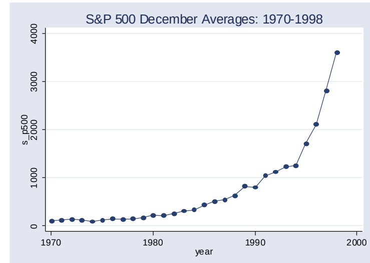
Stata Version 12 (& Later) Window

1. Review/Command Window: Lists previous session Commands
2. Stata Command Window: Enter a new command
3. Variables Window: Shows variables in the data set
4. Variable Properties Window: Shows properties of highlighted variable in Variables Window above.
5. Stata Results Window: Shows commands & results of commands

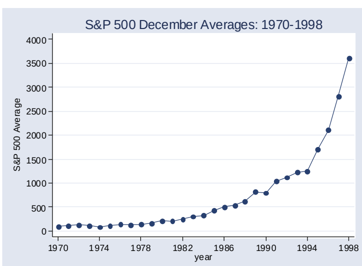
twoway scatter s_p500 year if(month==12), title("S&P 500 December Averages: 1970-1998")



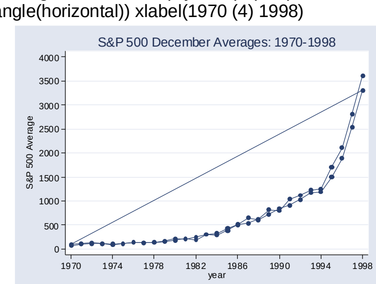
twoway scatter s_p500 year if(month==12), title("S&P 500 December Averages: 1970-1998") c(l)



twoway scatter s_p500 year if(month==12), title("S&P 500 December Averages: 1970-1998") c(l) ylabel(0 (500) 4000, angle(horizontal)) xlabel(1970 (4) 1998)



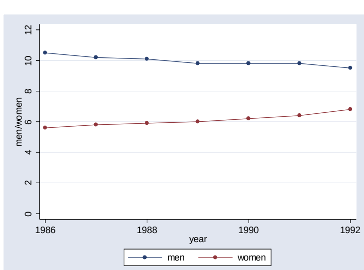
twoway connected s_p500 year if(month==12|month==6), title("S&P 500 December Averages: 1970-1998") ylabel(0 (500) 4000, angle(horizontal)) xlabel(1970 (4) 1998)



year	men	women	total
1986	10.5	5.6	16.1
1987	10.2	5.8	16
1988	10.1	5.9	16
1989	9.8	6	15.8
1990	9.8	6.2	16
1991	9.8	6.4	16.2
1992	9.5	6.8	16.3

Medical School Enrollment 1986 to 1992
(in Thousands)
(from Mendenhall, Beaver & Beaver, p. 63.)

Data view from Stata's Data Editor



Graphing two time series variables versus year.

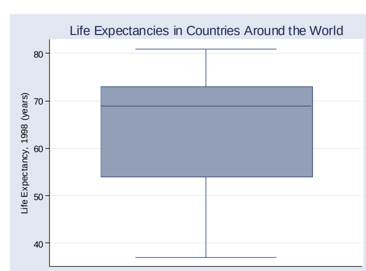
Try the following exercise:
- Start Stata and open the Data Editor
- Type in the 4 columns of data as shown
- Rename each variable from var1, etc. (Type "help rename" in the command line to see how to do this)
- Create the graph using command below
- Using the EC70tp.dot template submit a graph and the command used to create it.
- In particular, write a short paragraph explaining how "ylabel" & "xlabel" work in order to produce the labelling of the 2 axes.

graph twoway connected men women year, ylabel(0 2 to 12) xlabel(1986 1988 to 1992)

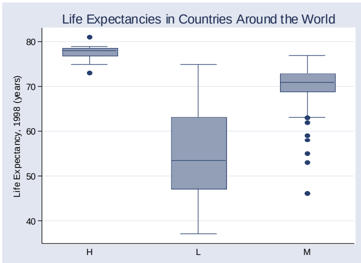
Exercise 2: Entering Data and Prettying Up Graphs (Revised Stata 10 Version)

label variable le98 "Life Expectancy, 1998 (years)"

graph box le98, title("Life Expectancies in Countries Around the World") ylabel(, angle(horizontal))



graph box le98, title("Life Expectancies in Countries Around the World") ylabel(, angle(horizontal)) over(inc)

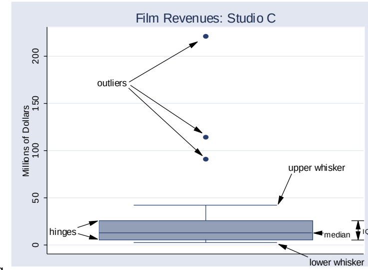


filmcrev
2.65 minimum (Lower Whisker)
3.55
3.55
3.57
3.64
4.21
4.95
4.99
5.3
5.73 Quartile 1 (hinge)
5.77
6.5
7.03
7.24
8.08
8.8
8.99
11.3
12.9 Quartile 2 (median)
14.2
15.3
17.8
21.6
21.8
24.3
24.9
25.6
26 Quartile 3 (hinge)
27.1
28.7
32
37.5
40.4
42.1 Upper Whisker
90.9
113.8
220.8 Maximum

Stata Constructs Boxplots according to the following procedure:

The data are for a film studio and represent the studio's revenues for all the pictures that it produced in a 3-year period.

First, sort the data from smallest to largest grossing picture.



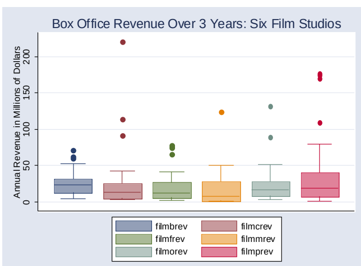
How to construct a boxplot:

- ✔ Compute the 5-number summary of the data: minimum, Quartile 1, Quartile 2, Quartile 3 Maximum
- ✔ Construct a box consisting of the middle quartiles
- ✔ Compute the **interquartile range (IQR): Q3-Q1 = 20.27**
- ✔ Multiply 1.5 times IQR = 30.4
- ✔ Add 30.4 to Q3: 26+30.4=56.4
- ✔ Find the last **adjacent value** that is less than 56.4
- ✔ Draw the upper whisker to that adjacent value
- ✔ Subtract 30.4 from Q1: 5.73 - 30.4 = -24.7
- ✔ Find the lowest adjacent value that is greater than -24.7
- ✔ Draw the lower whisker to that adjacent value
- ✔ Any data points outside the whiskers are **outliers.**

```
. summarize filmcrev, detail

                          filmcrev
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.65           2.65
 5%         3.55           3.55
10%         3.57           3.55       Obs                  37
25%         5.73           3.57       Sum of Wgt.          37

50%         12.9                      Mean           25.50135
                        Largest       Std. Dev.      40.28183
75%           26           42.1
90%         42.1           90.9       Variance       1622.626
95%        113.8          113.8       Skewness       3.567693
99%        220.8          220.8       Kurtosis       16.66538
```

graph box filmcrev, medtype(line) ytitle(Millions of Dollars) title("Film Revenues: Studio C")

graph box filmbrev filmcrev filmfrev filmmrev filmorev filmprev, ti("Box Office Revenue Over 3 Years: Six Film Studios") ytitle("Annual Revenue in Millions of Dollars")

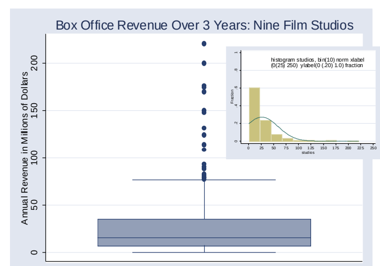



```
. use "D:\amifiles\econ70\data\film.dta", clear

. stack  filmbrev filmcrev filmfrev filmmrev filmorev filmprev filmtrev filmure
> v filmwrev, into(studios) clear wide

. summarize

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
      _stack |     441           5   2.584921          1          9
     studios |     304     26.5453   32.46495        .46      220.8
    filmbrev |      17    28.02353   20.79413        4.1       71.2
    filmcrev |      37    25.50135   40.28183       2.65      220.8
    filmfrev |      36    21.20889   21.69755       1.38       77.6
    filmmrev |      34    17.17265   23.79215       1.02      124.4
    filmorev |      28      24.815   27.86407       2.56      131.7
    filmprev |      37     38.3373   48.72081        .46      176.6
    filmtrev |      28    27.10071   28.48916        .86      150.4
    filmurev |      38    25.27421   33.91648       2.24      200.1
    filmwrev |      49    29.99776    28.1359       1.46      148.1

. describe

Contains data
  obs:           441
 vars:            11
 size:        37,926 (95.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
_stack          int    %8.0g
studios         double %10.0g
filmbrev        double %10.0g
filmcrev        double %10.0g
filmfrev        double %10.0g
filmmrev        double %10.0g
filmorev        double %10.0g
filmprev        double %10.0g
filmtrev        double %10.0g
filmurev        double %10.0g
filmwrev        double %10.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved

graph box  studios, ti("Box Office Revenue Over 3 Years: Nine Film Studios")
ytitle("Annual Revenue in Millions of Dollars")


 .  summarize  studios, detail

                          studios
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1.05            .46
 5%         2.24            .86
10%         3.68           1.02       Obs                 304
25%         7.29           1.05       Sum of Wgt.         304

50%           15                      Mean            26.5453
                        Largest       Std. Dev.      32.46495
75%           35            175
90%         59.5          176.6       Variance       1053.973
95%         81.1          200.1       Skewness       2.966599
99%          175          220.8       Kurtosis       13.95872
```

Stacking variables to create a new variable that includes all the values from a set of other variables:

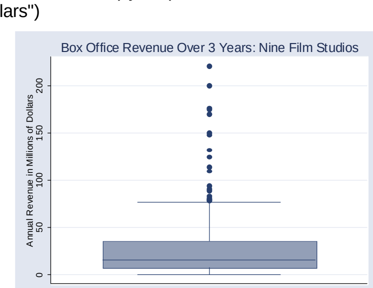
We start with a dataset, "film.dta" that has 9 variables, each of which shows the gross revenue from each picture produced by each of 9 film studios in a three year period.

We want to produce a boxplot that shows the distribution of gross revenues for all 9 studios. Use the stack command to concatenate all 9 variables ("filmbrev" through "filmwrev") into a new variable "studios". Note that "studios" has 304 non-missing values but the entire data set now has 441 rows. So, "studios" has 137 missing values.

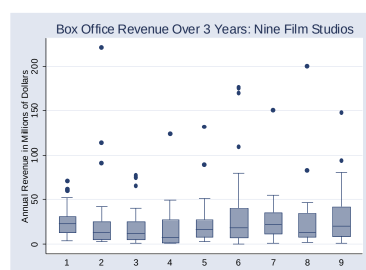


stack filmbrev filmcrev filmfrev filmmrev filmorev filmprev filmtrev filmurev filmwrev , into(studios) clear wide

graph box studios, ti("Box Office Revenue Over 3 Years: Nine Film Studios") ytitle("Annual Revenue in Millions of Dollars")

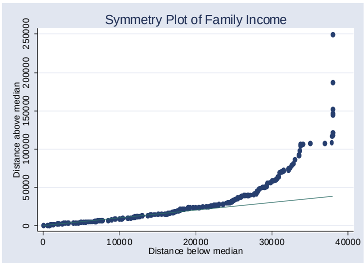


graph box studios, ti("Box Office Revenue Over 3 Years: Nine Film Studios") ytitle("Annual Revenue in Millions of Dollars") over(_stack)

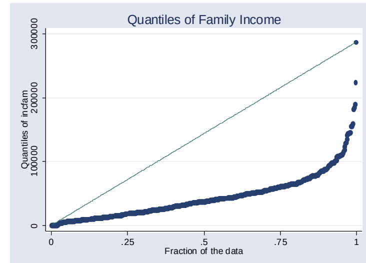


use
"D:\amifiles\econ70\lectures_2005\a_Introduction\pers75w_subs.dta",
clear

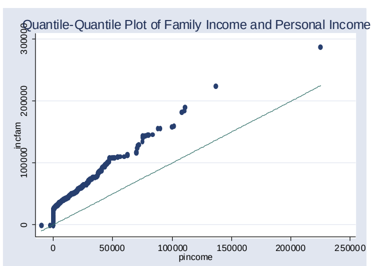
symplot incfam, ti("Symmetry Plot of
Family Income")



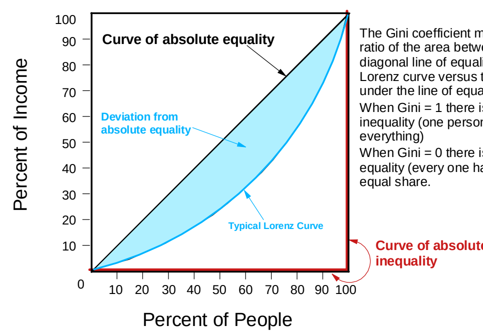
quantile incfam, ti("Quantiles of Family Income")



qqplot incfam pincome, ti("Quantile-Quantile
Plot of Family Income and Personal Income")

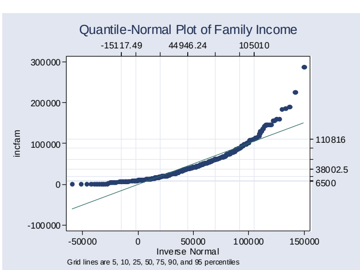


The Lorenz Curve As A
Measure of Inequality



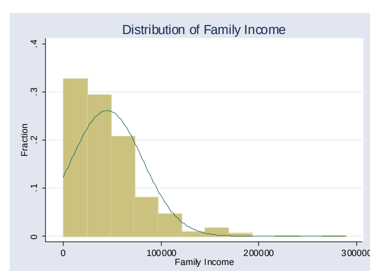
The Gini coefficient measures the ratio of the area between the diagonal line of equality and the Lorenz curve versus the total area under the line of equality.
When Gini = 1 there is absolute inequality (one person has everything)
When Gini = 0 there is absolute equality (every one has an exactly equal share.

qnorm incfam, grid ti("Quantile-Normal Plot of Family Income")
ylabel(, angle(horizontal) axis(1)) ylabel(, angle(horizontal)
axis(2))

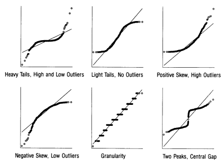


label variable incfam "Family Income"

. histogram incfam, bin(12) norm fraction
ti("Distribution of Family Income")



Interpreting the Quantile-Normal Plot



The quantile-normal figures illustrates how quantile-normal plots reflect distribution shape:

1. Heavy-tailed distributions are steepest at the top and bottom.

2. Light-tailed distributions are least steep at the top and bottom.

3. Skewed distributions have one heavy and one light tail. This gives plots of positively skewed distributions a downward-bowed appearance, like that of water use in Figure 1.9. Negatively skewed distributions bow upward.

4. Outliers appear as points toward the upper right or lower left, vertically separated from the rest of the distribution.

5. Granularity means that certain discrete values occur repeatedly in the data. These appear as plateaus separated by gaps.

6. Since less-steep areas indicate higher-than-normal data density, two less-steep areas separated by a gap or steep climb (lower-than-normal density) indicate that the distribution has two peaks (is bimodal).

Professor B. Turchi