



How Deep is the Ocean? (A song - cf Wikipedia)

## 1 What percentage of the world's surface is covered by water?

The data provided by the Scripps Institution of Oceanography [see [Oceanography Data](#) in the 'Resources' link opposite the BIOS601 topic "Sampling"] can provide an answer, but some work is required on your part.

- i. Draw a simple random sample<sup>1</sup> of 200 locations on the Earth's surface, and obtain from the SRTM30\_PLUS database the land elevation / ocean depth at each of these. From these 'readings', calculate a point estimate of the percentage. Also calculate a (probabilistic) margin of error (ME): do this by calculating a standard error, and multiplying it by say 1.96 so that you can make a probabilistic statement.
- ii. Are you worried about the appropriateness of using 1.96 (and the Normal distribution) for 95% confidence? Why/why not?
- iii. The root mean squared error includes both sampling variation and non-sampling errors. Your margin of error is limited to the sampling variation, and is modulated by the choice of 'n.' It does not include *non-sampling* errors. Describe one possible source of non-sampling error in this particular context [internet searching encouraged! – and try to find an unrelated example that you could describe to a lay person, and remember the concept by. If you find a striking example, share it with us!].

## 2 What is the average depth of the ocean?

- i. From the relevant observations (from among the 200), estimate the mean ocean depth, and calculate an accompanying ME. Even though there is a

<sup>1</sup>Previous year students have used the R `geosphere` package. Instead, 'roll your own' function; if need be, read the '[random points on a sphere](#)' notes in the 2 R functions by JH, found in the [Oceanography Data](#) link

random component to it, pretend that the sample size was predetermined.

- ii. Are you worried about the appropriateness of using 1.96 (and the Normal distribution) for 95% confidence? Why/why not?

## 3 Ensuring that a sample of $n'$ locations will yield $n = 200$ [or more] usable ones

- i. How big must  $n'$  be in order to have a good chance (say 80%) that it will yield at least 200 usable ones (i.e. ocean locations)?
- ii. What if you sampled sequentially until, at the  $n'$ -th draw, you reached the 200-th usable one? What distribution describes the random variable  $n'$ ? How could you calculate its 10-th and 90-th percentiles? (pretend you know the value of the parameter that determines its distribution).

## 4 More efficient (or more practical) sampling strategies

(*Very briefly*) describe the circumstances<sup>2</sup> in which a sampling scheme other than s.r.s (systematic, stratified, cluster) would offer either practical or statistical efficiency advantages; mention also the downsides of these schemes [text-book and internet searching encouraged – *if* you acknowledge the source!].

## 5 Oh Oh

(a) A researcher spent the entire research budget on a sample of 200 locations, but where the latitude locations were  $\sim U(-90, 90)$  and likewise the (independently selected) longitude locations were  $\sim U(-180, 180)$ . Are the data worthless? Could you recover something from them?

(b) At 'latitude'  $\theta \in [-\pi/2, \pi/2]$  on a (long <sup>$n'$</sup> -based) section of a sphere (e.g., an orange), the width  $w$  (and thus the no. of sampled locations should)  $\propto \cos(\theta)$ . Use this to justify the statistical algorithm used in the R function.

<sup>2</sup>The Cross-Canada Survey of Radon Concentrations in Homes [Resources] might help.