STAT 400 Data Analysis Sep 12, 2003

Today: 1. Short quiz (Ch 1 & 2)
2, Power Transformations (ch 2)
3, Aspect Ratio, Smoothing with Loess, … (Ch 3)

1. Quiz: (mark your own!)

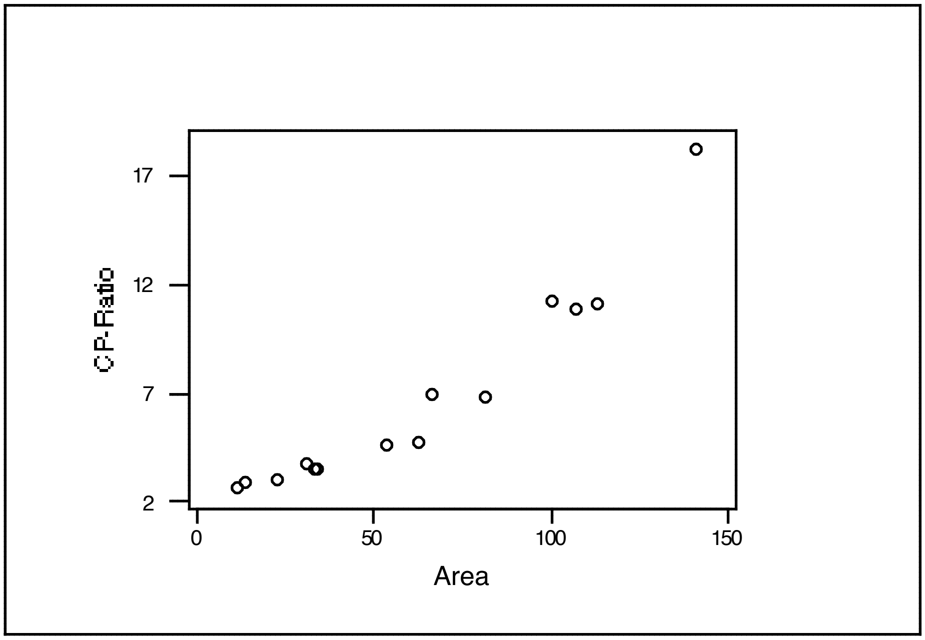
1. When you analyze a data set, what are the first two steps you should do?
2. If you were comparing the distributions of two independent univariate samples of unequal size, what advantage over two histograms are
a) two quantile plots?
b) two dotplots?
3. Is a "normal Q-Q plot" a "Q-Q plot"? Explain your choice.
4. How does Figure 2.17 (p 41) relate to the analysis of variance technique?

2. Power Transformations: Data x_1, x_2, ….,x_n

Let $y_i = x_i^t$. For these transformations with $t<1$, we need $x >0$.
If {x} are skewed right (long tail right) then for $0 \le t < 1$, {y} will be less skewed to the right.
If {x} are skewed right (long tail right) then for $t<0$, it is not clear what skewness y will have – it depends on how close to zero the smallest values of x are.

The 1/x transformation is often useful when the data is recorded in un-natural units.

3. Notes for Ch 3 (pp 86-101): Bivariate Data



CP Ratio is density of ganglion cells at center of retina to density at periphery of retina
Area is the area of the retina and is a proxy for age (older fetuses have bigger retinas).

Asymmetry in role of variables – see objective p 87 para 2. Usual in regression.

Notice that the MINITAB default plot does not have the same aspect ratio as on p 86.

Is there a proper aspect ratio =Height/width of enclosing rectangle? Fig .3.3 shows why aspect ratio is important. To bank to 45 degrees, choose the scaling such that the dot-to-dot slopes are as close to ±1 as possible. (A weighted least squares criterion). See details p 90-91 but do not worry too much if you don't follow it exactly. Note though on p 91 that it should be the **absolute** value of □ that should be close to 1. So wavy curves are candidates for this criterion.

Fig 3.5 shows that a quadratic plot does a good job of fitting the data. How is this done? Example of parametric fitting.

What is non-parametric fitting? Loess is the most used method. Very useful in situations like Fig 3.6. See details of method p 95. Understand this – it is not too complex. The following refers to the "steps" graphs on p 95.

1. The first step is to define a grid of points across the x axis – the first graph just shows what to do for **one** of these grid points. The thin dotted line is at the grid point. The other dotted lines are equi-spaced from the grid point – what determines this distance is the deviation needed to reach the 10th most distant point. (The number 10 is computed from the parameter □ , which in turn is the proportion of data to be included in the interval around each grid point. In this case, we have 20 points, and □ equals 0.50, so the □ times 20 is 10.)

We are going to use these 10 points to estimate the smoothed fit at the grid point itself. But we want to make the data close to the grid point to count more than the more distant points. Step 2 accomplishes this….

2. In this step we define a weighting function that we want to apply to the data in the interval of "close" points found in step 1. Note that the shape needs to be spread over the particular interval defined in 1. and that this will usually vary across the grid points.

The details of this weight function are given on p 101.

3. We use the weight function of 2. and the interval of data in 1. to do a weighted regression. Wherever that regression function (usually a line or a quadratic curve) crosses the grid point line is the value of the smoothed data at that grid point.

If we go through this whole process for each grid point, we have the outline of the smoothed data. Joining line segments makes this look like a curve. Supressing the smooth points leaves a picture of the original raw data and the smoothed curve.

The curve is called the loess curve (really joined line segments) or sometimes lowess.

The parameter Lambda (□) is the degree of the WLS polynomial to be used. Usually □ is 1 or 2.

Alpha is the proportion of the data to use around a given grid point. Usually □ is in the range .1 to .5 but can be anything in (0,1]
Control constants alpha and lambda must both be set to use loess.

See effect of changing alpha Fig 3.9 p 99.

Can we do analyses on nonparametric fits?

Residual Analysis
Parametric Inference
Visualization Inference

What is weighted least squares?
When might it be useful in data analysis?

Next – More about residual analysis

R Tutorial Monday.

MINITAB tutorial?