Statistics 215a – 9/8/03 – D. R. Brillinger

Summaries of location

mean = Σ y / Σ 1, mean()

median = middle of ordered values, median()

midmean = mean of middle half of ordered values

trimmed mean, mean(…, trim=…)

biweight = Σ $(1 - u^2)^2$ y / Σ $(1 - u^2)^2$, triweight, lowess()

u = measure of deviance, e.g. standardized residual

Symmetric case

Behavior under transformation?

Data = summary + deviation

Summaries are useful, but …

Statistics 215a – 9/8/03 – D. R. Brillinger

Spread vs. level plot

Batches of data

Plot

log(IQR) versus log(median)

for each

Then

$$\text{Log(IQR)} \approx a + b \log(\text{median})$$

suggests transformation

$$Y = X^{1-b}$$

Better suited for comparison and visual exploration

Various procedures require constant spread

Perhaps: more symmetric, fewer outliers

Statistics 215a – 9/8/2003 – D. R. Brillinger

Q/Percentile plot aka empirical cdf

Data $x_1, x_2, \ldots, x_n$

Order statistics $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

Plot the points $(x_{(i)}, p_i = i/n)$

$Q(p_i) = x_{(i)}$

May join by straight lines so can interpolate

Might plot at (i-.5)/n, (i-.33)/(n+.33) or i/(n+1)

Can locate quantiles

Local slope gives local density – steeper the tangent the greater the density

Symmetry – do points in top right stretch off the same way as points in bottom left?

Can compare quantiles

Sometimes axes reversed

Statistics 215a – 9/8/2003 – D. R. Brillinger

Empirical Q-Q plots

Data $x_1, x_2, \ldots, x_m$; $y_1, y_2, \ldots, y_n$

Order statistics $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(m)}$; $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

Interpolate the points $(x_{(i)}, p_i = i/m)$

$Q_x(p) = x$; $Q_y(p) = y$

$(Q_x(p), Q_y(p)\ ;\ 0 \leq p \leq 1)$ is the empirical Q-Q plot

Compare distribution shapes:

y = x?
y = a+x?
y = a + bx?

Differing spreads, skewnesses, …

Statistics 215a – 9/8/2003 – D. R. Brillinger

Empirical Q-Q plots

Question – regulation needed to control traffic and associated emissions?

"Sunday and workday variations in photochemical air pollutants in NJ and NY". Cleveland et al. (1974) *Science* 186, 1037-1038.

Data - daily maxima ozone

They found ozone slightly higher on Sundays than weekdays

A surprise since vehicular traffic / pollutants much lower on Sundays

Ozone generation mechanism not understood!

Q-Q plot compares entire distributions, not just means

Base line y = x

Sunday quantiles of NO, NO_2, CO markedly lower

Very highest O_3 maxima occur workdays, rest highest on Sundays

No p-values in paper