**BIOS 6621 Statistical consulting I – Conditional distributions, graphs and models**

This is a very informal description of conditional distributions, graphs, and models, not worrying much about formal notation.

**Conditional distributions**

Recall from probability Pr(A|B) is the conditional distribution of event A given event B, and f(y|x) is the conditional density of random variable Y given X=x. One way to think of conditional distributions is as subgroup distributions, where the levels of the conditioning variable represent the subgroups. For example if Y is income and R is racial group, f(y|r) represents the density of income for racial group R=r. If income varies across racial group, these densities are different for different races. Adding a third variable, say age A, the conditional density of income given race and age is f(y|a, r).

**Graphs**

Conditional distributions can be visualized using the graphical tools in Section 2.4 of the notes, or summarized in the paper by Kozak. For example,

* The unconditional density of income, f(y), could be shown by a histogram.
* The conditional density of income given race, f(y|r), could be shown by several histograms, one for each racial group.
* The conditional density of income given age, f(y|a), could be shown by a scatterplot of income versus age.
* The conditional density of income given age and race, f(y|a, r) could be shown by a scatterplot with different symbols for races, or by several scatter plots, one for each race (i.e. panels).

**Models**

Models often summarize conditional distributions, for example with means estimated by regression, linear models or some other method. Model estimates are sometimes superimposed on graphs to highlight these features. For example,

* The unconditional mean of income, E(Y), could be shown as a dot or line on a histogram of f(y).
* The conditional mean of income given race, E(Y|R=r), could be shown as dots or lines on each histogram of f(y|r).
* The conditional mean of income given age, E(Y|A=a), could be shown as a line connecting values of E(Y|A=a), i.e. of the mean of Y for each value of A.
* The conditional mean of income given age and race, E(Y|A=a, R=r), could be shown as lines on a scatterplot of y versus a, with different colored lines and points for each race. Or as lines on the panel plots of y versus a, one panel for each race.