

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 sample¹ of 200 locations on the Earth's surface, and obtain from the SRTM30_PLUS database the land elevation or 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 random component to it, pretend that the sample size was predetermined.

¹Previous 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

- 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 circumstances² 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 [textbook 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 (\log^{n_l} -based) section of a sphere (e.g., an orange), the width w (and thus no. of sampled locations should be) $\propto \cos(\theta)$. Using 'Figure 4B' (Oceanography Resources – it was dropped from the final version of the *Significance* article), or the R plot with the **# longitude lines laid end to end**, explain the inverse CDF method in words. [related idea: in a distribution, are there more/fewer people between the 55th & 56th percentile than there are between the 5th & 6th, or 95th & 96th?]

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

6 Physical Activity: JH 2010-2015



2012						
Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
9593 3844	5157 2494	3202 3202 Battery	2605 2605	4856 6445	7487 363	6779 956
6640 571	10052 4484	7347 7899	7146 299	7529 649	4628 729	6750 1500 Battery
8230 1576	6759 5057	12926 8804	15278 15393	10538 11645	11255 11447	15954 8348
9595 6971	12432 10475	7421 5567	14000 12649	8280 7586	7274 2791	18318 15690
8760 5819	7044 10864	9337 6535				

Since 2010, JH has used a ‘step-counter’ (pictured above left) to record how many steps he takes each day. His spouse AM has done the same, and has entered the pairs of daily counts onto a log book.

Refer to the three files (2010-2011, 2012-2013 and 2014-2015) under the heading “Physical Activity: How many steps a day has JH being doing since 2010?” near the top of the Resources webpage. The 2010-2011 .csv file has the paired recordings for 2010, as well as JH’s ones for 2011. The 2012-2013.pdf and 2014-2015.pdf files have scanned images (see above right) of the pages of paired recordings from the log-book.

(a) Who had more daily recorded steps in 2010? and by how much? (report the 2 means, as well as the higher:lower ratio, e.g., 1.27:1).

(b) Ignore the fact that it is a census of 2010 (i.e. a 100% sample – so the finite population correction factor would make the standard errors zero. Calculate a standard error for the mean difference, and (if you are able to: if you are not, ask JH) an approximate one for the ratio.

(c) Describe some possible ‘errors’ that are not included in each standard error.

(d) Look up, and provide a verbal description of Benford’s Law, and how the

first person to notice it was led to it. (Before testing it out) do you think it should apply to recorded step counts? Why/why not? Then test it out on the computerized step count data for 2010-2011.

(e) In order to assess any trends in JH’s activity, it would be nice to have the mean daily steps for each of the 6 years. Those for the first 2 years are easy to obtain, but to obtain exact values for 2012 to 2015 would take more data entry work than is reasonable for a single assignment.³

* Suggest two sampling methods (the simple random sampling method, and one other sampling method), each of which samples approximately 30 days per year, to obtain estimates of the mean daily number of JH steps for each of 2012 to 2015.

* Carry out ONE of these methods, preferably using R to select the days (report the starting seeds used, so that JH can replicate the sampling plans – also try to co-ordinate with other students so that you don’t all draw the same sample of days). Make a time-graph of the values/estimates for the years 2010-2015. Accompany each estimate with an error bar, taking care to say what the error bar represents.

* Mention any ‘savings’ in time/effort that you thought of as you did the sampling, and the extraction and computerizing of the sample values.⁴

7 What was the point of each of the assignments?

For each of the assigned questions, use one sentence to describe what you think the learning objective was; use another to describe in what situations the concepts and techniques will be of use to you and to those you will work with.

³Unfortunately, the OCR function in Adobe Acrobat cannot reliably read AM’s handwriting – even if it does a reasonable job at printed material, such as the (automated) Blood Pressure and Pulse (Heart Rate) Measurements discarded by customers of the Jean Coutu pharmacy, shown further down the Resources webpage.

⁴JH has already started to test computer dictation as a way to enter the numbers of steps, and hopes to find an efficient way that he and the class can divide the labour and computerize the numbers for all days of each of the years 2012 to 2015.