Biostatistics 208
Data Exploration

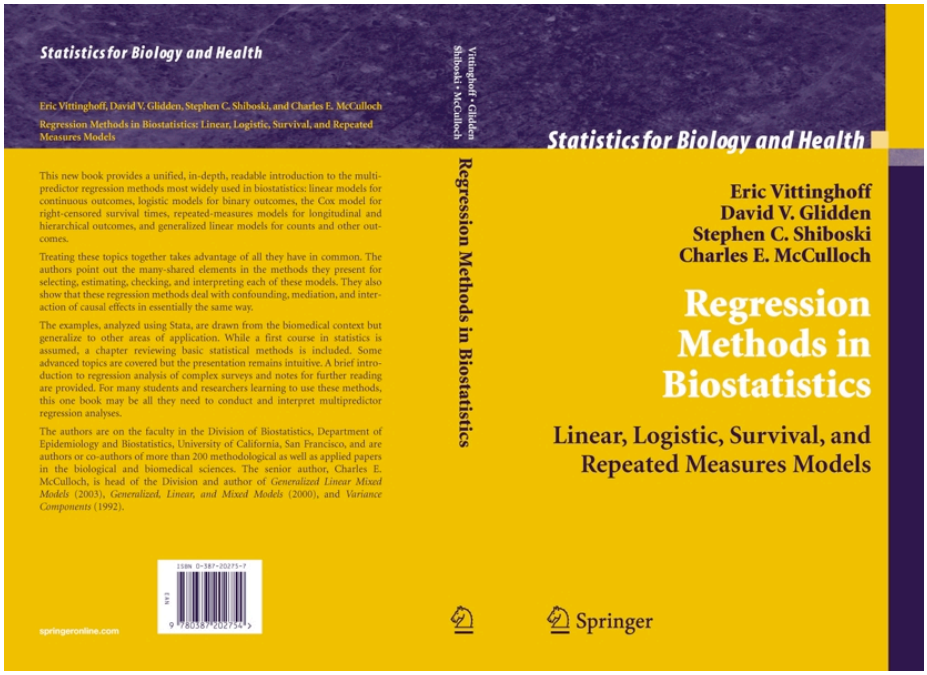
Dave Glidden
Professor of Biostatistics
Univ. of California, San Francisco

January 8, 2008
http://www.biostat.ucsf.edu/biostat208

Organization

- Office hours by appointment (CBL 5724)
- E-mail me to make an appointment or get questions answered quickly: dave@biostat.ucsf.edu
- Download lecture slides, labs, data from http://www.biostat.ucsf.edu/biostat208

Textbook



Lectures

- Descriptive Statistics (Glidden)
  3 lectures
- Linear Regression (Vittinghoff)
  4 lectures
- Logistic Regression (Shiboski)
  4 lectures

Computer labs

- Thursdays in CBL 6704
- Two sections: 10:30-11:30, 11:30-12:30
  by assignment
- Need a laptop or terminal server account
- Labs show how to use STATA to implement methods discussed in class
- We supply the data and commands, plus an interpretive handout at the end of the lab

Homework

- Counts for 70% of the grade
- Five total homework assignments
  missing one can making passing difficult
- Your responsibility to meet deadlines
  if out of town, make arrangements
- Handed back quickly. Late not accepted.

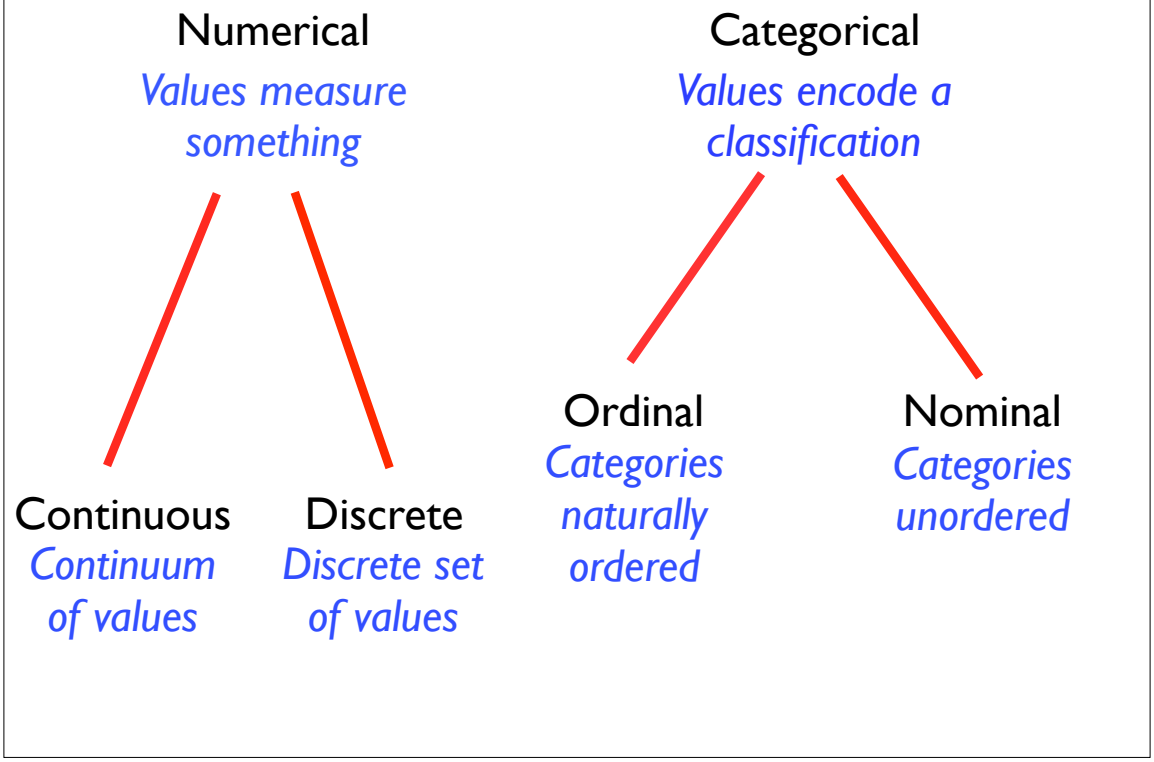
Class Philosophy

- Emphasis: Practical not theoretical
- Outlook: pragmatic not dogmatic
- Applied statistics is an art
- Often no single correct method
- Computing is a valuable tool
  not going to 'teach' Stata per se

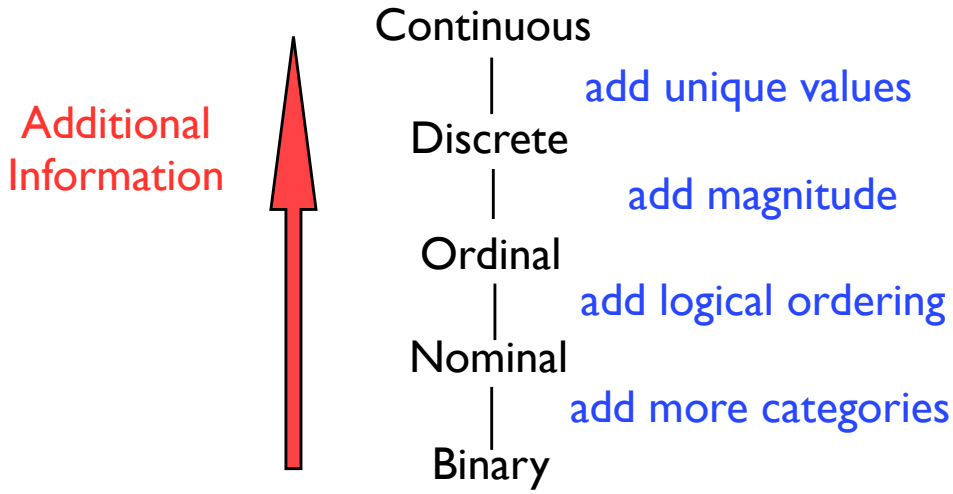
Multiple Predictor Regression

- Assess the relationship between an outcome and multiple predictors
- Powerful tool for:
  - understanding complex relationships
  - controlling confounding
  - prediction / risk stratification
- Regression models differ by outcome type, but all have much in common

Data types



Hierarchy of data types



Name the data type

- Blood pressure
- Likert scale
- Presence or absence of disease
- Number of hospitalizations
- Age (*in years*)
- Genotype (*wild type, heterozygote, homozygous mutant*)

More on Data Type

- Data type implies plausible probability distn
- Different data summaries
  the sample mean *not always interpretable*
- Distinctions between different types can be flexible

Model depends on outcome type

- Continuous -- linear or gamma model
- Discrete (counts)- Poisson/negative binomial
- Binary -- logistic, relative risk models, survival models when follow-up varies
- *All easily implemented in Stata*

Data Exploration

- Find data errors
- Assess missingness
- Detect anomalous observations and outlying data values
- Select appropriate analysis methods
- Support a formal data analysis

Data example

- Western Collaborative Group Study
- Large early observational study (n=3154)
- Association between "type A" behavior and coronary heart disease (CHD)
- Example variable: systolic blood pressure

Descriptive Output

```
summarize sbp, detail

                        systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           104             98
 5%           110            100
10%           112            100       Obs                 3154
25%           120            100       Sum of Wgt.         3154

50%           126                      Mean           128.6328
                        Largest       Std. Dev.      15.11773
75%           136            210
90%           148            210       Variance       228.5458
95%           156            212       Skewness       1.204397
99%           176            230       Kurtosis       5.792465
```

Descriptive Output

```
summarize sbp, detail

                        systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           104            98
 5%           110           100
10%           112           100      Obs                3154
25%           120           100      Sum of Wgt.        3154

50%   median  126                    Mean           128.6328
                       Largest       Std. Dev.      15.11773
75%           136           210
90%           148           210      Variance       228.5458
95%           156           212      Skewness       1.204397
99%           176           230      Kurtosis       5.792465
```

four smallest values

four biggest values

Summary statistics

- Mean and standard deviation capture 'location' and 'spread', *but are sensitive to skewness, outliers*
- More robust five-number summary:
  1. minimum
  2. 25th percentile (lower quartile)
  3. 50th percentile (median)
  4. 75th percentile (upper quartile)
  5. maximum

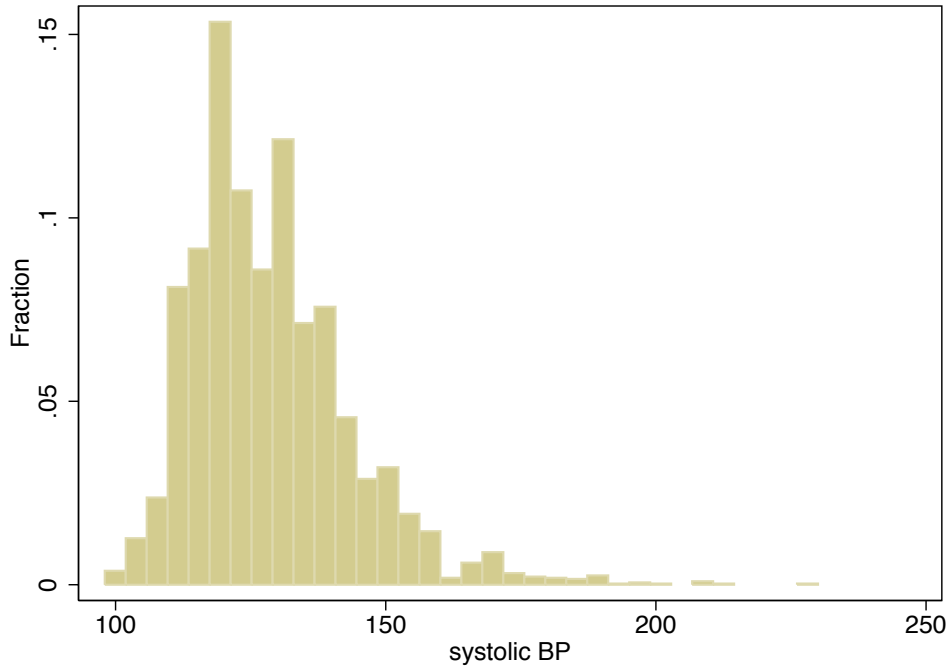
Cholesterol data w/ outlier

	mean	SD	min	25%	50%	75%	max
with outlier	241	88	170	198	229	237	645
omit outlier	224	27	170	198	228	235	294

Graphical data summaries

- Can effectively communicate the distribution
- Methods have different strengths:
  - Histogram: *captures location, spread, shape of the distribution; "density" plot as a smoothed histogram*
  - Boxplot: *gives 5-number summary, shows outliers, skewness*
  - Normal Q-Q plot: *assesses Normality*

Histogram of SBP



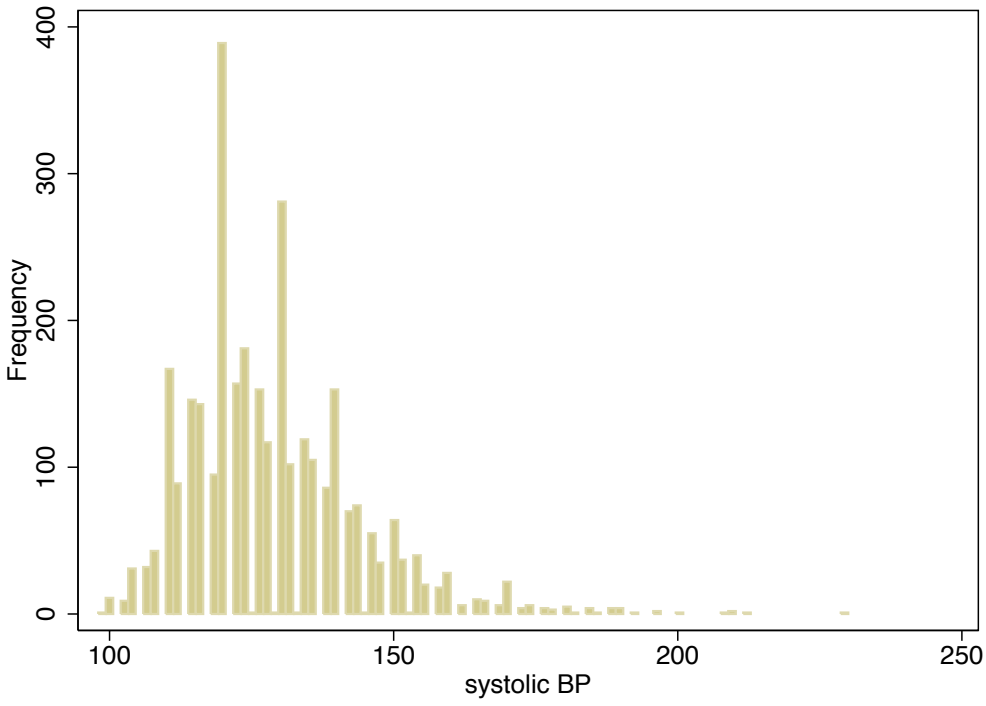
Histogram

- Shows location, spread, and shape of the distribution
- Horizontal axis: *intervals or "bins" in which data values are grouped*
- Vertical axis: *number, fraction, or percent of the observations in each bin*

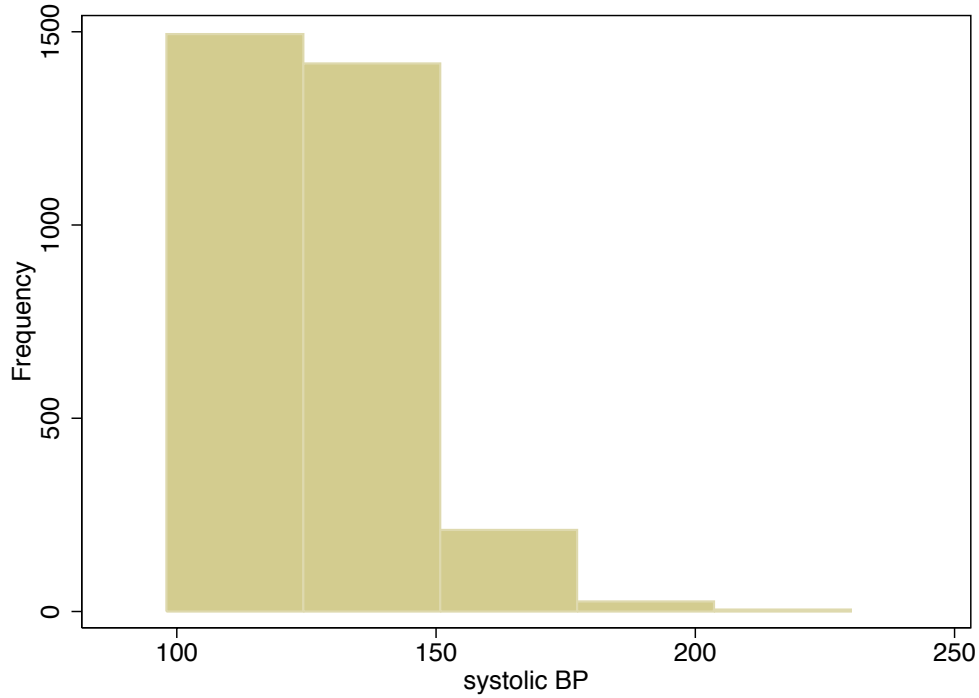
Interpreting Histograms

- Pattern of bar heights conveys shape of distribution:
  - *number of modes*
  - *skewness*
  - *long or short tails*
- Usefulness depends on number of bins
  - *too many defeats goal of summarization*
  - *too few obscures shape of distribution*

Too many bins



Too few bins

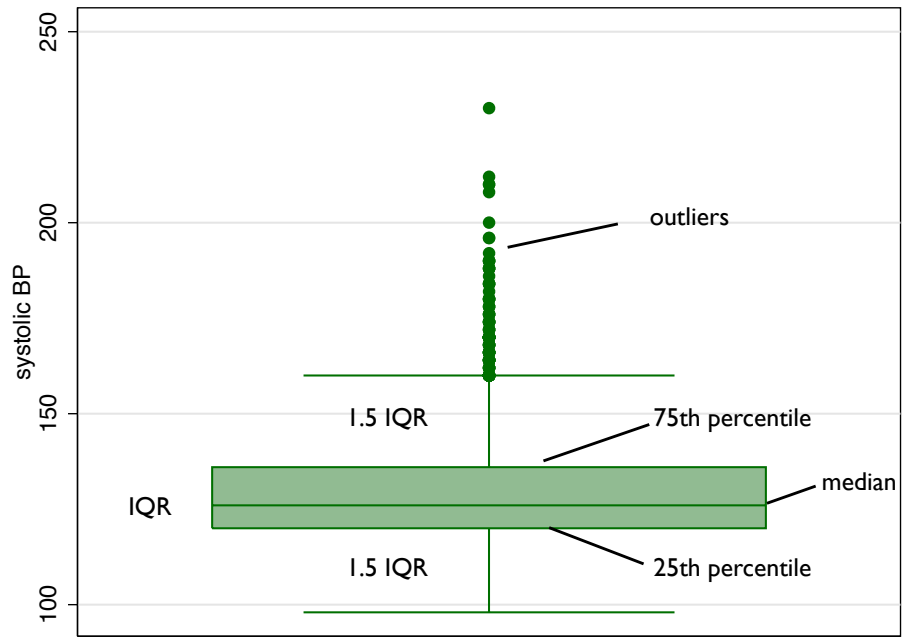


Stata Commands

- histogram *varname*
  to graph a histogram
- histogram *varname*, bin(x)
  histogram with x bars
- histogram varname, freq
  histogram with frequency not fractions

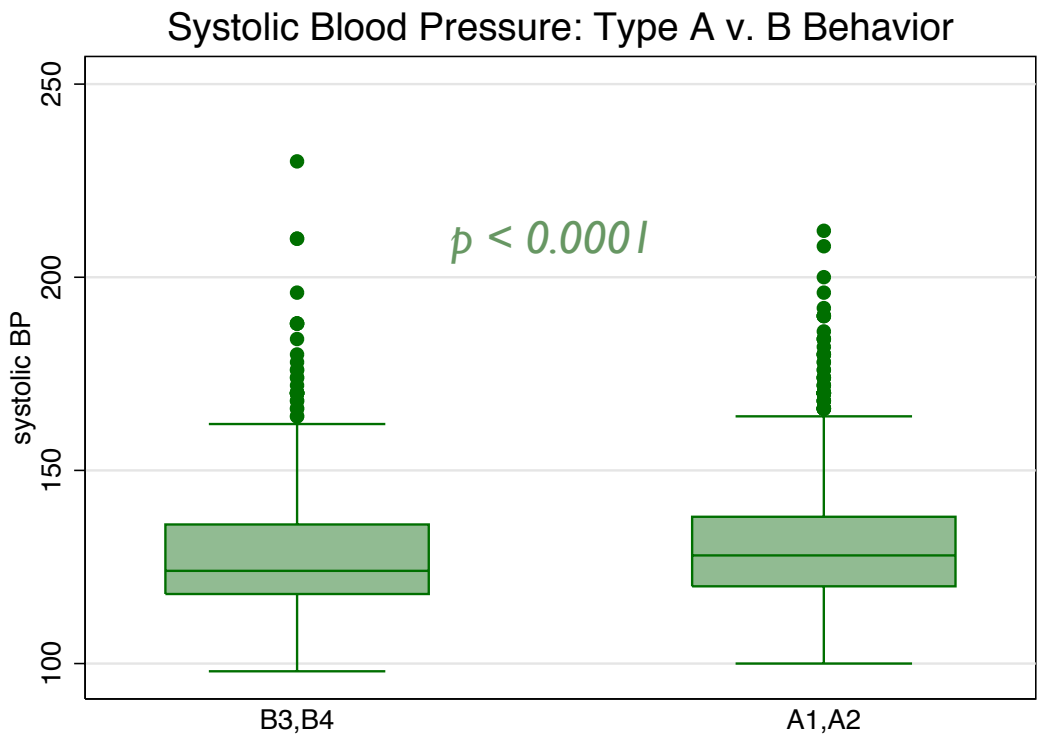
Boxplot

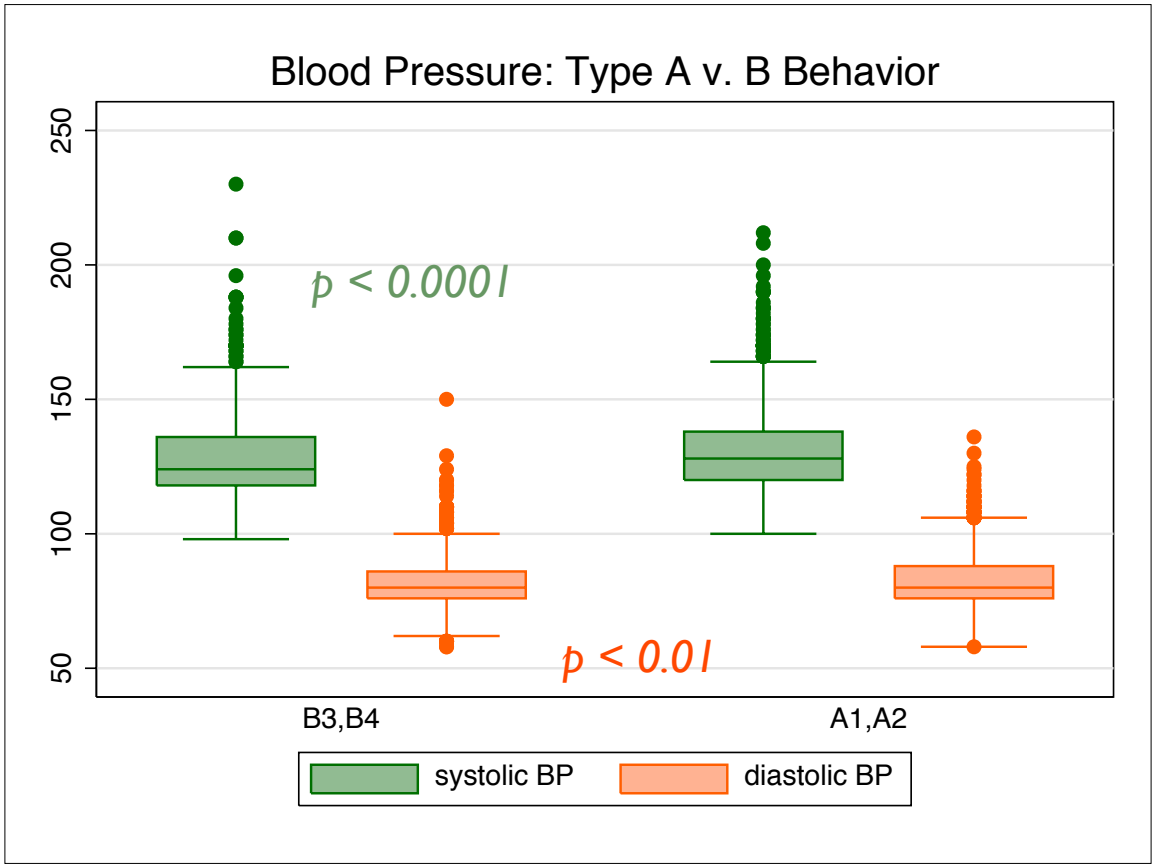
- Box with upper & lower hinges
- Box: 25% tile, median, 75% tile
- Length of box: interquartile range (IQR)
- Lower hinge: 25% tile minus 1.5*IQR
- Upper hinge: 75% tile plus 1.5*IQR
- Values outside hinges: outliers



Using a Boxplot

- Location: given by lines in box, *median*
- Spread: given by size of box, *IQR*
- Skewness: distance between the lines
- Outliers are clearly marked
  can usually tell how many and their values



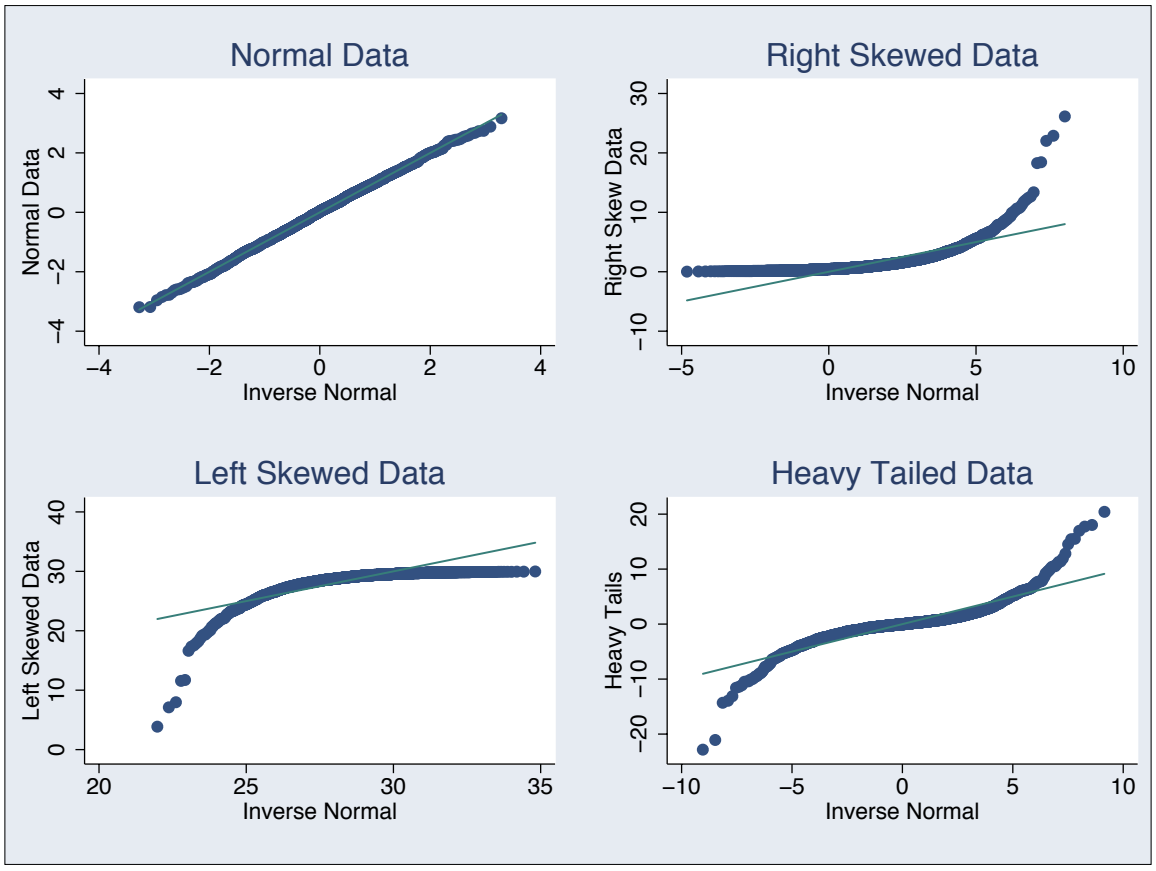


Stata Command

- `graph box` *varname*
  to graph a boxplot
- graph box *varname*, over(*grpvar*)
  side-by-side boxplots based on grpvar
- group *varname1 varname2*, over(*grpvar*)
  side-by-side boxplots for two variables

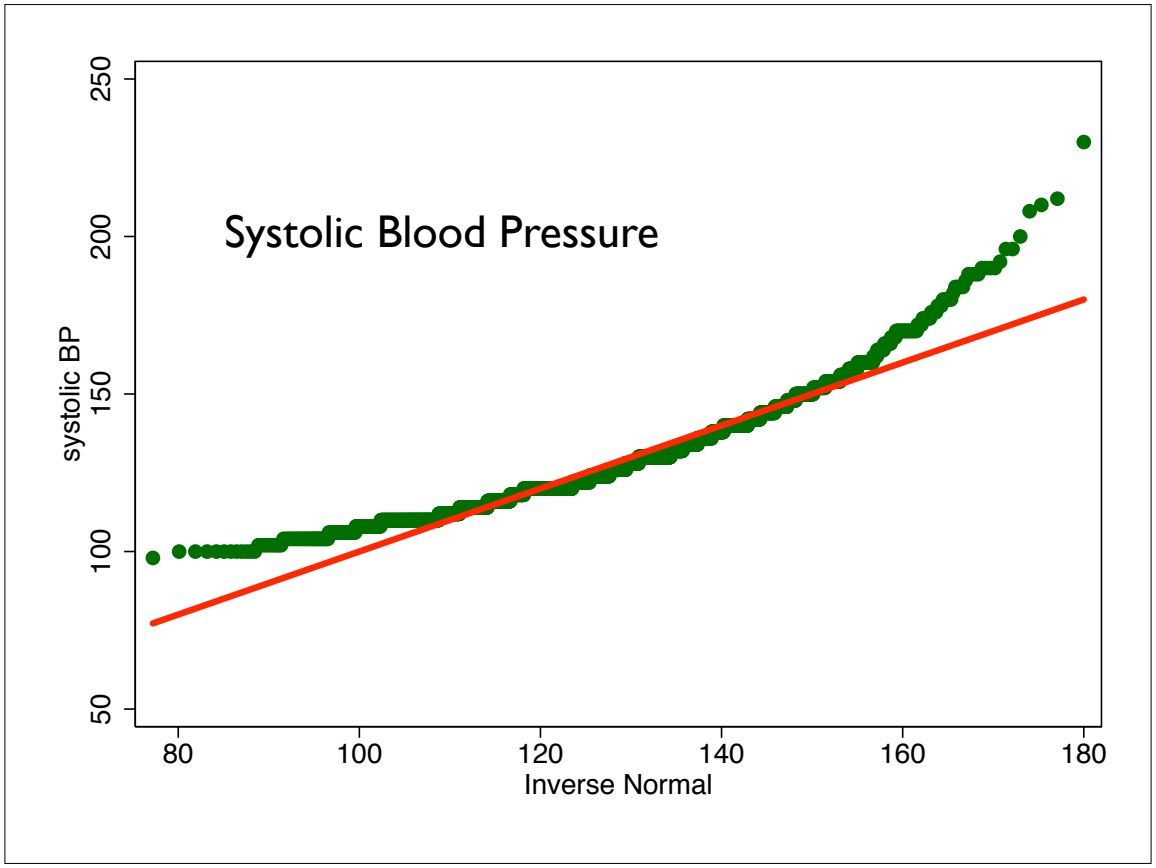
qq Normal Plot

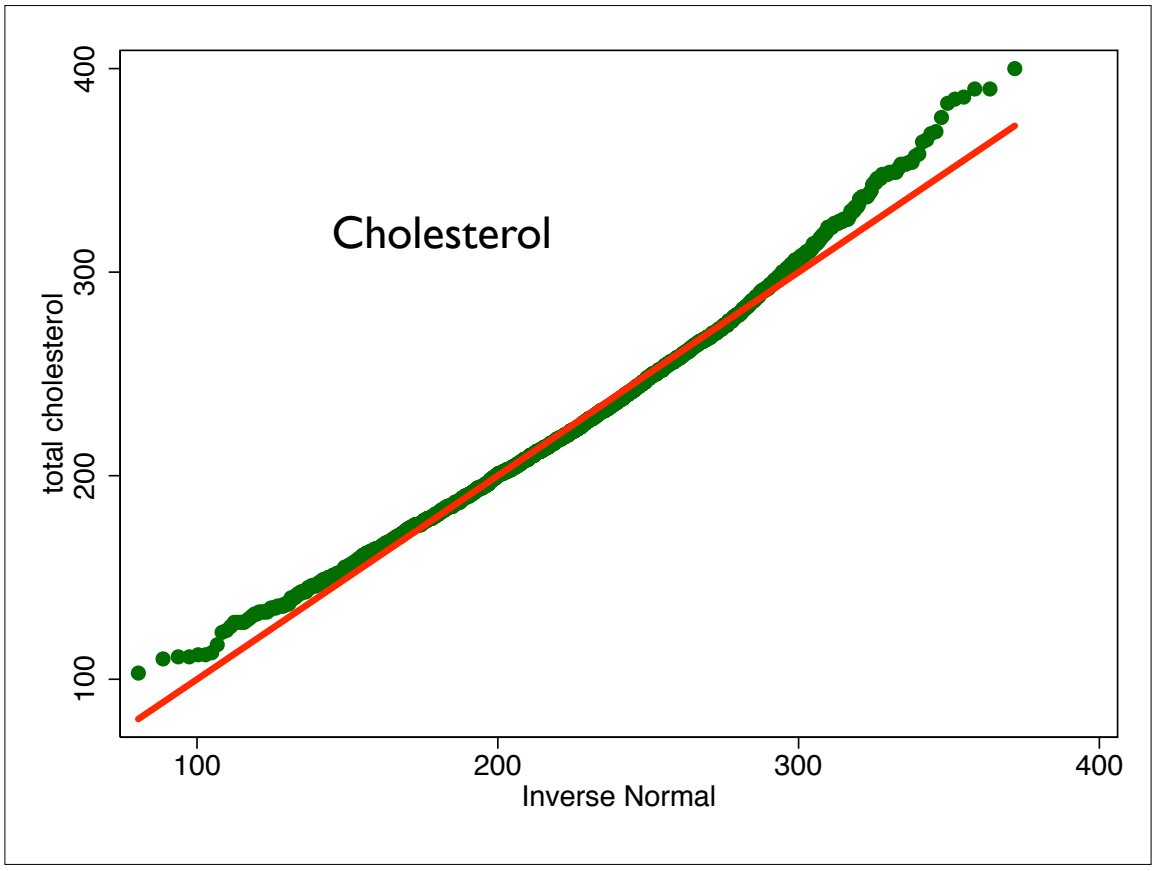
- Graphical approach to assessing Normality
- Horizontal axis (x axis)
  sorted data values
- Vertical axis (y axis)
  expected data values if data Normal
- If plot straight, data is nearly Normal
- Shape indicates nature violation, if any



Using qq Normal Plot

- Right skew: plot curved up
- Left skew: plot curved down
- Outlier: values far off line
- STATA: `qnorm varname`


Systolic Blood Pressure
systolic BP
250
200
150
100
50
80
100
120
140
160
180
Inverse Normal


Cholesterol
total cholesterol
400
300
200
100
100
200
300
400
Inverse Normal

Transforming variables

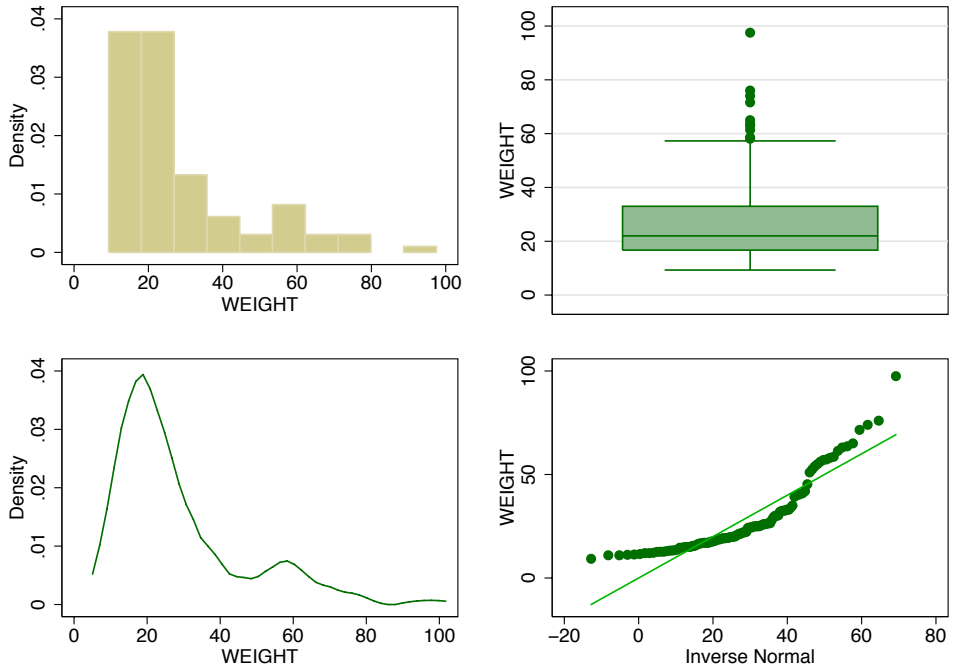
Rationale:

- Make outcome more normally distributed
- Linearize predictor effects, remove interactions, equalize outcome variance
- Much more about this later

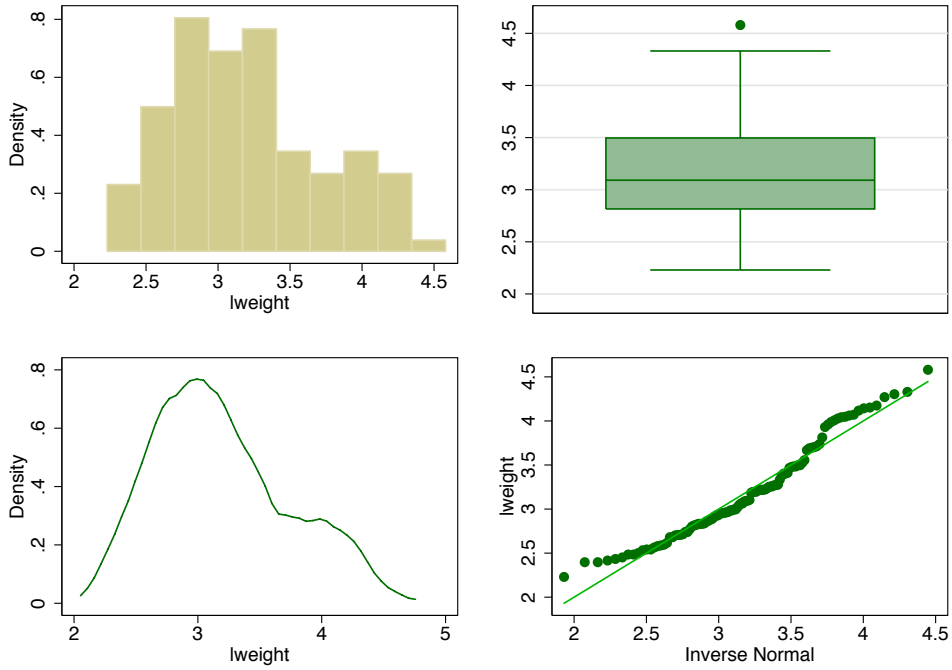
Drawbacks:

- Untransformed variable more credible, interpretable
- Natural scale may be more meaningful: *cost vs log cost*

Weight before transformation



Log-transformed weight



Log transformation 'pulls in the tail'

Value	*Difference*	*$\log_{10}$ value*	*$\log_{10}$ Difference*
0.01	-	-2	-
0.1	0.09	-1	1
1	0.9	0	1
10	0	1	1
100	90	2	1
1000	900	3	1

Can be used to linearize a relationship *(more on this later)*

Frequency Tables

- Used for categorical data
  loses no information
- Display raw numbers of percentages
- Can be used for continuous data
  discards lots of information
  may create relevant groups

SBP and behavioral pattern

```
. tab sbpcat

 systolic BP |      Freq.     Percent        Cum.
-------------+-----------------------------------
  < 120 mmHg |        767       24.32       24.32
120-139 mmHg |      1,694       53.71       78.03
140-159 mmHg |        567       17.98       96.01
 >= 160 mmHg |        126        3.99      100.00
-------------+-----------------------------------
       Total |      3,154      100.00

. tab behpat

 behavioral |
 pattern (4 |
     level) |      Freq.     Percent        Cum.
------------+-----------------------------------
         A1 |        264        8.37        8.37
         A2 |      1,325       42.01       50.38
         B3 |      1,216       38.55       88.93
         B4 |        349       11.07      100.00
------------+-----------------------------------
      Total |      3,154      100.00
```

Summary

- Types of Data: Numerical v. Categorical
  Categorical: ordered or not
  Numerical: discrete v. continuous
- Numerical: mean, SD, 5 numbers
- Numerical: histogram, boxplot, qq normal
- Categorical: Tables
- Transformations: potentially useful

Available on Website

- Syllabus, due dates, note about 209 project
- Lecture slides
- WCGS dataset *(wcgs.dta)*
- Stata commands to make graphs *(lecture1.do)*
- Instructions and data for Thursday's lab
- http://www.biostat.ucsf.edu/biostat208