Homework 3: Estimation

1. In an age of open APIs, data sharing between services is a common exercise. Users of these services have expectations, however, about how their data should be distributed. Respecting the privacy of users can be difficult, especially when shared data is combined with secondary sources of information. One approach to this problem recommends only releasing or sharing "safe" data and let it flow in an unrestricted fashion.

   Making data safe is an area of considerable research (although it's not often worded exactly in that way). For example, let's suppose we run a service that administers online surveys. Users are asked questions on a variety of topics, perhaps politically sensitive topics, and releasing their answers publicly could be embarassing. Here is a simple scheme to make sharing yes/no answers a bit more "safe."

   (a) Have the respondent toss a coin with probability of showing heads $\alpha \neq 0.5$. You choose the coin and so you set the value of α; in short, it is a known quantity.

   (b) The respondent is to keep the results of this coin toss private; have them

      i. Answer the question posed if the coin was heads, or

      ii. Answer its complement if the coin landed tails

   These data are now "safe" in the sense that we are not sure whether a respondent's answer refers to a question or its complement. But what price have we paid for this safety? Do we still have the ability to estimate the chance that people will respond afirmatively to the original survey question?

   To study this, suppose we have n respondents follow this recipe and let X denote the number of "yes" responses you collect in this way. Let θ represent the probability that someone will answer "yes" to the question we're interested in and let π denote the probability that they respond "yes" using our two-stage process.

   (a) For one respondent, write down an expression for the chance that they answer "yes"; this should involve θ and α.

   (b) Viewing X as the result of n trials, each with this probability, come up with an estimate of π.

   (c) Use this estimate and the expression in (a) to form an estimate of θ. Keep in mind that α is something you fix and is known for this analysis.

(d) Is this estimate unbiased? What is its standard error? Address these questions first through simulation. Then, using the expression for the estimate, reason directly and compute mathematically the bias and standard error of your estimate.

(e) How does this compare to asking the question directly? Is one more efficient than the other?

2. Let's change the procedure slightly; now instead of answering the complement of the question when the coin comes up tails, we instead have respondents answer a question for which we know the probability of a "yes" in the general population is δ.

   (a) Follow the same general procedure as in the previous question and derive an estimate for θ.

   (b) Finally, compare the mean squared errors of these two estimators; which do you prefer?