CLIP/CLASP WORKSHOP

Adam Carlson, Paul Cohen, David Hart, David Westbrook

Experimental Knowledge Systems Laboratory
University of Massachusetts, Amherst

CLASP

1. The clasp interface
2. Clasp for exploratory data analysis
3. Testing differences between groups
4. Testing for interactions between factors
5. Modeling the relative contribution of factors

CLIP

1. The clip interface
2. An experiment with TRANSSIM
3. Working groups

CLASP FOR EXPLORATORY DATA ANALYSIS

Purpose of EDA: Make your data give up its secrets

Tools: Lots of visualizations, descriptive statistics, partitioning and point coloring, tests of independence, transformations, smoothing.

Phoenix dataset:

SD	Shut-down time; time to contain a fire, or time at which trial is abandoned
WS	Wind speed; one of 3, 6, 9 kph
ScaledRTK	Ratio of thinking-time to environment-change time; one of .33, .53, .6, .8,1,1.75, 5
Plan1	The first plan Phoenix tries; one of MODEL, MBIA, SHELL
#Plans	The number or replanning events less one
FB	The amount of fireline built
Area	The area claimed by the fire at time SD
Success	1 if fire is contained in 120 hours or less, 0 otherwise
Util	The proportion of time during which the fireboss (planner) is busy

Basic operations

Open the dataset
Open a variable within the dataset
Describe a variable or dataset
Delete a result
Menu of options

Partitioning

Partition to look at successful trials only.

Two options:

```
    :Partition On (Dataset) PA-SCALED (Variable) success
    :Partition (Dataset) PA-SCALED (Partition Clause)
       (.and. (.<<. sd 120) (.==. success 1) )
```

Each produces a new dataset. The first gives data for successful trials, the second for successful trials for which shut-down time is less than 120 hours.

More examples soon...

Histograms and Point Coloring

Always look at the distribution of variables of interest.

```
:Histogram (x) Area
```

You may notice a gap in a histogram. This suggests the influence of another variable.

A trick to find the "other factor": Plot Area against Area in a Scatterplot, and color the points by the other variable:

```
:Scatter Plot (y) Area (x) Area (keywords) :Color-By-Sequence WS
```

Now you see that low area is associated with low wind speed. Now overlay the plots:

```
:Overlay Graphs (Graphs) GRAPH-DATA-ICON-10, GRAPH-DATA -
ICON-11,
```

and see how the gap in the histogram separates low and medium wind speed from high wind speed trials

Histograms and Point Coloring

You can also color histograms by another variable:

Scatterplots

To look at the joint distribution of two continuous variables:

```
:Scatter Plot (y) Area (x) FB
```

shows the relationship between the area burned and the amount of fireline built. Notice the anomalous vertical lines. Let's try to identify what's causing them:

```
:Scatter Plot (y) Area (x) FB (keywords)  :Color-By-Sequence PLAN1
```

Now you see two vertical lines are associated with a single initial plan, MODEL. Let's partition the data to look at those associated with the MODEL plan:

```
:Partition Dataset (Dataset) SUCCESS-1 (Partition Clause)
(.==. plan1 model) (Includes variables [default all])
```

Now run the scatterplot again with the MODEL data, coloring by, say, WS:

```
:Scatter Plot (y) Area (x) FB (keywords)  :Color-By-Sequence WS
```

Zooming

It looks like we have three clusters, one for each wind speed, but it's difficult to see if the leftmost one is a lot of points or just a few.

Zoom!

Move the mouse into the graph, click right, select Zoom In, draw a box around the desired points.

You can do it a second time to expand them even more.

Click on a point in the scatterplot to get its coordinates.

In sum, when Plan1=MODEL and WS = 3, FB = 11km (roughly); when WS = 6, FB = 17km; when WS = 9, FB = 27km. Is the MODEL plan efficient?

You can Move the mouse into the graph, click right, select Zoom Out to get back to where you were.

LinePlots

Line plots give the value of a function over some range.

Here's a nice application: When you plot a histogram, its appearance, and particularly the appearance of gaps, depends on bin size. To find out whether a gap exists at any bin size, do this:

```
:Sort (Variable to sort) Area-2
```

```
:Line Plot (y) SUCCESS-1-ROW-NUMBER (x) SORTED-AREA-2
```

Horizontal (or nearly horizontal) regions of this function say sorted area increases but the number of cases doesn't. For example, sorted area increases from 10 to 20 but there are almost no cases in this region. Hence we would expect to see a gap in a histogram between Area = 10 and Area = 20.

Should our histogram fail to disclose a gap, we can change bin size as follows:

Select the legend in the legend box and click left. A pop-up window of graph parameters will open. Change bin size or number of bins, and click OK at the bottom of the box. The graph will redisplay. Note you can "zoom" by selecting max and min values.

Other uses of lineplots: Time series, as we'll see later

Univariate Descriptive Statistics

All the usual statistics can be had (mean, median, mode, standard deviation, etc.)

```
:Statistical Summary (x) SD
```

Robust Statistics

Some robust statistics are available: Trimmed mean, Median, Interquartile range.

Let's do an experiment to show how these are robust. We will use the SAMPLE> menu and functions:

```
:Sample Normal (Mean) 0 (Standard Deviation) 1 (Size of
Samples) 100 (Number of Samples) 1
```

```
:Sample Normal (Mean) 10 (Standard Deviation) 1 (Size of
Samples) 10 (Number of Samples) 1
```

```
:Merge Datasets (Datasets) NORMAL-DISTRIBUTION-1,NORMAL -
DISTRIBUTION-2
```

Now plot the histogram of NORMAL-DISTRIBUTION in the merged data set

Robust Statistics, cont.

Get the trimmed mean of two datasets—the one that includes the outliers and the one that doesn't (both will have the same name but in different datasets):

`:Trimmed Mean (x)` NORMAL-DISTRIBUTION,NORMAL-DISTRIBUTION
`(trimming factor (real from 0 to .5))` .25

Now look at the means of the same distributions—the one that includes the outliers and the one that doesn't (both will have the same name but in different datasets):

`:Mean (x)` NORMAL-DISTRIBUTION,NORMAL-DISTRIBUTION

You can see that the mean is seriously affected by the outliers but the trimmed mean is not. The trimmed mean is robust against outliers. So is the interquartile range and so is the median.

Bivariate Statistics

The usual ones are available: Covariance, correlation, the regression line.

For time series we also have autocorrelation and cross correlation.

For categorical data we have contingency tables, chi-square and G

We'll describe these in order.

Covariance, Correlation, Simple Regression

Lots of options:

:Covariance (y) SD (x) FB

:Correlation (y) SD (x) FB

:Linear Regression Brief (y) SD (x) FB

:Linear Regression Verbose (y) SD (x) FB
(you can open up the scatterplot, later)

:Regression Plot (y) SD (x) FB

Regression and Transformations

Often, relationships are not linear but can be transformed to be linear.

```
:Regression Plot (y) SD (x) SCALEDRTK
```

Note that this plot is concave downward, plus the points on the left are bunched up and then there's a big empty area. A log transform spreads the points better:

```
:Natural Log (Variable) SCALEDRTK
```

```
:Regression Plot (y) SD (x) LOG-SCALEDRTK
```

The regression line appears to fit these points better. To confirm:

```
:Linear Regression Brief (y)  SD,SD (x)  SCALEDRTK,LOG-SCALEDRTK
```

(this performs two regressions, SCALEDRTK on SD and LOG-SCALEDRTK on SD)

You see that the coefficient of determination is higher for the transformed data; Log(SCALEDRTK) accounts for more of the variance in SD (32%) than SCALEDRTK does (23%).

Regression and Outliers

Linear regression, correlation, and many other statistics are very sensitive to outliers. You can see a few in the plot of LOG-SCALEDRTK against SD.

To get rid of outliers, simply partition the dataset, selecting the variables we want and the values we want:

```
: Partition (Dataset) SUCCESS (Partition Clause) (.<<. SD 115)
(Includes variables [default All]) LOG-SCALED-RTK, SCALEDRTK,
SD
```

This gives us a new dataset with just three variables, including only those data for which shut-down time is less than 115 hours

Now we can run the regressions again. We will find that LOG-SCALED-RTK accounts for 39.7% of the variance in SD and SCALEDRTK accounts for 27.4% of the variance in SD.

If we compare Data Length (x) SD for the original and partitioned data we find we have deleted four outliers from 215 points. Thus, 2% of the data reduces the coefficient of determination from roughly 40% to roughly 32%. BEWARE!

Time Series

```
:Load Dataset (Pathname) pauls-time-series
```

```
:Row Line Plot (y) SHIPS-IN-DOCK
```

Messy, huh?

Smooth the series:

```
:Smooth Variable 4253h (x) SHIPS-IN-DOCK
```

```
:Row Line Plot (y) SMOOTH-OF-SHIPS-IN-DOCK
```

Looks better ... Now overlay them:

```
:Overlay Graphs (Graphs) GRAPH-DATA-ICON-19, GRAPH-DATA -
ICON-18
```

Maybe this will look nice, maybe not. You can zoom, as described earlier. You can also click the legend, open a graph-description window, and edit colors etc.

Smoothing Series

You can smooth series in other ways than 4253h by using clasp lisp functions directly:

```
(setf a (smooth-median-4 SHIPS IN DOCK))
```

or

```
(setf a (smooth-median-2 SHIPS IN DOCK))
```

And of course you can repeat these commands simply by selecting them and clicking in the clasp window. The net effect is to smooth the smoothed series, repeatedly.

Mean smooths haven't been implemented yet.

Autocorrelation and Cross Correlation

Consider the relationship between FULL-SHIPS and SHIPS-IN-DOCK: the former eventually become the latter and the latter eventually become the former. How strong is the predictive relationship between them?

`:Cross Correlation (y)` FULL-SHIPS `(x)` SHIPS-IN-DOCK `(min-lag)` -5 `(max-lag)` 5

You see the strongest relationship is for a lag of zero, which is to be expected: if ships are in dock, they aren't full at sea, and vice versa.

It is often more informative to look at the first derivative of series—how the number of ships in dock and at sea change over time, and how these are correlated:

`:Discrete Derivative (x)` FULL-SHIPS, SHIPS-IN-DOCK

`:Cross Correlation (y)` DIFFERENCED-FULL-SHIPS `(x)` DIFFERENCED-SHIPS-IN-DOCK `(min-lag)` -5 `(max-lag)` 5

Once again we see a negative cross-correlation at lag zero, but we also see a positive correlation at lag -3. This suggests that DIFFERENCED-FULL-SHIPS on thursday is positively correlated with DIFFERENCED-SHIPS-IN-DOCK on monday.

Autocorrelation and Cross Correlation

The previous result suggests a three-day turnaround time for ships in dock. This impression is confirmed by an autocorrelation of DIFFERENCED-SHIPS-IN-DOCK:

```
:Autocorrelation (x) DIFFERENCED-SHIPS-IN-DOCK (min lag
[default 0]) 0 (max lag) 13
```

lag 0 is positive, lag 3 negative
lag 4 is positive, lag 7 negative

Note that you don't get this result if you look at SHIPS-IN-DOCK instead of DIFFERENCED-SHIPS-IN-DOCK.

The reason is yesterday and today are always highly correlated, yesterday and tomorrow slightly less so... The autocorrelation function for a raw series is swamped by "inertia" or "trend." Differencing removes trend.

Bivariate Distributions of Categorical Data

Some dependent variables, such as success, are categorical. Similarly, some independent variables, such as wind speed or first plan are interval (with few bins) or ordinal (with few levels) or categorical.

Contingency tables are cross-tabulations of these variables.

Go back to the original pa-scaled data:

```
:Chi-Square Rxc (y) SUCCESS (x) WS
```

This gives the contingency table for the number of trials at each pairing of success (levels are success, failure) and wind speed (levels are 3, 6, 9 kph)

Percentages are also shown and easier to interpret.

Expected values are also shown, and will be discussed later.

Recoding Continuous Data as Categorical

CLASP FOR HYPOTHESIS TESTING

So far we have looked at descriptive statistics, now we change our focus to answering questions.

Are two groups different?

Are several groups different?

Are two variables independent?

Do two variables interact to produce an effect on a third?

Testing Whether Two Groups Are Different

Phoenix knows three general plans, MODEL, SHELL and MBIA, and we saw earlier evidence that the amount of area lost to fires when using the MODEL plan is perhaps higher than the amount lost to the others.

Partition the dataset PA-SCALED into two groups of successful trials:

```
:Partition Dataset (Dataset) PA-SCALED (Partition Clause)
(.and. (.==. success 1) (.==. plan1 model)) (Includes variable [default
all]) AREA
```

```
:Partition Dataset (Dataset) PA-SCALED (Partition Clause)
(.and. (.==. success 1) (./=. plan1 model)) (Includes variable [default
all]) AREA
```

This will produce two datasets each including a variable called AREA-1, but one variable will contain fires fought under the MODEL plan and the other will contain fires fought under the SHELL and MBIA plans

```
:T Test Two Sample (y) AREA-1 (x) AREA-1 (Tails: [Both,
Positive or Negative]) Both
```

This runs a two-sample t test on the two groups and returns a highly significant result. Run statistical summary to see which plan loses the more area.

The D Test

An alternative to the T test is the D test. It is based on computer-intensive randomization resampling. To estimate the sampling distribution of the difference of two means:

1. throw all the data for the two samples A and B into a single bucket
2. shake it up
3. draw two new "pseudosamples" A* and B*
4. calculate the means of A* and B* and their difference, D*
5. repeat these steps until you have several hundred values of D*. These are the estimated sampling distribution of the difference of two means under the null hypothesis that the samples A and B were drawn from the same population.

```
:D Test (y) AREA-1 (x) AREA-1 (Tails: [Both, Positive or
Negative]) Both
```

You will see that the significance level is effectively zero; in other words, there's really no way to have achieved the sample difference D = mean(A) - mean(B) by chance under the null hypothesis. Therefore, reject it.

Testing Whether Several Groups are Different

Let's look at three wind speeds and their affect on area burned for successful trials:

```
:Anova One Way Variables (y) AREA-2 (x) WS-2
```

The one way analysis of variance provides the usual results: sums of squares, mean squares, an F ratio and a p value. Clearly, different amounts of area were burned at different wind speeds.

Exhaustive Pairwise Comparisons of Means

The analysis also provides a plot of means, and also provides Scheffe pairwise comparisons. You can see that all pairwise comparisons are significant. Clicking on the asterisks gives the value of the Scheffe statistic and its p value.

(Axes, rows and columns will be labelled in the next release.)

Testing Whether Variables are Independent

Clearly, a significant one way anova tells us that the x and y variables are not independent: Wind speed affects area burned.

An alternative, appropriate for continuous variables, is to ask whether the correlation of the variables is significantly different from zero.

Fisher's r to z transform is often used to test this hypothesis.

We will use clasp's easy lisp interface to write a randomization test.

1. Let R be the correlation of two samples A and B
2. Shuffle the elements of B yielding B* (or A) and calculate R*, the correlation of A and B*
3. Do this several hundred times. The distribution of B* estimates the sampling distribution of the correlation under the null hypothesis that A and B are independent.
4. Sort the distribution of B* to find the percentile at which B lies. This is your significance level.

Testing Whether Variables are Independent

Anova tells us whether x, with a smallish number of levels, is independent of continuous y. Tests on correlations tell us whether continuous x is independent of continuous y. What about x and y, both with a smallish number of levels?

Is success independent of wind speed?
Is plan1 independent of wind speed?

Chi-square tests tell us whether the distribution of counts in a contingency tables are significantly different that we'd expect by chance.

```
:Chi-square Rxc (y) SUCCESS (x) WS
```

The distribution is marginally different than we'd expect by chance (p = .0688)

```
:Chi-square Rxc (y) FIRSTPLAN (x) WS
```

(use FIRSTPLAN because it's a number and chi-square can't handle category labels yet)

There's no evidence that these variables are not independent.

Testing Whether Variables are Independent

Sometimes we want to test whether a frequency distribution differs from a standard. For example, we think roughly equal numbers of fires should be fought at each wind speed (because we set WS ourselves). Let's check whether this is credible:

The total numbers of fires fought at each wind speed are 120, 117 and 106 for wind speeds of 3, 6 and 9 kph, respectively (see previous analysis)

343 fires, total, were fought. If equally divided among wind speeds the number fought at each should be 343/3 = 114.33

Into lisp:

```
(setf a (make-array '(2 3) :initial-contents '((120 117 106)(114.33 114.33 114.33))))
```

```
(chi-square-rxc-counts a)
```

Two numbers are returned: the value of chi square for the table and the p value for the statistic. You can see there is no evidence to suggest that wind speed is not distributed evenly over trials.

Effects of Two Variables on a Third

You have two variables, x1 and x2 that might affect y. A one-way analysis of variance tells you whether x1 (or x2) affects y. What if y is affected by some interaction of x1 and x2?

Start with SUCCESS-=-1

```
:Anova Two Way Variables (y) SD-2 (x1) WS (x2)
RECODEOF#PLANS
```

You can see a significant main effect of RECODEOF#PLANS, no interaction effect, and a marginal effect of WS.

```
:Anova Two Way Variables (y) NUMPLANS-2 (x1)
CATEGORICALRTK-2 (x2) PLAN1-2
```

Here you can see main effects of RTK and PLAN1, and also an interaction effect: the mean amount of replanning increases with RTK, but more quickly for MBIA than SHELL and more quickly for MODEL than MBIA.

Additional Topics

- Modeling with multiple linear regression
  - one-level models
  - multi-level models
- Parameter estimation
  - Confidence intervals
- Sampling