Chapter 1A: Formulating and refining questions (will renumber later)

"*A problem well stated is a problem half solved.*" Charles Kettering

1A.1 The fundamental importance of questions

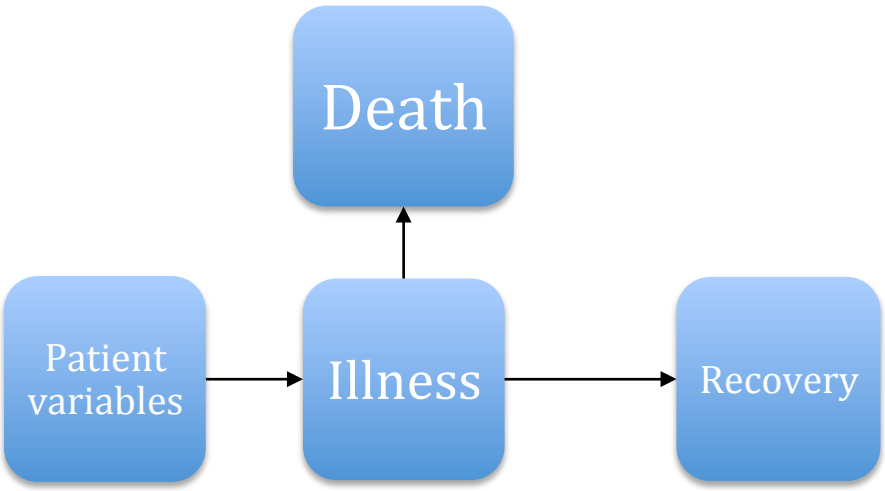
Sometimes investigators will ask statisticians to 'analyze the data', or even more distressing, to 'run the stats'. Sometimes the study and goals are so clear that these are reasonable requests, but more often it is necessary to think about, discuss, and carefully define the questions of interest. Questions drive the analysis, and in the best cases the questions and answers tell a coherent and interesting story.

Some investigators need help organizing the data and posing a set of interesting questions. Other times, investigators do pose good, clear questions, and even in these cases it may be necessary to refine the questions so that they are stated carefully enough to be answered by the available data. Investigators are experts in their subject, but statisticians have typically seen a lot of studies and have a different way of thinking about data, so can suggest questions the investigators may not have thought of. Working together can then be very rewarding.

1A.2 An early step: Grouping variables

Many studies involve a lot of variables, particularly a lot of outcomes. In these cases one of the most useful and earliest steps can be to organize or classify the variables in some way based on the subject matter. An example is classifying pre-operative variables for heart surgery patients into demographics, general subject characteristics, cardiac-related variables, and comorbid conditions. Another example is classifying variables about subject characteristics, behaviors, physiological outcomes, and disease status. There can be several benefits to such classifications: a) It forces the analyst to consider the meaning of each variable, which leads to increased understanding of the subject matter; b) It allows better organization of tables and figures, around subject matter components; c) It suggests more general questions, for example questions about heart failure rather than about a particular pre-operative variable that happens to be related to heart failure; d) It allows us to think about a few sets of variables rather than large number of individual variables, making it easier to suggest interesting and more general questions. While there is not usually one unique way of classifying variables, it is very often worth using some classification.

Sketching a conceptual model can also help with formulating general questions. For example the chart below describes a situation where some patients become ill, and then some recover and some die. It suggests general questions like 'What types of patients become ill?", "What types of illnesses tend to lead to death, and what types tend to lead to recovery?".



1A.3 Formulating interesting questions

Taking the time to think about the subject matter and what 'ordinary people' might be interested in knowing can pay off. Sometimes these simple sounding questions are forgotten in the details and technical work of the analysis. Sometimes more technical questions can be phrased in simpler terms. For example, a technical question like 'What factors are associated with higher risk of heart attack?' might be thought of as 'What types of people tend to have heart attacks?' While you might want to use the first phrasing in a technical paper, it might be easier to think and talk about the second.

1A.4 Refining questions

One very common situation is when questions seem clear, even obvious, but are not formulated carefully enough to allow analysis. Investigators may say "The question is simple, I just want to know if the groups are different", not thinking of the many ways the groups could differ. Do the groups differ in which outcome? At what measurement time? In their patterns of response over time? In their response to an intervention? Each of these questions suggests and requires a different analysis.

In grant proposals, the general questions are often referred to as 'Aims' or 'Goals', and the more specific questions as 'Hypotheses'. In fact, one way to define a hypothesis is that it is specific enough to determine an appropriate statistical analysis. The hypothesis needs to be specific enough to make a direct connection to the results of some analysis.

A good way to determine if questions or hypotheses are specified carefully enough, and to guide investigators in doing so, is to consider hypothetical results. This may be done in the form of mock results tables, with or without

hypothesized values. For example, in a two group study involving an intervention, results can be arranged in a two by two table of means.

	Group A	Group B
Control	a	b
Intervention	c	d

Common questions can be illustrated by pointing to differences involving the four means. For example, within-group intervention effects are a versus c or b versus d, combined group intervention effects are a+c versus b+d, and group differences in intervention effects (interaction) are (c-a) versus (d-b). This becomes very useful in designs with more factors. This example also illustrates one reason why statisticians can help investigators to formulate questions, even when the statistician is not so familiar with the subject matter: In some common data structures, certain questions are standard to consider.

1A.5 Two common situations

Mediation

Sometimes an association between a covariate or exposure and an outcome is hypothesized or found, and it is of interest to examine whether the path from exposure to outcome may be explained by another variable. The other variable would be called a mediator and would be said to mediate the relationship between exposure and outcome. An example would be a drug that reduces the probability of death. If it acts by reducing blood pressure, which in turn reduces chances of death, blood pressure would be a mediator.

Effect modification (interaction)

Interaction means the effect of one variable is different depending on levels of a second variable. Said another way, the effect of the first variable is modified (changed) by levels of the second variable. An example would be an intervention that has a different effect for men and women. Statisticians usually use the term interaction, epidemiologists often use effect modification. These variables may be continuous, categorical, or one of each, the issues are the same. This is the correct way to answer questions about differences in the intervention effect betweeh groups, rather than fitting separate models in the two groups and comparing them. The latter can be useful in understanding the interaction effect, but the test should be done using an interaction model.

1A.6 Case studies and examples

Exercise 1: Formulate some interesting questions about the following study. For each of 1000 enrollees aged 60-70 at an HMO, their health care costs during the previous year were recorded. Costs for enrollees who did not use any services during the year are listed as 0. Members were classified as retired or working,

and the study would like to study how health care costs differ between retirees and workers. Focus on the distribution of cost, particularly that some members had 0 cost for the year. Consider only retirement status and cost, not other variables such as age or sex.

Exercise 2: Formulate some interesting sets of questions about the following study. About 80 patients with idiopathic fibrosis, a lung disease, were studied at a baseline visit and again one year after baseline (unless they had died during the year – death is recorded). At each visit their lungs were tested with Xray scans, where each Xray was read by a radiologist for several types of lung disease, and also each Xray was passed through an automated computer program that calculated certain features of the Xray (eg ratio of white to black pixels). Each patient was also tested for total lung capacity and forced vital capacity using physical tests. Consider only variables that are specifically mentioned in this description.

Exercise 3: Sketch a conceptual model describing the questions yellow marked on p32-33 in the Wyatt 2008 paper.

Exercise 4: A cardiologist would like to study the association between race and adverse outcomes. Explain and give an example of what it means for a) the level of care the patient receives is a mediator for the relationship between race and adverse outcomes, and b) the sex of the patient is an effect modifier of the relationship between race and adverse outcomes.