Graphical Methods for Data Analysis & Multivariate Statistics

Michael Friendly
Psychology 6140
Why plot your data?

Graphs help us to see

patterns, trends, anomalies and other features

not otherwise easily apparent from numerical summaries.

Source: http://xkcd.com/523/
How graphs can change your life (n=1)

Personal analytics

15 yr. blood sugar, pre-diagnosis

A statistician contracts diabetes, and uses graphs to monitor his blood sugar.

Visual feedback on diet & exercise reinforce behavioral change

→ Residual plots show unexplained events, possibly important

daily average, after diagnosis


average hourly variation

residuals: - daily average and hourly
Different graphs for different purposes

Graphs (& tables) as communication:
• What audience?
• What message?

• **Analysis graphs**: design to see patterns, trends, aid the process of data description, interpretation

• **Presentation graphs**: design to attract attention, make a point, illustrate a conclusion
Different graphs for different purposes

Presentation graphs: single image for a large audience
Exploratory graphs: many images for a narrow audience (you!)
Comparing groups: Analysis vs. Presentation graphs

Six different graphs for comparing groups in a one-way design
- which group means differ?
- equal variability?
- distribution shape?
- what do error bars mean?
- unusual observations?

Never use dynamite plots
Always explain what error bars mean
Consider tradeoff between summarization & exposure
Florence Nightingale: Deaths in the Crimean war from battle vs. other causes (disease, wounds)

She used this to argue for better field hospitals (MASH units)

The best presentation graphs pass the **Interocular Traumatic Test**: The message hits you between the eyes!
Presentation: Turning tables into graphs

Graphs of model coefficients are often clearer than tables

Source: tables2graphs.com
Common Sense Revolution


(National Council of Welfare & the Toronto Disaster Relief Committee)

© Scott Sparli
Side-by-side boxplots of variables in the baseball data show the shapes of distributions --- aid to transformation

- Each variable is standardized to allow comparison.
- Plot is produced by `datachk` macro.

See: [http://datavis.ca/sasmac/datachk.html](http://datavis.ca/sasmac/datachk.html)
Exploratory graphs: Transformations

Data often needs to be transformed to meet analysis assumptions:

- Symmetry (~ Normality)
- Linear relations
- Constant variance

For symmetry, a **symbox** plot shows a variable transformed to various powers.

SAS: symbox macro
R: car package: symbox()
Diagnostic plots can be used to suggest corrective action, often by a power transformation: $y \rightarrow y^p$

**Symmetry transformation plot:**

- Constructed so symmetric data plots as horizontal line
- Slope ($b$) of data line → power: $p = 1 - b \rightarrow y^p = y^{(1-b)}$

Other diagnostic plots use the same idea: slope ($b$) → $y^{(1-b)}$
Model diagnosis: Influence in regression

Multiple regression model: prestige ~ income + education

Influence plots can show:
- model residual
- leverage (potential impact)
- influence ~ residual x leverage (Cook D statistic)
- contour map of influence
Model diagnosis: regression quartet

Statistical software should make it easy to get informative diagnostic plots

In R, plotting a \texttt{lm} model object → the “regression quartet” of plots

\begin{verbatim}
> model <- lm(prestige ~ income + education)
> plot(model)
\end{verbatim}

(SAS has similar, using ODS graphics)
Scatterplots: A basic workhorse for quantitative data

- Show the relation between two Q variables (ignoring all others!)
- More useful when enhanced to show visual summaries
- Vary point color/shape to show strata/groups
- Combine in multi-panel displays to show more
  - Scatter plot matrix: all pairs
  - Conditional relations: Y vs. X stratified by Group
Scatterplots: Scales matter

Computer plots are usually generated with a given \textit{aspect ratio}, to conform to the page or screen.

A better idea is to scale the plot so that slopes of lines or curves average \textasciitilde 45 degrees.

In the \textit{rescaled} version, we can see that, within each cycle, sunspots tend to increase more quickly than they decline.
Scatterplots: Annotations enhance perception

Data from the US draft lottery, 1970

• Birth dates were drawn at random to assign a “draft priority value” (1=bad)
• Can you see any pattern or trend?

Me (May 7): 127 → priority = 35
Scatterplots: Annotations enhance perception

Drawing a smooth curve shows a systematic decrease toward the end of the year.

- The smooth curve is fit by loess, a form of non-parametric regression.

Visual explanation:
Scatterplots: Data ellipses

Galton’s (1886) semi-graphic table, showing relation of mid-parent’s height to children’s height.

As shown:
- Contours of equal frequency formed ellipses
- Regression lines of Y on X and X on Y are the loci of vertical and horizontal tangents
- Major/minor axes are the principal components
Scatterplots: Data ellipses

Galton’s data on child & mid-parent heights, shown as a sunflower plot: each sunflower symbol shows the number of observations in the (x, y) cell.

2D density estimate of bivariate surface
Any scatterplot can be summarized by data ellipses (assuming normality). These show: means, standard deviations, and allow correlations & regression lines to be visually estimated.

Data ellipse:

$$D^2(y) \approx \chi^2_p(1 - \alpha)$$

Galton data, 40%, 68% & 95% data ellipses. Sizes are:
- $$\chi^2(0.40) = 1.0$$
- $$\chi^2(0.68) = 2.28$$
- $$\chi^2(0.95) = 6.0$$
Visualizing multivariate data

Showing relations among 3 or more variables:

• Scatter plot matrices (enhance with visual summaries, thin for many variables)
• Conditional plots: $Y \sim X \mid (Z, \text{Group})$
• Seeing multivariate profiles, clusters:
  ▪ Star plots, face plots, parallel coordinates
• Biplots: project data into low-D view
Scatter plot matrix

- **Fitness data:**
  - Oxy $\sim$ Age + Weight + Runtime + Rstpulse
- Each panel shows row var vs. col var
- Reg line shows *linear* relation

**Questions:**
- What is the best predictor of Oxy?
- Which two predictors are most highly correlated?
• **Occ. prestige:**
  
Prestige ~ %women + Educ + Income

• Box, rug plots show univar. distributions

• Quadratic regressions show linear/non-linear relations (loess would be better)

Questions:
- How should Educ be modeled?
- How should Income be modeled?
Larger data sets: Visual thinning

Baseball data: log(Salary) ~ performance variables

- Too much data to show individual points
- Each scatterplot is summarized by a loess smoothed curve and a data ellipse

Questions:
- Which variables are most strongly related to logSal?
- Which relations are strongly nonlinear?
- Which predictors are too highly correlated?
Larger data sets: Corrgrams

Correlation diagram shows **pattern** of correlations for many variables.

Variables are re-ordered to make the groupings most visually apparent.

This graphic assumes that all relations are linear, not necessarily always true.

Graph using SAS `corrgram` macro, [http://datavis.ca/sasmacro/corrgram.html](http://datavis.ca/sasmacro/corrgram.html)
Corrgrams: Different renderings

The value of a correlation may be rendered in different ways, with different visual impact.

- Shading levels: help detect similar values
- Pie symbols: make it easier to compare for larger/smaller

Graph using R **corrgram** package
Conditional plots: $Y \sim X \mid Z$

Often want to explore how the relation between $Y$ and $X$ depends on/ varies with some other variable(s) $Z$.

- Moderator variables
- Interactions

Emission of NOx from ethanol in relation to engine compression ratio and richness of air/ethanol mixture (EE)

Graph using R `lattice` package
Conditional plots: Y \sim X \mid Z

The same data is shown in a different format, with

- loess smooth curves
- curves banked to \sim 45^\circ

The joint dependence on CR and EE is now much clearer

(These are examples of lattice plots, produced using R software.)
3D plots

Often not useful, unless done with great care.

This plot shows the loess \textit{smoothed} predicted values of NOx in relation to EE and CR. (But, raw data not shown.)

Color is used to show the predicted NOx, using a “heatmap” color scale.
3D plots can be enormously useful with dynamic, interactive software & perspective

This plot shows a relation of occupational prestige to income & education.

• points are shown in perspective, connected to the fitted surface
• the fitted surface (linear, quadratic, smoothed) can be changed interactively
• the plot can be rotated dynamically to see other views
Seeing multivariate clusters: face plot

A faces plot assigns variables to facial features, to show **configural patterns** of many variables.

**Pros:** Easy to see similar patterns in large data sets.

**Cons:**
- Hard to connect features to variables for interpretation
- No good rules/ideas for assigning variables to features.

Graph using SAS **faces** macro, [http://datavis.ca/sasmac/faces.html](http://datavis.ca/sasmac/faces.html)
### Seeing multivariate clusters: face plot

#### Means, by make & origin

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Left Side Variable</th>
<th>Right Side Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye size</td>
<td>mpg</td>
<td>mpg</td>
</tr>
<tr>
<td>Pupil size</td>
<td>mpg</td>
<td>mpg</td>
</tr>
<tr>
<td>Pupil position</td>
<td>turn</td>
<td>turn</td>
</tr>
<tr>
<td>Eye slant</td>
<td>turn</td>
<td>turn</td>
</tr>
<tr>
<td>Eye X position</td>
<td>hroom</td>
<td>hroom</td>
</tr>
<tr>
<td>Eye Y position</td>
<td>hroom</td>
<td>hroom</td>
</tr>
<tr>
<td>Eyebrow curvature</td>
<td>rseat</td>
<td>rseat</td>
</tr>
<tr>
<td>Density of eyebrow</td>
<td>rseat</td>
<td>rseat</td>
</tr>
<tr>
<td>Eyebrow X position</td>
<td>dispa</td>
<td>dispa</td>
</tr>
<tr>
<td>Eyebrow Y position</td>
<td>length</td>
<td>length</td>
</tr>
<tr>
<td>Upper hair line</td>
<td>rep77</td>
<td>rep78</td>
</tr>
<tr>
<td>Lower hair line</td>
<td>weight</td>
<td>weight</td>
</tr>
<tr>
<td>Face line</td>
<td>weight</td>
<td>weight</td>
</tr>
<tr>
<td>Hair darkness</td>
<td>rep77</td>
<td>rep78</td>
</tr>
<tr>
<td>Hair shading slant</td>
<td>gratio</td>
<td>gratio</td>
</tr>
<tr>
<td>Nose line</td>
<td>length</td>
<td>length</td>
</tr>
<tr>
<td>Mouth size</td>
<td>price</td>
<td>price</td>
</tr>
<tr>
<td>Mouth curvature</td>
<td>price</td>
<td>price</td>
</tr>
</tbody>
</table>

#### Faces Plot of Automobile data

- **Means by Origin, Make**
  - American
  - European
  - Japanese

![Face Plot Diagram](image-url)
Biplots: variables and obs. in low-D View

- Based on PCA: data is shown in 2D (3D) view that accounts for greatest variance
- Observations: plotted as points
- Variables: vectors from origin (=mean)
- Angles between vectors ~ correlations
- Projection of point on vector ~ score
Biplot: US crime rates

Dim1: ~ Overall crime rate
Dim2: Property vs. personal

Note: clusters of southern, New England, western states

This 2D biplot only shows 76.5% of total variance.

Still, it gives a useful summary of 9 variables and 50 observations.
Biplot: Baseball data

Baseball hitters’ data:
• Dim2: fielding, -years
• Dim1: batting performance

Players identified by position, with data ellipses for each
• IF: more assists, errors
• DH: older

This 2D biplot only shows 63.7% of total variance.
HE plots for MANOVA, MMReg

HE plots provide a way to visualize hypothesis tests in MANOVA and multivariate multiple regression, using data ellipses for fitted (H) and residual (E) co-variances.

**Graphic ideas:** (a) Data ellipses summarize H & E (co)variation; (b) Scale H ellipse so it projects outside E ellipse *iff* effect is significant (Roy test)
HE plot matrices

HE plots in a scatterplot matrix show effects for all pairs of responses.

For the iris data, the Species means are highly correlated on all variables except Sepal length.
HE plots: 2-way MANOVA

Plastic film data: 2x2 MANOVA
(gloss, opacity, tear) ~ rate*additive

MANOVA tests show that both main effects are significant:

Type II MANOVA Tests: Roy test statistic

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>approx F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rate</td>
<td>1</td>
<td>7.5543</td>
<td>0.003034**</td>
</tr>
<tr>
<td>additive</td>
<td>1</td>
<td>4.2556</td>
<td>0.024745 *</td>
</tr>
<tr>
<td>rate:additive</td>
<td>1</td>
<td>1.3385</td>
<td>0.301782</td>
</tr>
</tbody>
</table>

HE plot shows the nature of these effects, e.g.,
high rate: ↑tear, ↑opacity, ↓gloss

1 df tests: H ellipsoid collapses to a line
Rohwer data: Cognitive ability and PA tests: $n=37$, Low SES group

(SAT, PPVT, Raven) ~ $n + s + ns + na + ss$

• Only one predictor, NA, is (barely) significant

• Yet, overall multivariate test: $H_0$: $B = 0$ is highly so!
HE plots: MMRA & MANCOVA

Rohwer data: Low SES & Hi SES groups

(SAT, PPVT, Raven) ~ SES + n + s + ns + na + ss
Dynamic, interactive graphics

Interactive graphics & data analysis provides:

- Identifying points
- Model & display controls

SAS/Insight: mpg ~ weight, linear fit
Dynamic, interactive graphics

Interactive graphics & data analysis provides:

- Identifying points
- Model & display controls

**SAS/Insight**: mpg ~ weight, **quadratic** fit
Dynamic, interactive graphics

Dynamic graphics provide multiple, linked views of a data set

Selecting points, regions in one plot (“brushing”) selects the same observations in all other plots

Image source: Data Desk (Paul Velleman)

Multivariate frequency data: mosaic plots

A contingency table can be visualized by tiles whose area ~ cell frequency.

**Shading:** ~ Pearson residual,

\[ d_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} \]

**Color:**
- blue: \( O_{ij} > E_{ij} \); red: \( O_{ij} > E_{ij} \)

Interp: + association (dark hair, dark eyes), (light hair, light eyes)
N-way tables

3+ way tables: split each tile ~ conditional proportions of the next variable.

Now, there are several different models that can be fit.

• Mutual independence: [H][E][S] → all vars unassociated

• Residuals: show associations not acct’d for by the model
N-way tables

Conditional independence: [Hair, Sex][Eye, Sex]

All models fit to the same table have \textit{same}-sized tiles \((O_{ijk})\), but \textit{different} residuals.

This model of conditional independence, [HS][ES] \(\rightarrow\) H, E independent \textit{given} Sex.
N-way tables

The model of joint independence, \([\text{HE}]\text{[S]}\) allows Hair, Eye color association, but \(\rightarrow [\text{HE}]\) assoc. is independent of Sex.

This model obviously fits much better, except for blue-eyed blonds, where females are more prevalent than the model allows.
Summary

• Goal of statistical analysis: summarization
• Goals of graphical analysis: exposure!
  ▪ Often more useful when enhanced with visual summaries (fitted curve, data ellipse)
• Different graphs for different purposes:
  ▪ Reconnaissance (overview)
  ▪ Exploration (detecting patterns, trends)
  ▪ Model diagnosis (assumptions, outliers)
Summary

• Multivariate data requires novel graphs to display increasing # of variables
  ▪ Enhanced scatterplot matrices
  ▪ Visual thinning: less is often more
  ▪ Low-D views (biplots)
  ▪ HE plots to visualize multivariate tests
  ▪ Mosaic plots to visualize n-way frequency tables.