Chapter 8 in new version of the text.

Chpt 9 Gentle

Graphical Methods in Comp. Stats.

One of the 1st steps in attempting to understand data is to visualize it. This provides a wealth of psychological tools that can be used to detect features, discover relationships, & retain knowledge gained.

Graphical displays have always been an important part of statistical data analysis → but with the advancement of computing their roles, have greatly increased. and usefulness

The number of variables & number of observations play a role in determining the way graphical displays are constructed.

- Data of ~~two~~ three or fewer dim can be portrayed in 2-d easily.
- higher dims. require projections, transformations and other techniques.
- large data sets are sometimes viewed in pieces.

Section 8.2

See 8.1 Viewing one, two, or three Variables.

Plots of one or two variables are easy to construct and interpret. & often we are able to use some of these same tech. w/ 3 variables. For data sets w/ more variables it is sometimes useful to look at 1, 2 or 3 at a time or projections into a 1, 2 or 3 variable subspace.

* graphs that represent the density have one more dim than the data?

Histograms

page 345

One of the most important properties of data is the shape of its distribution or density.* The basic tool for looking at the shape of the dist. of univariate data is the histogram.

Histogram
- a graph of the counts or the rel. freq. of the data w/in contiguous regions called bins
- the vertical axis is the counts in the bins or proportions so that the total area adds up to 1.
- the formation of the bins is fundamental to visualizing & understanding the data.

Choice of bins.
- Often binwidth is uniform, but this is unnecessary, ex. if there are only a small # of obs over a wide interval, the bins in that range can be made wider to smooth out the roughness & variation in the small counts.

move to next page.

(1) - # of bins can markedly affect appearance
Ex Fig 8.2 30 points from a gamma w/ shape para. 3 and scale para. 10. & fixed bin width.

Figure 8.2

- So how many bins?
  - too few - obscure structure
  - too many - rougher appearance - can be difficult to ascertain patterns

- In general the # of bins ↑ w/ ↑ obs.
  - A simple rule of thumb is to use $1 + \log_2 n$ bins when you have n obs.

* (2) → Insert from previous page. Binwidth

Figure 8.3

Ex/ Fig [redacted] Variable bin width.
- Same as (c) in 8.2 [redacted] but last two are combined to smooth.

(3) - Cutpoint location. can affect view

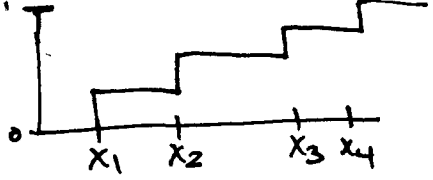
Ex/ 8.2 Fig [redacted] (c) cut points at 0, 10, 20, ..., 70 if we shift these by 2 to 2, 12, 22, ... 72 we obtain Fig 8.4 [redacted] (b). ∴ further shifting results in even diff histograms.

So when using histograms as a comp. stats. tool one should consider a # of different views (bin: #, size, location) with the emphasis on exploration and not confirmation of hypothesis or presentation.

Page [redacted]

The Empirical Cumulative Distribution Function or ECDF

The ECDF is one of the most useful summaries of a univariate sample. It is a step function w/ an increase of size 1/n at each point in a sample of size n.

Ex/ 

A variation of the ECDF that is often more useful is the broken line ECDF. Here lines connect the points $(x_i, i/n)$. (x_i sorted)

Ex/ Fig 8.5 (a)

Another variation - is the mountain plot, where the ECDF is folded at the median. It is often easier to see certain prop. such as symmetry in a mountain plot.

Ex/ Fig 8.5 (b).

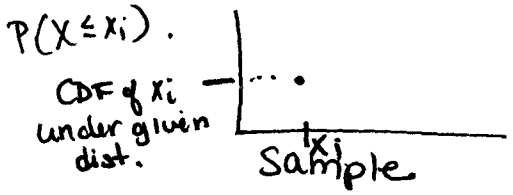
The plot of the ECDF provides a simple comparison of the sample to the uniform dist. → if $X \sim U$ then the broken-line ECDF is a straight line & the mountain plot is an isosceles Δ.

Also, the ECDF of a unimodal sample is concave. Multimodal - convex in some areas concave in others.

Ex/ 8.5 is a skewed unimodal pattern.

Q-Q plots & Prob. plots.

If the vertical axis is transformed so that it corresponds to the cum dist function of a given dist'n D then it is easy to compare the sample to it. The b-l ECDF will be close to a straight line if the sample is from D.



(xi, Eval CDF of xi w/ D).

This type of plot is called a probability plot.

A related plot is the quantile-quantile plot. Here the quantiles (or "scores") of the ref dist are plotted against the sorted data.

Basically: the 1/nth quantile is plotted against the 1st order statistic in the sample of size n & so on.

To calculate the empirical quantiles, the kth smallest value should corres to a value of p approx equal to k/n or (k-1)/n. A common way is to just split the diff & use

$$p_k = (k-1/2)/n$$

Then the kth smallest value is the p_kth sample quantile.

8.6

Ex/ Fig [redacted] q-q plot corres to gamma shape 4 [redacted]. compared to 1 & 4.

Points have form (quantile$_{i/n}$, $x_{(i)}$.)

If the sample quantiles compare closely to ref dist ⇒ a straight line. [redacted] ⑨.
O.w. doesn't match well. 8.6
- small below ⇒ heavier left tail
- large below ⇒ [redacted] right tail.

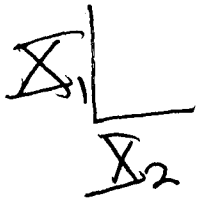
q-q plots are good at finding differences in tails.

- Also q-q plots are indep of location and scale (shape) of the data.
- Not a useful technique for multivariate data. (ECDF, q-q, etc)

(Book also discusses; Smoothing graphing continuous functions, ? Bezier curves)

Scatter plots.

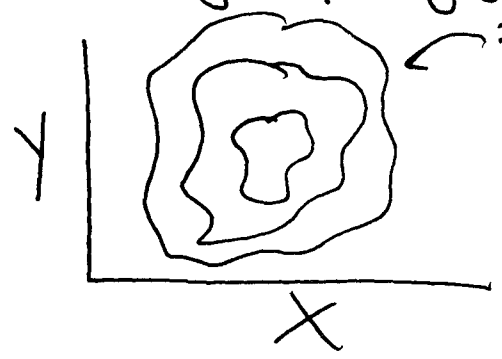
just a plot of points on cartesian axis rep. the variables → useful for showing dist ? is the same dim as the data.



Representations of the 3rd dimension

(1) Countour plots: Allow rep of 3-dim in a 2-dim graphic. A countour line (or band) represents a path over which the values in the dim not represented are constant. Particulary useful if one variable is a dep. var.

Ex/ Data (X, Y, Z)



② Image Plots – useful if one variable is a dependent variable – the third dim is represented by color or by a gray scale.
- useful in identifying structural dependencies
- reordering of the "ind." axes has a major effect esp. w/ categorical data.

Ex/ Fig 8.7 Data rep. gene expression for 500 genes from 60 cells
ⓐ arbitrarily pos, others try to order the cells or genes (or both) to find patterns.

R produces both <u>contour</u> & <u>image</u> plots.

③ Simulation of a visual perception of depth. surface rendering – viewpoint dependent.

④ Stereograms – use two horizontally juxtaposed displays of the same data set. One set (or both) are offset from the other. Idea: the viewer has to defocus each separate view & fuse the two views into a single view.

Ex/ Fig 8.8 (X, Y, Z). stereogram is formed by 2 side by side displays $\leftrightarrow x$ $\updownarrow y$ and z is the depth. The perception of depth occurs because the values of x are offset by an amount prop. to the depth

The depth at point i is $d_i = c(z_{max} - z_i)\frac{x_{max} - x_{min}}{z_{max} - z_{min}}$

C depents on sep of displays & on the units of measurements of the data.

Then $(x-d, y)$ is plotted on the left & $(x+d, y)$ on the right.

Sec 7.2 Viewing Multivariate Data

There are basically two ways of displaying higer dim data on a 2-d surface.

$O(d^2)$ small dataset.

① Use multiple 2-d views proj into cartes.

② use other types of graphical obj w/ char associated w/ each of the variables.

① Projections

ⓐ View two variables at a time using scatter plots. An effective way or arranging these plots is to lay them out in a square pattern w/ all plots in the same row having the same variable on the ↕ axis & all in the same col. having the same variabe on the ↔ axis. Called a scatterplot matrix or SPLOM (in R splom).

Ex/ Fig 8.9

ⓑ IMPLOM Ex/ Fig 8.10

Others in book

ⓒ Move a plane through space & proj points onto it. (Be careful Ball & Sphere are the same).

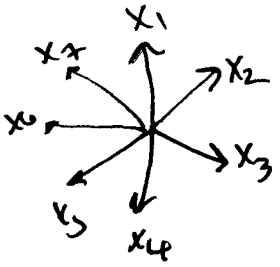
② Non Cartesian Displays

- Each obs. is rep. as a more complicated object than just a point. w/ the values of the ind. variables X_i (x_{i1} x_{i2} x_{i3} ···) represented by some aspect of the object.
- Downside - only useful w/ small datasets
- sometimes hard to see relationships @ 1st.

Ex/ Table 6.1

Obs	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	6	5	4	3	2	1	2
2	1	2	3	4	5	6	5
3	4	4	6	4	3	3	3
4	2	2	2	3	6	6	6
min	1	2	2	3	2	1	2
max	6	5	6	4	6	6	6

③ Glyphs & icons



- Star diagrams: To represent an obs of d-dim data you use rays pointing from a central point in d equally spaced directions to represent the values of the variables.
- variety of ways to rep magnitude
  - true values
  - scaled so that min = 0, max = 1.
  - "snowflakes" w/ connected end points

Ex/ Fig 6.1 "Stars" in S-Plus.

- Chernoff faces: Stylized human faces w/ each variable associated w/ some feature.
  - width of mouth, height of face, etc.
  - since we are so adept at recog faces often we can quickly see similarities

Ex/ Fig [redacted].

"faces" in splus

X_1 - area of the face
X_2 - shape of the face
X_3 - length of the nose
X_4 - location of the mouth
X_5 - curve of the smile
X_6 - width of the mouth
X_7 - location of the eyes.

- others Matlab - feathers, compass, & rose.
- # dos ≤ 20 or 30

(5) Parallel Coordinates: Points become Broken line segments.
- a piecewise linear curve joining the values of the variables on a set of parallel axes, each axis rep. the values of a given variable.

Ex/ Fig [redacted]

generally [redacted] scaled to cover the range of each variable. [redacted] min to max

Parallel Coord. help to identify relationships between variables.

- pairwise pos correlations between variables in adj coord lines result in line seg w/ sim slopes v/s neg correlations w/ diff slopes
- can reorder the data to find more.

- sim obs have similar paths
- useful to identify groups in data

Others: Trigonometric series : Points become curves
: cone plots
rotations

Displaying large Data Sets

- Be careful - too dense • info is missed.
  - overplotting.

Ideas: jittering - slightly offsetting data
grayscale
color, etc.

All depend on visualization cun ability of human eyes & brain to interpret the data.