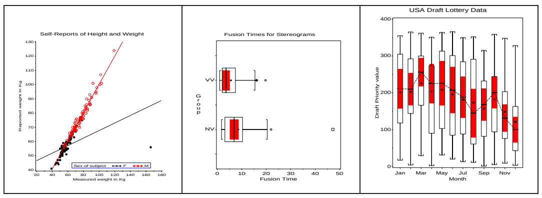
Data Screening



Michael Friendly

York University

SCS Short Course

October, 2004

Course Outline

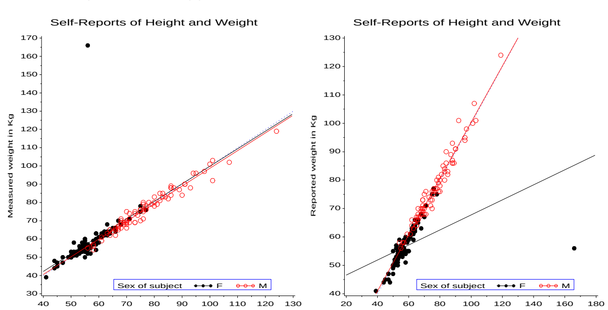
- Part 1: Getting started
  - Failures to screen data
  - Entering and checking raw data
    - Data entry
    - Creating a documented database
    - Checking data at input
  - Assessing univariate problems
    - Boxplots and outliers
    - Transformations to symmetry
    - Normal probability plots
- Part 2: Assessing bivariate problems
  - Enhanced scatterplots
    - Smoothing relations
    - Plotting discrete data
  - Transformations to linearity
  - Dealing with non-constant variance
- Part 3: Multivariate problems and missing data
  - Assessing multivariate problems
    - Multivariate normality
    - Multivariate outliers
  - Dealing with missing data
    - Estimation with missing data (EM algorithms)
    - Simple Imputation
    - Multiple Imputation
- SAS macro programs:
  - `http://www.math.yorku.ca/SCS/sssg/`
  - `http://www.math.yorku.ca/SCS/sasmac/`

Color versions of these slides:
`http://www.math.yorku.ca/SCS/Courses/screen/`

Failures to Screen Data

Data on Self-Reports of height and weight among men and women active in exercise

- Regression of reported weight on measured weight gave very different regressions for men and women
- Plotting the data suggested an answer



Failures to Screen Data

Fusion times for random dot stereograms

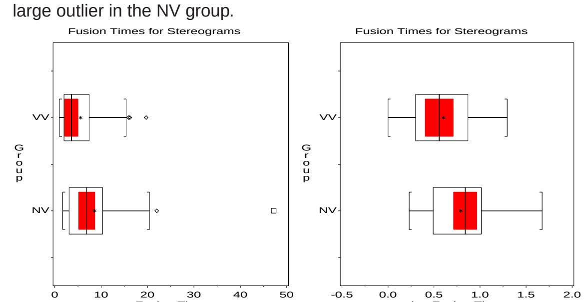
- Does knowledge of the form of an embedded image affected time required for subjects to fuse the images?
- Two group design: Group NV (no visual info), Group VV (visual and verbal info).
- t-test: $t(76) = 1.939, p = 0.0562$, NS!

```
                  TTEST PROCEDURE

Variable: TIME          Fusion Time

Variances          T        DF      Prob>|T|
----------------------------------------------
Unequal       2.0384      70.0        0.0453
Equal         1.9395      76.0        0.0562
```

- Boxplots show: times are positively skewed, differ in variance, and one large outlier in the NV group.



- Transforming the raw data to log (time) cured these problems, and led to the opposite conclusion!

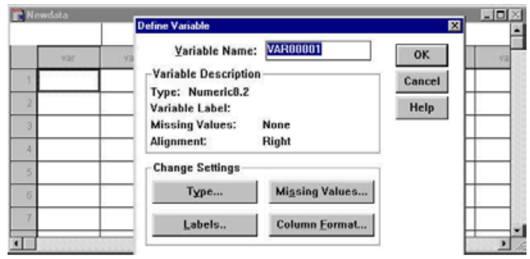
See `lib.stat.cmu.edu/DASL/Stories/FusionTime.html`

Entering Raw Data - Basic tools

- Ordinary editor (e.g., Notepad, WinEdt, UltraEdit)
  - Easy for small data sets
  - Manual alignment of input fields
  - No protection against input errors (wrong type, out of range, etc.)
- Spreadsheet (e.g., Excel)
  - Easy for small to moderate sized data sets
  - Automatic alignment of input fields
  - Automatic calculation of derived variables,
  - Keyboard macros for repetitive tasks
  - Programmable macros for checking input
  - Import .xls spreadsheet to SAS (File – Import)
  - Data conversion tools (e.g., `dbmscopy` → SAS, SPSS)
- Database packages (Access, dBase, etc.)
  - Easy for small to large sized data sets
  - Define fields (type, length, min, max)
  - Values can be verified as entered
  - Import .dbf database to SAS (File – Import)
  - Data conversion tools (e.g., `dbmscopy` → SAS, SPSS)

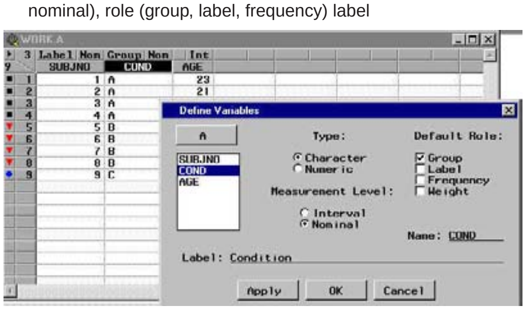
Entering Raw Data - Statistical packages

- SPSS
  - Startup: Newdata window
  - Define variable - name, type (num/char), variable label, missing values
  - Import .por (portable file) to SAS
  - Data conversion tools (e.g., `dbmscopy` → SAS)



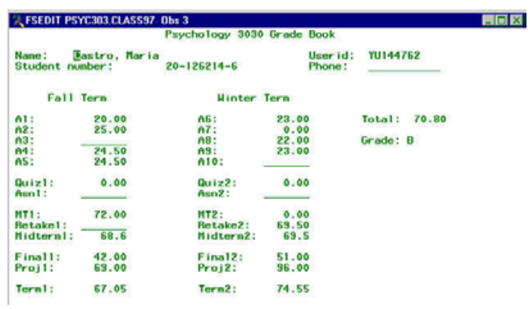
Entering Raw Data - Statistical packages

- SAS/Insight
  - Globals – Analyze – Interactive Data Analysis – New
  - Define variables - name, type (num/char), measurement level (interval, nominal), role (group, label, frequency) label



- SAS/FSP - Full Screen Product
  - Design display screen to suit the application
  - Define variable - name, type (num/char), variable label

```
filename psy303 '~/sasuser/psyc303';
proc fsedit data=psy303.class97 screen=psy303.screen97;
```



  - Assign name, type, min, max, required, etc. to each variable
  - Automatic range checking
  - Automatically computed fields

```
asn1 = mean( of A1-A5);
asn2 = mean( of A6-A10);
```

Creating a documented database

- Example: Baseball data

```
Andy Allanson ACLEC 293 66  1 30 29 14  1  293  66   1 ...
Alan Ashby    NHOUC 315 81  7 24 38 39 14 3449 835  69 ...
Alvin Davis   ASEA1B479130 18 66 72 76  3 1624 457  63 ...
Andre Dawson  NMONRF496141 20 65 78 37 11 56281575 225 ...
A Galarraga   NMON1B321 87 10 39 42 30  2  396 101  12 ...
A Griffin     AOAKSS594169  4 74 51 35 11 44081133  19 ...
```

- SAS
  - Assign descriptive labels to variables
  - User-defined formats (PROC FORMAT) for variable values

```
/* Formats to specify coding of variables (other=error) */
proc format;
  value $league
    'N' ='National'     'A' ='American'    other= ' ';
  value $team
    'ATL'='Atlanta      '     'BAL'='Baltimore     '
    'BOS'='Boston       '     'CAL'='California    '
    'CHA'='Chicago A    '     'CHN'='Chicago N     '
    ...                       other = ' ';
data baseball(label='1986 Baseball Hitter Data');
  input name $1-14 league $15 team $16-18 position $19-20
    atbat 3. hits 3. homer 3. runs 3. rbi 3. walks  3.
    years 3. atbatc 5. hitsc 4. homerc 4. runsc 4. rbic 4.
    walksc 4. putouts 4. assists 3. errors 3. salary 4.;
  label
    name = "Hitter's name"  atbat = 'Times at Bat'
    hits = 'Hits'           homer = 'Home Runs'
    runs = 'Runs'           rbi   = 'Runs Batted In'
    walks= 'Walks'         years = 'Years in Major Leagues'
    ...
  format league $league.  team $team.;
```


Checking data at input

- Check categorical variables using 'other' format

```
data baseball(label='1986 Baseball Hitter Data');
  input name $1-14 league $15 team $16-18 position  ...
  ...
  if (put(league, $league.) = ' ' then error;
  if (put(team, $team.) = ' ' then error;
  if (put(position, $position.) = ' ' then error;
  ...
```

- Check ranges of numeric variables

```
  if !(0 < atbat < 500) then error;
  if !(0 < hits  < 500) then error;
  if !(0 < years < 50)  then error;
  ...
```


Checking variables

- Descriptive statistics checks
  - SPSS - Frequencies
  - SAS - PROC UNIVARIATE
    - Min, Max, # missing
    - Mean, median, std. dev, skewness, etc.
  - Use plot option for stem-leaf/boxplot and normal probability plot
  - Use ID statement to identify highest/lowest obs.

```
proc univariate plot data=baseball;
   var atbat -- salary ;
   id name;
```

- Consistency checks (e.g., unmarried teen-aged widows?)
  - SPSS - Crosstabs
  - SAS - PROC FREQ

```
proc freq;
   tables age * marital;
```

- But: these can generate too much output!

Checking numeric variables - the DATACHK macro

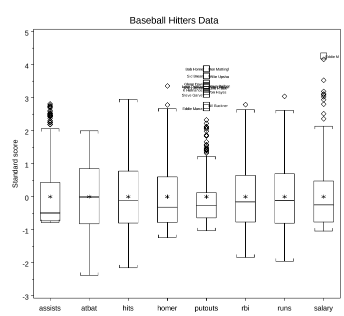
- Uses PROC UNIVARIATE to extract descriptive stats, high/low obs.
- Formats output to 5 variables/page
- Boxplot of standardized scores to show distribution shape, outliers
- Lists observations with more than nout (default: 3) extreme z scores, $|z| >$ zout (default: 2)
- Example:

```
  %include data(baseball);
  %datachk(data=baseball, id=name,
var=salary runs hits rbi atbat homer assists putouts);
```

Documentation:

http://www.math.yorku.ca/SCS/sasmac/datachk.html

Hebb lab (SAS): %webhelp(datachk);

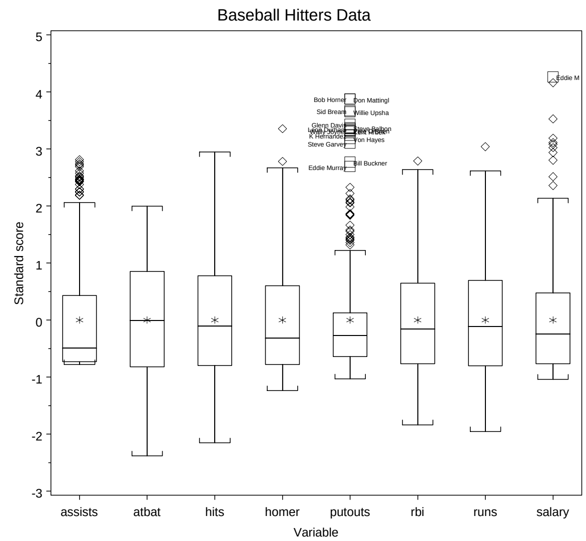


```
Variable              Stat    Value       Extremes Id

ATBAT                 N         322          16 Tony Armas
                      Miss        0          19 Cliff Johnson
Times at Bat          Mean   380.9286        19 Terry Kennedy
                      Std     153.405        20 Mike Schmidt
                      Skew   -0.07806
                                            663 Joe Carter
                                            677 Don Mattingly
                                            680 Kirby Puckett
                                            687 T Fernandez
                      ------------------------------------------
HITS                  N         322           1 Mike Schmidt
                      Miss        0           2 Tony Armas
Hits                  Mean   101.0248         3 Doug Baker
                      Std    46.45474         4 Terry Kennedy
                      Skew   0.291154
                                            211 Tony Gwynn
                                            213 T Fernandez
                                            223 Kirby Puckett
                                            238 Don Mattingly
                      ------------------------------------------
RBI                   N         322           0 Doug Baker
                      Miss        0           0 Mike Schmidt
Runs Batted In        Mean   48.02795         0 Tony Armas
                      Std    26.16689         2 Bob Boone
                      Skew   0.608377
                                            113 Don Mattingly
                                            116 Dave Parker
                                            117 Jose Canseco
                                            121 Joe Carter
                      ------------------------------------------
RUNS                  N         322           0 Mike Schmidt
                      Miss        0           1 Cliff Johnson
Runs                  Mean   50.90994         1 Doug Baker
                      Std     26.0241         1 Tony Armas
                      Skew   0.415779
                                            108 Joe Carter
                                            117 Don Mattingly
                                            119 Kirby Puckett
                                            130 R Henderson
                      ------------------------------------------
SALARY                N         263          68 B Robidoux
                      Miss       59          68 Mike Kingery
Salary (in 1000$)     Mean   535.9658        70 Al Newman
                      Std     451.104        70 Curt Ford
                      Skew   1.589077 *
                                           1975 Don Mattingly
                                           2127 Mike Schmidt
                                           2413 Jim Rice
                                           2460 Eddie Murray
                      ------------------------------------------
```


The `datachk` macro

Boxplots of standard scores show the 'shape' of each variable, with labels for 'far-out' observations.



Sidebar: Using SAS macros

- SAS macros are high-level, general programs consisting of a series of `DATA` steps and `PROC` steps.
- Keyword arguments substitute your data names, variable names, and options for the named macro parameters.
- Use as:
  `%macname(data=dataset, var=variables, ...);`
  e.g.,
  `%boxplot(data=nations, var=imr, class=region, id=nation);`
- Most arguments have default values (e.g., `data=_last_`)
- All *SSSG* and *VCD* macros have internal and/or online documentation,
  http://www.math/yorku.ca/SCS/sssg/
  http://www.math/yorku.ca/SCS/sasmac/
  http://www.math/yorku.ca/SCS/vcd/
- Macros can be installed in directories *automatically* searched by SAS. Put the following `options` statement in your `AUTOEXEC.SAS` file:

```
options sasautos=('c:\sasuser\macros' sasautos);
```


Sidebar: Using SAS macros

E.g., the `SYMBOX` macro is defined with the following arguments:

symbox.sas ···

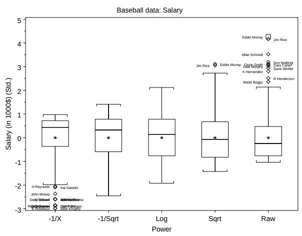
```
%macro symbox(
  data=_last_,          /* name of input data set           */
  var=,                 /* name(s) of the variable(s) to examine*/
  id=,                  /* name of ID variable               */
  out=symout,           /* name of output data set           */
  orient=V,             /* orientation of boxplots: H or V   */
  powers=-1 -0.5 0 .5 1,  /* list of powers to consider      */
  name=symbox           /* name for graph in graphics catalog */
  );
```

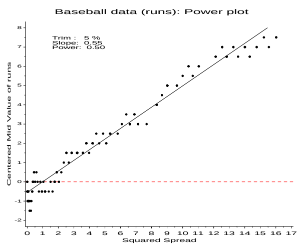
Typical use:

```
%symbox(data=baseball,
   var=Salary Runs,   /* analysis variables  */
   id=name,           /* player ID variable  */
   powers =-1 -.5 0 .5 1 2);
```

Assessing univariate problems

- Boxplots
- Transformations to symmetry
- Outliers
- Normal probability plots

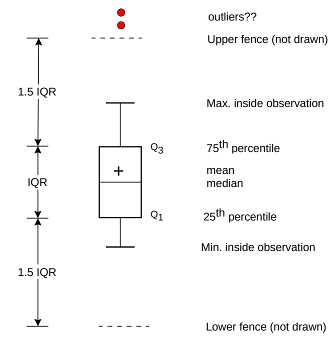




Boxplots

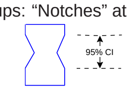
Boxplots provide a *schematic* graphical summary of important features of a distribution, including:

- the center (mean, median)
- the spread of the middle of the data (IQR)
- the behavior of the tails
- outliers (plotted individually)



- Notched boxplots for multiple groups: "Notches" at

$$\text{Median} \pm 1.58 \frac{\text{IQR}}{\sqrt{n}}$$

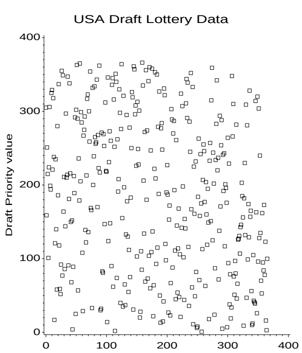


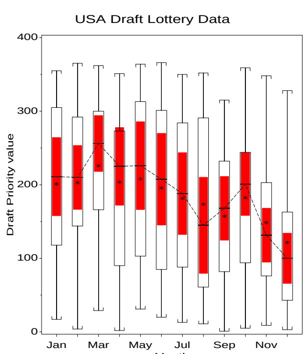
show approximate 95% confidence intervals around the medians. Medians differ if the notches do not overlap (McGill et al., 1978).

Boxplots - Example

1970 USA Draft Lottery

- Each birth date assigned a "random" priority value for selection to the military
- Ordinary scatterplot does not reveal anything unusual
- Boxplots by month show those born later in the year more likely to be drafted





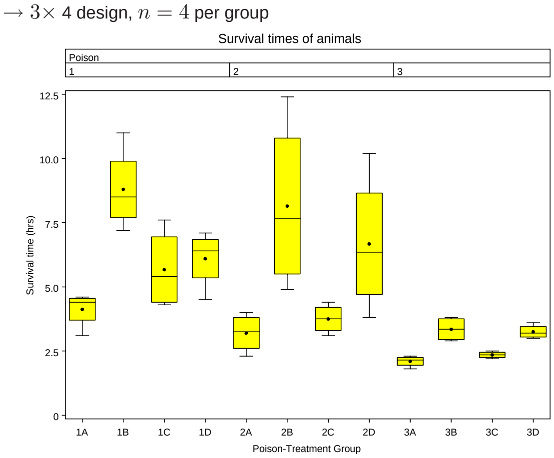
See Friendly (1991), "SAS System for Statistical Graphics" §6.3.

Boxplots - ANOVA data

- Boxplots are particularly useful for comparing groups
- ANOVA: Do means differ?
- ANOVA: Assumes equal within-group variance!

Example: Survival times of animals (Box and Cox, 1964)

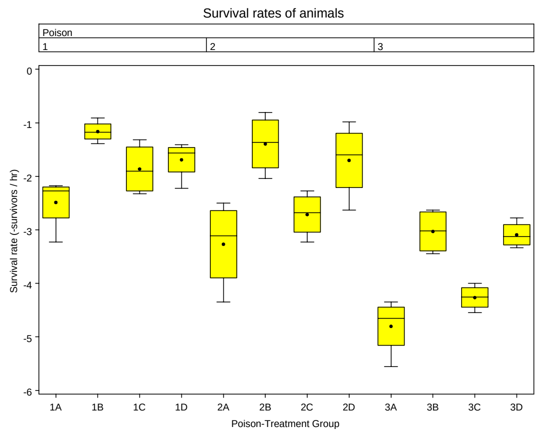
- Animals exposed to one of 3 types of poison
- Given one of 4 treatments
- $\rightarrow 3\times$ 4 design, $n = 4$ per group



- Boxplot shows that variance increases with mean (why?)

Boxplots - ANOVA data

- Methods we will learn today suggest that power transformations, $y \longrightarrow y^p$ are often useful.
- Methods we will learn next week suggest rate = 1 / time



Boxplots with SAS

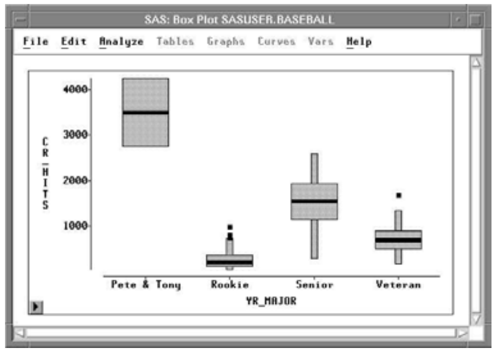
- `BOXPLOT` macro - Graphics plots with many options

```
%boxplot(data=draftusa, class=Month,
   var=priority, id=day,  connect=3, cnotch=red);
```

- `PROC BOXPLOT` (Version 8)

```
proc boxplot;
   plot priority * month;
```

- SAS/INSIGHT - Analyze – Box Plot, select response as Y, class variable(s) as X. Selecting highlights obs. in all other views.



Transformations to symmetry

- Transformations have several uses in data analysis, including:
  - making a distribution more symmetric.
  - equalizing variability (spreads) across groups.
  - making the relationship between two variables linear.
- These goals often coincide: a transformation that achieves one goal will *often* help for another (but not *always*).
- Some tools (Friendly, 1991):
  - Understanding the *ladder of powers*.
  - `SYMBOX` macro - boxplots of data transformed to various powers.
  - `SYMPLOT` macro - various plots designed to assess symmetry. POWER plot: line with slope $b \Rightarrow y \longrightarrow y^p$, where $p = 1 - b$ (rounded to 0.5).
  - `BOXCOX` macro - for regression model, transform $y \longrightarrow y^p$ to minimize MSE (or maximum likelihood); influence plot shows impact of observations on choice of power (Box and Cox, 1964).
  - `BOXGLM` macro - for GLM (anova/regression), transform $y \longrightarrow y^p$ to minimize MSE (or max. likelihood)
  - `BOXTID` macro - for regression, transform $x_i \longrightarrow x_i^p$ (Box and Tidwell, 1962).

Transformations – Ladder of Powers

- Power transformations are of the form $x \longrightarrow x^p$.
- A useful family of transformations is *ladder of powers* (Tukey, 1977), defined as $x \longrightarrow t_p(x)$,

$$t_p(x) = \begin{cases} \frac{x^p - 1}{p} & p \neq 0 \\ \log_{10} x & p = 0 \end{cases} \tag{1}$$

- Key ideas:
  - $\log(x)$ plays the role of x^0 in the family.
  - $1/p \rightarrow$ keeps order of x the same for $p < 0$.
- For simplicity, usually use only simple integer and half-integer powers (sometimes, $p = 1/3 \longrightarrow \sqrt[3]{x}$); scale the values to keep results simple.

Power	Transformation	Re-expression
3	Cube	x^3 /100
2	Square	x^2 /10
1	NONE (Raw)	x
1/2	Square root	$\sqrt{x}$
0	Log	$\log_{10} x$
-1/2	Reciprocal root	$-10/\sqrt{x}$
-1	Reciprocal	$-100/x$

Ladder of Powers – Properties

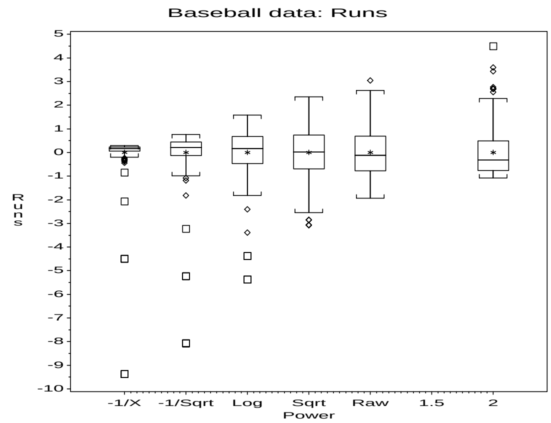
- ***Preserve the order of data values***. Larger data values on the original scale will be larger on the transformed scale. (That's why negative powers have their sign reversed.)
- ***They change the spacing of the data values***. Powers $p < 1$, such as $\sqrt{x}$ and $\log x$ compress values in the upper tail of the distribution relative to low values; powers $p > 1$, such as x^2, have the opposite effect, expanding the spacing of values in the upper end relative to the lower end.
- ***Shape of the distribution changes systematically with*** p. If $\sqrt{x}$ pulls in the upper tail, $\log x$ will do so more strongly, and negative powers will be stronger still.
- ***Requires all*** $x > 0$. If some values are negative, add a constant first, i.e., $x \rightarrow t_p(x + c)$
- Has an effect only if the ***range of*** x ***values is moderately large***.

Ladder of Powers – Example

Baseball data - runs

- `SYMBOX` macro - transforms a variable to a list of powers, show standardized scores using the `BOXPLOT` macro

```
%include data(baseball);
title 'Baseball data: Runs';
%symbox(data=baseball, var=Runs, powers =-1 -.5 0 .5 1 2);
```



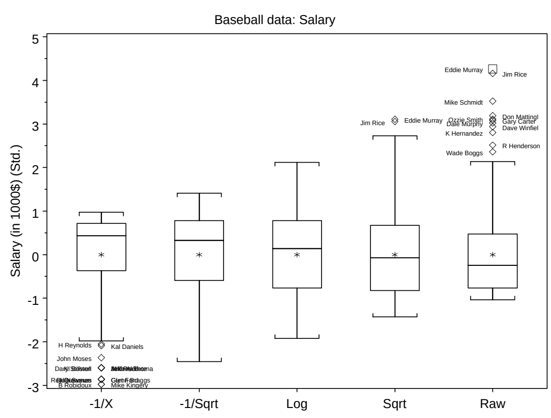
- runs $\longrightarrow$ $\sqrt{\text{runs}}$ looks best.

Ladder of Powers – Example

Baseball data - salary

- `SYMBOX` macro - transforms a variable to a list of powers, show standardized scores using the `BOXPLOT` macro

```
title 'Baseball data: Salary';
%symbox(data=baseball, var=Salary,
    powers =-1 -.5 0 .5 1, id=name);
```



- salary $\longrightarrow$ $\log(\text{salary})$ looks best.

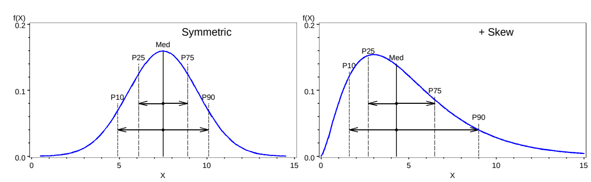
See `http://www.math.yorku.ca/SCS/sasmac/symbox.html`

Plots for assessing symmetry

Upper vs. lower plots

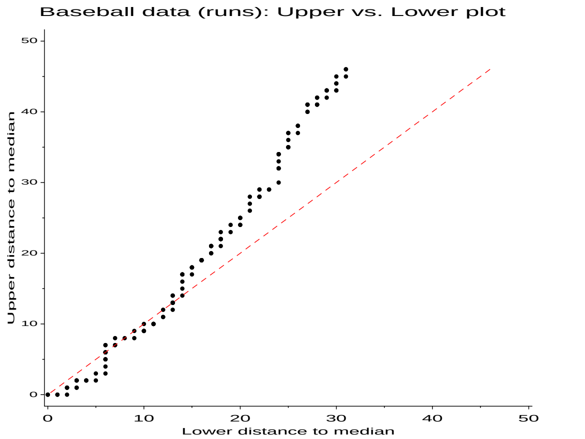
- In a symmetric distribution, the distances of points at the lower end to the median should match the distances of corresponding points in the upper end to the median.

$$\text{Lower distance to median} = \text{Upper distance to median}$$
$$\text{Med} - x_{(i)} = x_{(n+1-i)} - \text{Med}$$



- SYMPLOT macro - Upper vs. lower plot (`plot=UPLO`). Points should plot as a straight line with slope = 1 in a symmetric distribution.

```
title 'Baseball data (runs): Upper vs. Lower plot';
%symplot(data=baseball, var=runs, plot=uplo);
```



- For skewed distributions, the points will tend to rise above the line (positive skew) or fall below (negative skew).

Plots for assessing symmetry

Untilting: Mid vs. spread plots

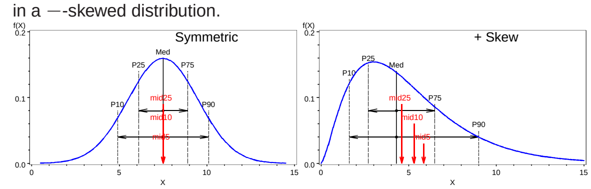
- In the Upper vs. Lower plot we must judge departure from symmetry by divergence from the line $y = x$.
- Change coordinates, so that the reference line for symmetry becomes horizontal. Rotate 45^o, by plotting:

$$\text{mid} \equiv [x_{(n+1-i)} + x_{(i)}]/2 \quad \text{vs.} \quad x_{(n+1-i)} - x_{(i)} \equiv \text{spread}$$

- In a symmetric distribution, each mid value should equal the median

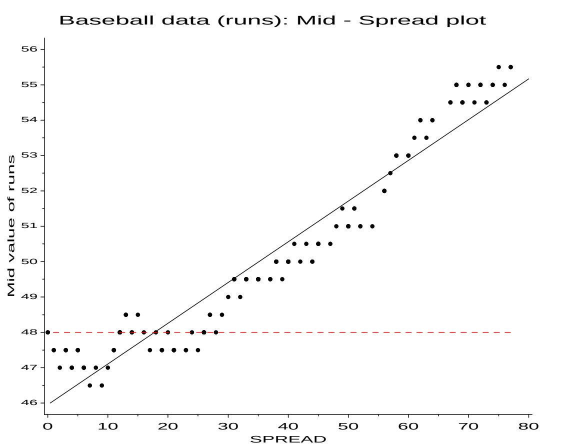
$$[x_{(n+1-i)} + x_{(i)}]/2 = \text{Median}$$

- Mid values will increase with i in a $+$-skewed distribution, decrease with i in a $-$-skewed distribution.



- SYMPLOT macro - Mid vs. spread plots (`plot=MIDSPRD`). Points should plot as a horizontal line with slope = 0 in a symmetric distribution.

```
title 'Baseball data (runs): Mid - Spread plot';
%symplot(data=baseball, var=runs, plot=midsprd);
```



- Because the plot is untilted (slope = 0) when the distribution is symmetric, expansion of the vertical scale allows us to see systematic departures from flatness far more clearly.

Plots for assessing symmetry

Power plot: Mid vs. z^2 plots

- Emerson and Stoto (1982) suggest a variation of the Mid vs. Spread plot, scaled so that a slope, b indicates the power $p = 1 - b$ for a transformation to approximate symmetry.
- In this display, we plot the centered mid value,

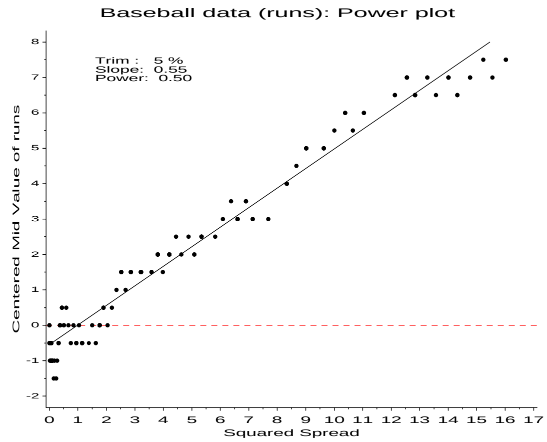
$$\frac{x_{(i)} + x_{(n+1-i)}}{2} - \text{M}$$

against a squared measure of spread,

$$z^2 \equiv \frac{\text{Lower}^2 + \text{Upper}^2}{4M} = \frac{[M - x_{(i)}]^2 + [x_{(n+1-i)} - M]^2}{4M}$$

- SYMPLOT macro - Power plots (`plot=power`). Points should plot as a horizontal line with slope = 0 in a symmetric distribution.

```
title 'Baseball data (runs):  Power plot';
%symplot(data=baseball, var=runs, plot=power);
```



- Symmetry is indicated by a line with slope=0 and intercept=0.
- The `SYMPLOT` macro rounds $p = 1 - b$ to the nearest half-integer.
- It is often useful to exclude (trim) the highests/lowest 5–10% of observations for automatic diagnosis.

See http://www.math.yorku.ca/SCS/sssg/symplot.html

Normal probability plots

- Compare observed distribution some theoretical distribution (e.g., the normal or Gaussian distribution)
- Ordinary histograms not particularly useful for this, because
  - they use arbitrary bins (class intervals)
  - they lose resolution in the tails (where differences are likely)
  - the standard for comparison is a curve
- **Quantile-comparison plots** (Q-Q plots) plot the quantiles of the data against corresponding quantiles in the theoretical distribution, i.e.,

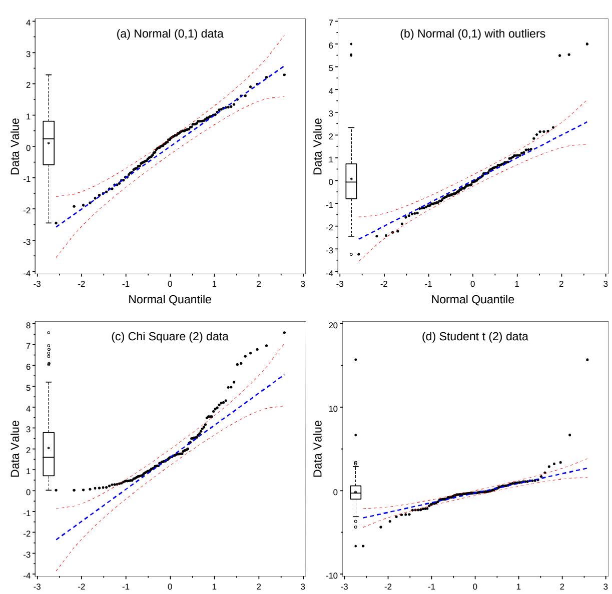
$$x_{(i)} \quad \text{vs.} \quad z_i = \Phi^{-1}(p_i)$$

  where $x_{(i)}$ is the i-th *sorted* data value, having a proportion, $p_i = \frac{i-1/2}{n}$ of the observations below it, and $z_i = \Phi^{-1}(p_i)$ is the corresponding quantile in the normal distribution.
- When the data follows the normal distribution, the points in such a plot will follow a straight line with slope = 1.
- Departures from the line shows *how* the data differ from the assumed distribution.

Normal probability plots

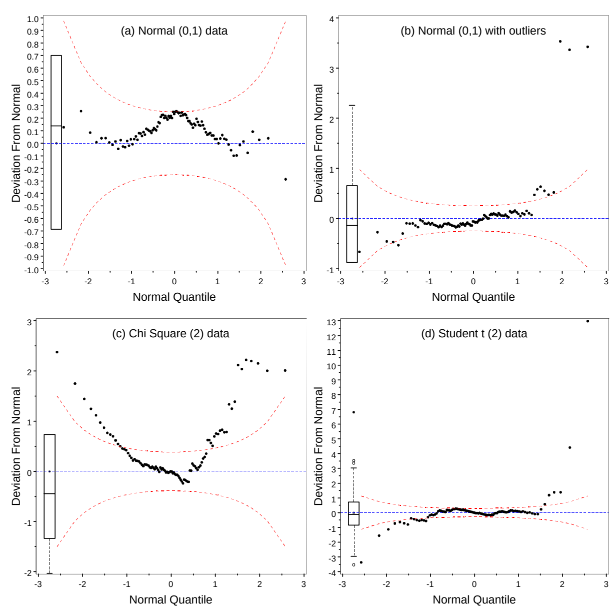
Patterns of deviation for Normal Q-Q plots:

- ***Postive (negative) skewed***: Both tails above (below) the comparison line
- ***Heavy tailed***: Lower tail below, upper tail above the comparison line



Normal probability plots

- De-trended plots show the deviations more clearly
- Plot $x_{(i)} - z_i$ vs. z_i.

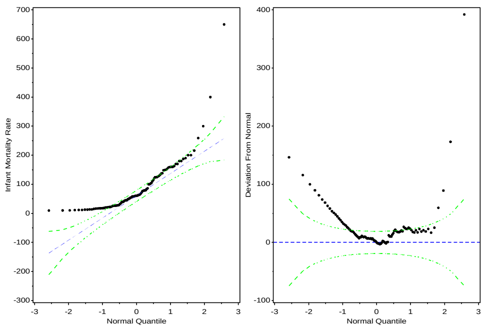


Normal probability plots: confidence bands

- Points in a Q-Q plot are not equally variable—observations in the tails vary most for normal.
- Calculate estimated standard error, $\hat{s}(z_i)$, of the ordinate z_i and plot curves showing the interval $z_i \pm 2\,\hat{s}(z_i)$ to give approximate 95% confidence intervals. (Chambers et al. (1983) provide formulas.)

$$\hat{s}(z_i) = \frac{\hat{\sigma}}{f(z_i)} \sqrt{\frac{p_i\,(1-p_i)}{n}}$$

- Confidence bands help to judge how well the data follow the assumed distribution



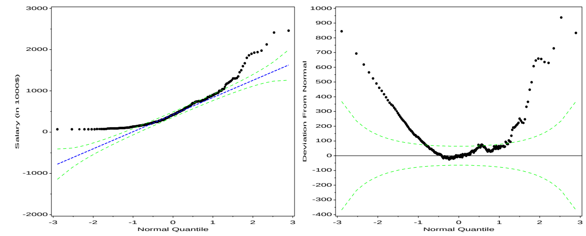
See `http://www.math.yorku.ca/SCS/sssg/nqplot.html`

Normal probability plots

Baseball data - salary

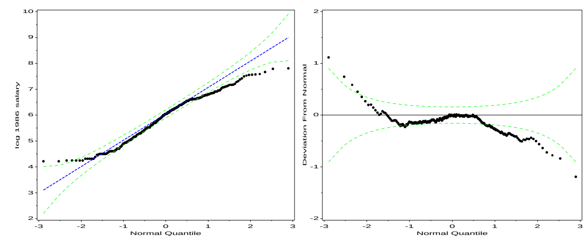
- Raw data

```
%nqplot(data=baseball, var=salary);
```



- Try $\log$ salary — better, but not perfect (who is?)

```
data baseball;
   set baseball;
   label logsal = 'log 1986 salary';
   logsal = log(salary);
%nqplot(data=baseball, var=logsal);
```



References

Box, G. E. P. and Cox, D. R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26:211–252, 1964. 19, 22

Box, G. E. P. and Tidwell, P. W. Transformation of the independent variables. *Technometrics*, 4:531–550, 1962. 22

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983. 36

Emerson, J. D. and Stoto, M. A. Exploratory methods for choosing power transformations. *Journal of the American Statistical Association*, 77:103–108, 1982. 31

Friendly, M. *SAS System for Statistical Graphics*. SAS Institute, Cary, NC, 1st edition, 1991. 18, 22

Gnanadesikan, R. and Kettenring, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

McGill, R., Tukey, J. W., and Larsen, W. Variations of box plots. *The American Statistician*, 32:12–16, 1978. 17

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, 1987.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977. 23