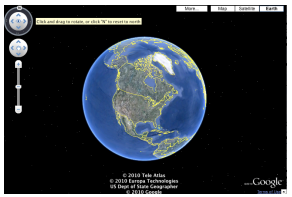
How Deep is the Ocean?

(A song – see Wikipedia – and an article – see Significance Dec, 2014)



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

The data provided by the Scripps Institution of Oceanography [accessible via `http://www.biostat.mcgill.ca/hanley/bios601/Surveys/Oceanography/` can provide an answer, but some work is required on your part.

i. In previous years, students drew a simple random sample of 200 locations on the Earth's surface,[1] and obtained from the SRTM30_PLUS database the land elevation or ocean depth at each of these. This year, to save some time[2], both the drawing of the sample, and the 200 database lookups have already been done for you – there are several `.csv` files (62 if JH has counted correctly) from previous years at the bottom of the 'Oceanography Data' webpage. To avoid picking the same datafile as another student, use your birthday `yyyymmdd` as the eight digit argument in the `set.seed()` function in `R`, and sample 1 from the numbers 1 to 62 using the `sample(1:62,1)` call. From the 'readings' in your selected file, calculate a point estimate of the percentage.[3] 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. [*Again, use a sensible no. of decimal places*]

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

iii. If you are – and even if you are not – find an online calculator/table that yields an 'exact' confidence interval. Compare the 'exact' interval with the 'approximate' one above.

iv. Using what you are able to find online or from your textbooks, explain to a relative who is an engineer how exactly this 'exact' confidence interval is calculated and how the principles behind it differ from those behind the usual one. [We will come back to this in a later class].

v. 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.[4] Describe one possible source of non-sampling error in this particular context of ocean depths.

Also, describe an unrelated example you would use to describe non-sampling errors to a lay person. [internet searching is encouraged, but please cite the source if you found this example online, or in a textbook]

2 What is the average depth of the ocean?

i. From the relevant observations (from among your 200), estimate the mean ocean depth, and calculate an accompanying ME. [5] Even though there is a 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

[1] Ch 9 in Gelman & Nolan's Teaching Statistics: A Bag of Tricks (an ebook at McGill Library) has an interesting way of sampling, and other useful remarks on this problem. Today, one could simply zoom all the way out in Google Maps, and spin!

[2] You are still welcome to get your own 'from scratch.'

[3] Do not show off how many decimals you (`R`) can calculate. If your parents asked you what percentage you got, how many digits would you give them? Same applies to other Qs!

[4] Some define a 'non-sampling' error as one that is not minimized by taking bigger and bigger n; indeed, if there is some 'systematic error in the measurements, taking an even bigger n will just make the answer more precisely wrong!

[5] In doing so, make sure to reduce your 'n' accordingly. Some students in previous years continued to use an n of 200!

n' ? Calculate its 10-th and 90-th percentiles (pretend you *know* the value of the parameter that determines its distribution).[6]

4 More efficient (or more practical) sampling strategies

(*Very briefly*) describe the circumstances[7] 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) One way to obtain random (λ=longitude, ϕ=latitude) locations is as $\lambda \sim U(-\pi, \pi)$, and $\phi \sim pdf(\phi) = (1/2)\cos(\phi)$ on $(-\pi/2, \pi/2)$.

Figure 4B (p. 10) was considered too technical for the Significance article. It is included on p.10 below to help explain how one could sample the ϕ's. Think of a longitudinal-based section of a (perfectly spherical!) orange, and of how wide it is at 'latitude' ϕ compared with how wide it is at 'latitude' 0 (on the equator). It is the same as the relationship between the west-east distance between two locations at the same latitude (e.g., $\phi = 45.5$N), but say 1 degree longitude apart, and the (approx. 100km) west-east distance between two locations on the equator, also 1 degree longitude apart. If we take the distance at the equator to be 1, the the distance at 'latitude' ϕ is $\cos(\phi)$. Thus, in any chosen longitude-based section the number of sampled locations at latitude ϕ should be $\cos(\phi)$ times the number sampled at the equator.

- Use your dataset of 200 to check that the random locations produced by this method [implemented in the `R` code used to create the personalized datasets for question 1(i)] appear to be sensible.

[*Question 10, The Locations of the Stars, takes up the question of randomness, from a different viewpoint!*]

(b) A researcher spent the entire research budget on a sample of 200 locations, using $\lambda \sim U(-\pi, \pi)$, but $\phi \sim U(-\pi/2, \pi/2)$.

- Explain why this sampling scheme is flawed. [Gelman and Nolan have a few words on this]. Are the resulting data worthless? Or, do you think we could recover something from them?

- Using the information in (a), suggest a way to correct for the researcher's oversampling of locations further from the equator.

(c) Search online (or in your textbooks) for ways to draw random samples from a non-uniform continuous distribution. List ones that are easy to implement when only the (i) pdf, (ii) CDF has a closed form.

(d) The rationale behind the 'inverse CDF' method is often missed – and not easily recalled years later – if students go through the 'proof' as a mere 'math-stat' or calculus exercise.

Figure 4B (p.10) tries to explain the 'inverse CDF' method in pictures rather than via calculus.[8]

Pages 11-12 are notes from 2010, with yet another plot of *west-east lines laid end to end* – another attempt to 'explain' it in this same 'sampling latitudes' context.

The attempt on page 13 uses an unnamed continuous random variable, but starts with a simple discrete version that might make the methods more intuitive.

• Now the test of whether any of these three attempts succeeded: ***in your own words***, explain to that same relative of yours how exactly the inverse CDF methods works. If you don't like the examples/explanations JH has provided, feel free to make up your own.[9]

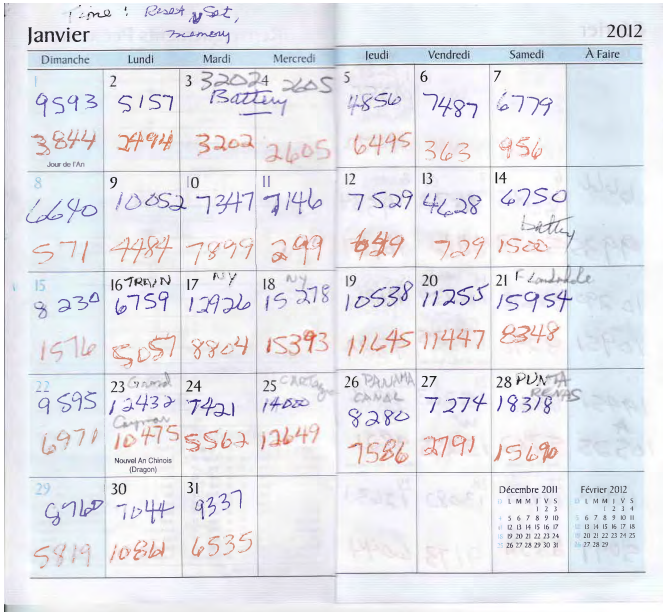
[6] `R` has 'exact' `d- p-` and `q-` functions for this distribution, but – given the numbers involved – the calculations can also be reasonably approximated if you know just its mean and variance.

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

[8] This article `http://www.biostat.mcgill.ca/hanley/Reprints/HowDeepIsTheOcean.pdf` originally had the data-mining challenge, and a description of the method to generate random locations. But the diagram (now on page 7 below) was considered too complex and too technical for the Significance Magazine readership.

[9] In the past, JH has heard a teacher start by asking students whether in a distribution – any distribution – there are more/fewer people between the 55th and 56th percentile than there are between the 5th and 6th, or 95th and 96th? This teacher was also quite fussy about words, and about using the word 'percentile' correctly; so he would probably have taken exception to JH's saying *percentiles* are numbered 1-100

6 Physical Activity: JH 2010-2021



Since 2010, until it stopped working in Fall 2021, JH has used a 'step-counter' (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 files (2010-2011, 2012-2013 , 2014-2015, 2016-2017, 2018, 2019-2020, ...) 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, 2014-2015, 2016-2017, 2018 and 2019-2021 pdf files have scanned images (see above right) of the pages of paired recordings from the log-book.

i. 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).

ii. 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.

iii. Describe some possible 'errors' that are not included in each standard error.

iv. 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.

v. In order to assess any trends in JH's activity, it would be nice to have the mean daily steps for each of the 10 full years. Those for the first 2 years are easy to obtain, but to computerize all of the values for 2012 to 2019 would take more work than this level of precision is worth.[10]

Thus, 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 2010 to 2019 (and from these, an estimate of the trend over the 10 years).

(modified in 2020) *Each of you is asked to focus on the calendar assigned to you, and to 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). The calendar year assigned to you is 201x, where x is the last digit of you McGill ID number – unless your ID number ends ends in 0, in which case you are asked to do 2017, or, if it ends in 1, you are asked to do 2018. Accompany your estimate with an error bar, taking care to say what the error bar represents.*

JH will use these to share a time-graph of the estimates for each of the years 2010-2020.

vi. Mention any ways to 'save' in time/effort that you thought of as you did the sampling, and the extraction and computerizing of the sample values.[11]

vii. ***(added in 2020)*** In order to assess how much JH's activity changed since COVID-19,

(a) compare the mean steps per day for the period March 14 to June 2020 with the mean steps per day for the period January 1 to

[10] Unfortunately, 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.

[11] JH 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 future classes can divide the labour and computerize the numbers for all days of each of the years 2012 to 2020.

March 13, 2020. Be careful how you report the difference, and avoid using causal language. Instead, better to say 'the activity was .. higher /lower when'.

(b) Also fit a 'regression discontinuity' (or 'interrupted time series') model, and (again carefully) interpret the coefficients of the model.

(c) Compare the difference (and the relevant regression parameter estimate(s)) with the corresponding ones for each of the 3 previous years (2017, 2018 and 2019).

Notes:

See this link for a wider view. [If you have data on your own activity, please use them rather than JH's)

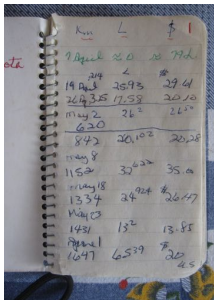
The daily step counts for the 1st half of each of the years 2017-2020 can be found in this `R` file, `http://www.biostat.mcgill.ca/hanley/bios601/Surveys/JHsteps1sthalf2020201920182017.R.txt`.

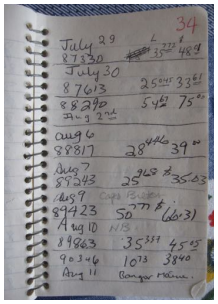
You will notice that to save on data entry (why he did by using the dictation facility in Mac OS), JH entered just the number of hundreds of steps, rounded down, so that for example 4,523 became 45, and 3,678 became 36. How much noise do you think this added? What if he had only entered the number of thousands of steps, rounded down, so that for example 4,523 became 4, and 3,678 became 3?

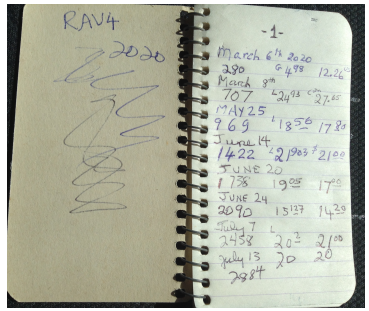
For (b) see Figure 3 in section IV of this classic and still-very-relevant textbook by Campbell and Stanley. Back in 2007, when he began to teach bios601, JH compiled this list of resources for how to carry out experimental, non-exp.tl, and 'quasi-exp.tl comparative studies. He still ranks the *Planning of Experiments* by David Cox as one of the best textbooks he studied as a student (back in 1970 when preparing for his (oral) comprehensive exam). Campbell and Stanley has since been added to his favourites on design, and Bradford Hill's *Short Textbook of Medical Statistics* to his list of favourite medical statistics books.

For a neat and well done example of the regression discontinuity approach, see this report on the *Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study*, part of the first author's doctoral work. For a possible abuse of the regression discontinuity approach, see "No, I don't believe that claim based on regression discontinuity analysis that... " by statistics textbook author and blogger, Andrew Gelman.

7 Gasoline consumption of family minivan 2006-2020, and replacement 2020-







Under the heading "Automobile Fuel Purchases: Toyota Sienna April 2006 - March 2020" near the top of this webpage you will find the documentation of the gasoline purchases for a Toyota Sienna (cf. extracts above), contained in a pdf file of 83 pages. Also on the website is the log showing the purchases for the Toyota Rav4, March 2020 to July 2022. *For didactic purposes only, the bios601 exercises from previous years are shown in blue, but are not part of the 2020 warmup.* The ones for 2020 and 2022, in black, are found further on.

Exercises from previous bios601 courses

In order to analyze the fuel consumption over the 14 years, JH had already entered the purchase dates, the odometer readings, and the purchased amounts (litres, or gallons) into a .csv file – but only for the first 27 pages and the last page. See the .csv file. He did not – and you need not – enter the dollar amounts, but they will be helpful in determining whether each purchased amount was in litres or gallons.

He has left the task of extracting and computerizing the remaining 40 pages (p28 - p.67) as an sampling exercise for the students in this class. The amalgamated dataset, consisting of approximately 550 individual entries from all

of the pages, will be used during the course to illustrate several concepts and principles, including sampling designs, data extraction and management, quality control, and behaviour of statistical estimators.

Each student is asked to use the `sample` function in `R` (with his/her McGill ID number – the 9-digit one that starts with 260 – as a seed) to select which 5 pages (s)he will computerize.

```
set.seed(your mcgill.id)
sort( sample(28:67,5) )
```

Using the .csv file put together by JH as a template, fill in the data from the 5 pages the random sample function allocated to you (again, do not enter the dollar amounts).

To avoid errors, or extractor-to-extractor variations that could affect the amalgamation of the individual .csv files, please follow these guidelines

- Insert your ID number as your 'data-extractor' identifier

- Enter dates as 3 separate columns, with year as an integer, month spelled out in full starting with a uppercase letter, and day as an integer. You may need to look back/ahead to other pages to see which year is involved.

- Be very careful to determine whether the purchase as in litres or gallons – in some cases it is noted, but you can also use the fact that litres were approx. 1 Canadian dollar per litre (L), and gallons (G) were approx. 2-3 US dollars per gallon. Enter them as an L and a zero, or a zero and a G, and we can convert the G's to litres later, at the analysis stage.

The **first quantity of interest** is the total number of litres purchased, so that we can calculate the standard measure of fuel economy – litres per 100 Km – by dividing the estimated total L by the total of 171,232 Km.

For the first 27 pages already compiled by JH, this estimate comes out to 7321.6L /(72,625/100) or 10.1L/100Km.

Once you have computerized your 5 pages of data – and converted each purchase to litres[12] – ***you are asked to make two new estimates***. Each one is a combination of the known amount for pages 1-27, your estimated amount for the 40 pages sampled from, and the known amount for page 68.

The first uses the *page* as the sampling unit, and blows up the mean, $\bar{l}$, of your 5 page-specific totals $l_1, \ldots, l_5$ by a factor of 40:

$$\frac{7321.6 + 40 \times \bar{l}_5 + 89.9}{171,232/100},$$

where $\bar{l}_5$ is the mean of these 5 page-specific totals.

The second uses the *individual purchase* as the sampling unit. Since – if JH's counting is to be trusted – the 40 pages contain a total of 322 purchases, it is

$$\frac{7321.6 + 322 \times \bar{l}_n + 89.9}{171,232/100}$$

where $\bar{l}_n$ is the mean of the n individual purchases $l_1, \ldots, l_n$ in the pages you sampled. For now we will ignore the fact that n is a random variable, since there is a slight variation from page to page in the number of entries per page.

Once you have computed these, you are asked to accompany each one by a 'rough' [13] ***standard error.***

For the first, which uses the *page* as the sampling unit, estimate $\text{Var}[\bar{l}_5]$ as $(1/5)$ × the s^2 of the 5 l's × the finite population correction, (40-5)/(40-1).[14]

For the second, which uses the *individual purchase* as the sampling unit, estimate $\text{Var}[\bar{l}_n]$ as $(1/n)$ × the s^2 of the n individual l's × the finite population correction, (322-n)/(322-1).

Once you have computed these, you are asked to form a 50% confidence interval[15] ***to accompany each of the two point estimates.*** Justify your choices of sampling distributions, and the resulting multipliers applied to the two SEs.

If it were left to you to sample from the purchases that have not yet been computerized, how would you have gone about it?

Finally, use you 9 digit McGill ID number as the name of your .csv file and email a copy to JH.

Exercises for 2020

i. Compare the fuel consumption of the Rav4 and Sienna over their first 2900 Km or so.

ii. You will no doubt realize that an average taken over a larger number of Km would be less noisy, since it would be less affected by the uncertainties about how many Km's were on the odometer and how full the tank was when the car was was bought, and how close to full it was after the last purchase. To address the uncertainty in your answer to (i), provide a plan for a 'sensitivity analysis.'

[12] litres = LitresPurchased + 3.78541 × GallonsPurchased

[13] An expert in sampling might make a somewhat more refined estimate of the L/100Km, and a more refined standard error calculation

[14] Textbooks vary as for the expression to be used.

[15] No, the 50% is not a typo.

iii. Suppose you compute a slope/gradient $\Delta A/\Delta Km$ by dividing the difference ΔA between independent altitude readings A_2 and A_1 by the distance ΔKm between the 2 locations at which they were measured. Suppose ΔKm is measured without error, but that A_2 and A_1 have independent measurement errors centered on 0, with variance σ_A^2. How does the SD of the slope depend on the magnitude of ΔKm?

iv. Summarize the messages from page 4 of these (old) notes

v. After reading pages 5-7 of these notes, rewrite the usual formula for the SD of a fitted slope in a simple linear regression (top of 1st col. of page 7) so that it isolates the 3 factors involved.

vi. Relate one of these factors back to question (ii) above.

Exercise for 2022

i. Now that there are an additional 24 months of data, the uncertainties mentioned in part (ii) in 2020 regarding the starting (and ending) values have less impact.

Assume each purchase fills the tank (generally true). Suggest a way to avoid some of these uncertainties.

8 An interesting (and disturbing) example from Nicholas Horton

At the Montreal SSC meeting in June 2018, Horton shared this item from his 2015 article: *Challenges and opportunities for statistics and statistical education: looking back, looking forward.* A full 2015 manuscript is available here: `https://arxiv.org/pdf/1503.02188.pdf` but for this warmup, it suffices to quote a small section of it.

> Consider an example from the excellent probability and mathematical statistics text by John Rice (2006). I've repeatedly adopted this book, plan to do so in the future, and continue to highly recommend it. But one exercise is highly illustrative of the challenges and opportunities of what and how we teach.
>
> (Problem 3.11) Let A, B, and C be independent random variables each distributed uniform in the interval [0,1]. Question: What is the probability that the roots of the quadratic equation given by $Ax^2 + Bx + C = 0$ are real?

Exercise for bios601:

i. Before you read Horton's paper or blog [16], [or look up solutions of the Web], describe the way you would calculate / arrive at this probability. [It is a nice test of you math-stat training and other talents]

ii. After having read Horton's paper, and numerically compared his two solutions, search the Web for any other solution(s), and provide links to those you find.

iii. What message do you take away from this story?

What message would you like to pass on to teachers of math-stat in 2019?

9 How well can you generate a random sequence 'out of your head', and can someone tell if that is what you did?

The first assignment in the probability course in Berkeley used to be to toss a fair coin 200 times, record on a sheet of paper the sequence of heads and tails, and hand it in. [McGill's David Wolfson also assigned this task] Many students took a 'shortcut' and made up the sequence 'out of their head.'

- Out of your head, 'make up' [17] a random sequence of the results (Head=1, Tail=0) of 200 tosses. enter this sequence of two hundred 0's and 1's into an `R` vector named '`sequence`'.[18] and save it as an `R` object: for example, JH would type `save(sequence,file="sequenceJH.Rdata")`.

  Email the `.RData` file to JH (name your filename `sequenceYourName.Rdata` so that JH can process all the student sequences in the same way).

- Think of a way the teacher might judge whether the student 'made up' the sequence, or actually took the time and tossed the coin 200 times.

[16] Horton also blogged about this in 2011: `https://www.r-bloggers.com/example-8-36-quadratic-equation-with-real-roots/`

[17] Making it up is, of course, much faster than actually tossing the coin 200 times, and recording the sequence.

[18] `sequence=c( , , )`

10 The locations of bright stars

Simon Newcomb was a Canadian-born astronomer, applied mathematician and autodidactic polymath who was Professor of Mathematics in the United States Navy and at Johns Hopkins University. In 1860 he derived, as a limiting case of the binomial, the $e^{-\mu}\mu^y/y!$ i.e., 'Poisson' distribution 'from scratch' and used it to determine the "probability that, if the stars were scattered at random over the heavens, any small space selected at random would contain s stars." Arguments at the time focused on the six brightest stars in the Pleiades. He considered a space of 1 square degree and supposed that there were approximately 1500 stars of the fifth and higher magnitudes spread at random over the entire 41,253 square degrees of heavens.

i. Explain how he arrived at the number 41,253. [If you need to, you can look it up online, a resource his readers did not have. If you do, provide the `url`]

ii. Explain carefully how he arrived at the probability of 0.000000128 that, if the heavens were divided at random into square degrees, *some one of those square degrees* would contain six stars.

iii. How might you compute the "probability that six stars should be found so near together that a square degree could be fitted on so as to include them" ?

iv. In his section 21, Newcomb explains what he means by 'distributed at random.' Do you like his way of 'speak[ing] with more philosophical accuracy' ? Can you improve on it?

v. At the very end of (his final) section 21, on page 140, Newcomb gives us an idea for a physical (rather than a computer) simulation of 'distributed at random,' as well as a possible interpretation of finding an unusual cluster. Suggest a modern-day 'experimental illustration' along the same lines.

Note: A related suggestion, but with an even mix of black and white, arose when statisticians were consulted about modifications for the Vietnam Draft Lotteries. At the top of the second column of page 193 of this article, possible fixes for the Vietnam Draft Lotteries we read:

> John Tukey recommended testing the method of stirring the objects to be drawn so that the mixing procedure worked as advertised; to perform this test, he proposed using black and white capsules in a test draw so that one could visually examine the mixing process and determine its adequacy.

11 Weibull

Refer to the 1951 article by Waloddi Weibull, available here.

i. In biostatistics, Weibull's distribution is now nearly always though about in the context of ***time***-measurements i.e. his 'X' (or 'Y' in bios601) refers to a ***duration of time***. Which of his several the examples involve(s) a ***time-duration*** measurement?

ii. Consider one of these, and give a reason/circumstance why/where the observation could be right-censored.

iii. Consider one of the others, and give a reason/circumstance why/where the observation could be (a) right- (b) interval- or (c) left-censored.

iv. The following questions are taken from the PhD comprehensive examination in May 2021.

(a) Weibull used the following reasoning to derive the distribution now called after him. Consider the distribution function F of a positive random variable X. "Any such F may be written in the form $F(x) = 1 - e^{-\phi(x)}$. Now we have to specify the function $\phi(x)$."

"The only necessary general conditions this function has to satisfy is that it be ... " (fill in the rest of the sentence)

(b) In the context of 'survival analysis', what is the name given to the function $\phi(x)$?

(c) "The most simple function satisfying this condition is $\phi(x) = (\lambda x)^k$, where $\lambda > 0$ and $k > 0$." In the context of survival analysis, what is the form of the hazard function?

(d) In this same context, explain to a non-statistician what the parameters λ and k mean.

(e) If 'X' refers to the survival time of persons who test positive for COVID-19, why would this model be (in)appropriate?
[For those who are interested, see Fig1c for the survival patterns in a very large dataset from the UK]

12 Galton, on meteorology

Refer to Francis galton's article "On an error in the usual method of obtaining meteorological statistics of the ocean" published in a Report of the British

Association for the Advancement of Science 36 : 16-7, in 1866, and available here. (If interested, further papers of his on meteorology can be found here.)

Today, the kind of sampling that produces errors of this kind, namely, where the distribution of a positive random variable Y is estimated using sampling probabilities that should be independent of, but unfortunately involve, the values of Y, is referred to as "??????-biased" sampling.

i. Find out what the missing word is, and when and why the bias first got to be called this.

ii. Paraphrase Galton's message.

iii. Use a simple numerical example involving a meteorological quantity to illustrate his point.

iv. Use a simple numerical example involving word lengths or durations of hospital stays (or family sizes) to illustrate his point. Here are some simulated data on durations of hospital stay, and some real data on word lengths in five books (the frequencies in the five named columns refer to these five books.)

v. Using your example, suggest a way to correct the bias.

When we meet in class, remind JH to tell the story of how, in the 1970s, researchers at the Montreal General Hospital fell into this sampling 'trap' when studying patients who sought mental health care. They applied a 'quick and easy' sampling method to their sampling frame, which consisted of a tray/drawer of '3 × 5 cards', similar to the 3 × 5 index cards used in libraries at the time. The sad part is that – once they realized their blunder – they were not sophisticated enough to correct for the bias, so they simply discarded the data. But, at least, one of the investigators, who held an appointment in our department, was good enough to admit this years later, and to tell the story to JH, in the hope that telling it will spread the lesson to others.

vi. Does this topic have any connection with issues raised in the Oh Oh question (*déjà vu*)?

vii. Use the internet to come up with a few other examples that would be easy to explain to a non-statistician, and that would also be easy to remember and use if you were teaching this topic.

13 'Overlapping' observations

JH has rrecently analyzed data from a study that is investigating how much a university course for future health professionals changed their attitude to patients. Because of the large number of students (180) in the cohort, and the need teach it to smaller groups, the 2-month course was given to a random 1/3 in January-February, to another 1/3 in March-April, and to the final 1/3 in May-June.

The 55 students who volunteered to participate in the study were surveyed 4 times: at the end of December 2020 (t_0) and at the end of February (t_1), April (t_2) and Jun 2021 (t_3). The surveys measured their attitude (and their stress level). For the purposes here, refer to the reported attitude as A.

The primary interest was in whether the (mean) difference of $D = A_t - A_0$ was greater in those who had just taken the course than in those who had not yet taken it.

At the end of February 2021 the investigators were in a position to compare (via an independent samples t-test) the mean D in the 21 participants who had just taken the course (index category) and the 34 who had not yet done so (reference category).

With the new wave of data obtained at the end of April the investigators were in a position to compare (again via an independent samples t-test) the mean D in the 18 participants who had just taken the course and the 13 who had not yet done so (3 of the 34 who were in the reference category in the first comparison did not answer the survey at the end of April, thus the total of only 18+13=31 available for this comparison).

BUT, rather than report multiple comparisons, the investigators preferred to 'merge' or 'pool' the observations involved in the 2 comparisons, and so they compared the mean of the (21+18) = 39 D's from those in the index category with the mean of the (34+13) = 47 D's from those in the reference category, via a test statistic, the numerator of which is the observed quantity

$$num = \frac{\sum_1^{39} D_{index}}{39} - \frac{\sum_1^{47} D_{ref}}{47}.$$

HOWEVER, it is no longer a '2-independent samples' comparison, since some 31 participants contribute twice. Thus, the challenge is to work out a SE ($Variance^{1/2}$) for this test statistic. We won't fuss about the fact that we should probably refer the $num/SE[num]$ ratio to some t distribution rather than to the z distribution, or about 'unequal variances' or any of those complications.

To work out the (sampling) variance of *num*, we can begin by numbering the 55 unique participants, making sure to distinguish the 24 participants who only contributed once (21 to the index category and 3 to the reference category) and the 31 who contributed twice (these 31 all contributed to the reference category at the end of February, and then at the end of April, 18 of them contributed to the index and 13 to the reference category). This way, using the unique ID numbers, and using superscripts 1 and 2 to denote 1st and 2nd contributions, we can rewrite the *num* statistic as

$$num = \frac{\sum_1^{21} D^1_{index} + \sum_{25}^{42} D^2_{index}}{39} - \frac{\sum_{22}^{24} D^1_{ref} + \sum_{25}^{55} D^1_{ref} + \sum_{43}^{55} D^2_{ref}}{47}.$$

If we now write each D_{index} as $\Delta_I + \epsilon$, and each D_{ref} as $\Delta_R + \epsilon$, and move the difference in Δ's to the front, and shorten $index$ and ref to I and R to save space, we can rewrite d as

$$num = (\Delta_I - \Delta_R) + \frac{\sum_1^{21} \epsilon^1_I + \sum_{25}^{42} \epsilon^2_I}{39} - \frac{\sum_{22}^{24} \epsilon^1_R + \sum_{25}^{55} \epsilon^1_R + \sum_{43}^{55} \epsilon^2_R}{47}.$$

Exercise:

i. Assuming that $Var(\epsilon) = \sigma^2$ and that $Covar(\epsilon^1, \epsilon^2) = \rho\sigma^2$, work out $Var(num)$. Then compare it with the $Var(num)$ one would have if there were no overlap in the contributors to the two sides of the comparison (if the 86 were distinct students). Then, in plain language, address any reviewers [19] who might object to the statistical 're-cycling' of subjects, and to the fact that it is not an 'independent samples' comparison.

ii. How else might you go about approximating $SE[num]$?

[19] Some reviewers like to show off what they learned in statistics 101 (the only statistics course they may have taken), where students learn to check independence, but are not shown what to do when it does not hold. JH has seen reports where authors, scared of these powerful 'gatekeepers', have removed data from a second family member, since it 'violates' some statistical assumptions. See the stories at the top of the first column of the second page of this article, and in particular the reference 12. The tax-payer paid for all of these data, and all of the siblings deserve to have their data used.

14 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.

****

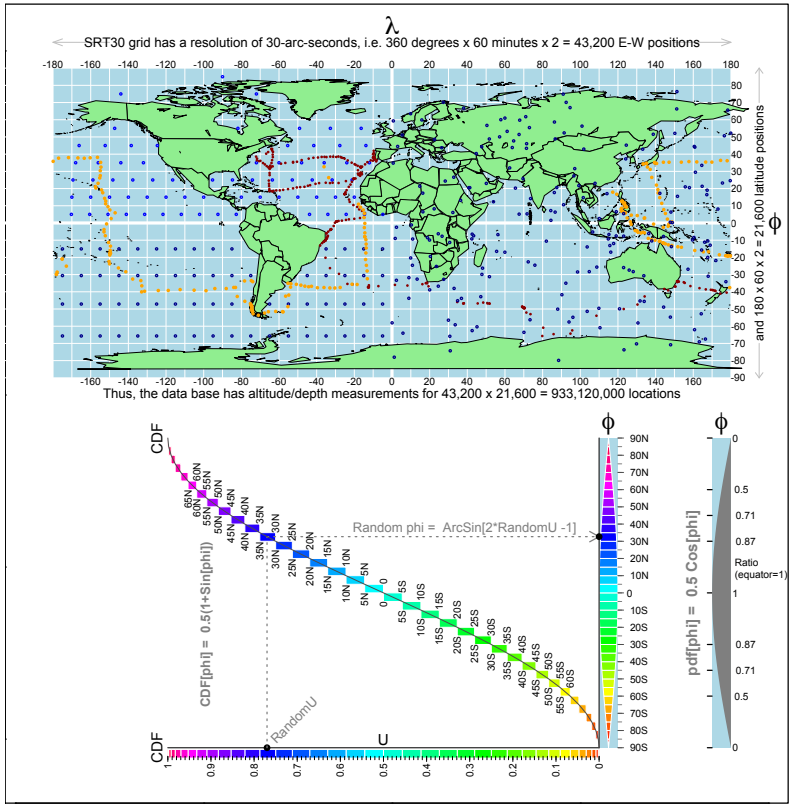
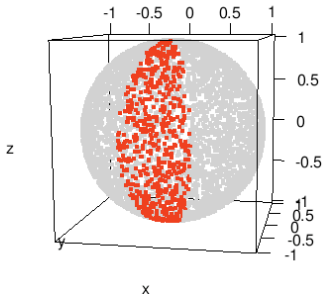
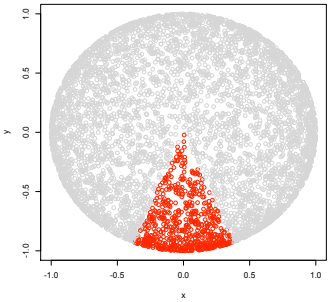


Figure 4: **A**. The resolution in the modern-day SRTM30PLUS database. Schematic representation of the rectangular grid of 933 million recordings in the SRTM30PLUS database, along with the locations of the soundings taken by the outward (red) and return (orange) portions of the 1872-76 Challenger Expedition. The soundings ranged from 4 to 4,475 fathoms: mean approx. 1400 (2700 metres, 1.6 miles). The locations, and the recoded depths, of all 500 soundings can be found online (see `http://19thcenturyscience.org/HMSC/README.htm`). The blue dots are for B.

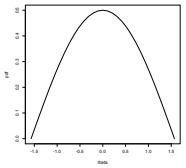
B. (For the Data Mining Challenge) Some ways one might sample from the database to obtain a suitable sample of locations on the earth's surface. The sampling needs to reflect the fact that relative to the length of equator, the length of the corresponding 'line/circle' at latitude ϕ is Cosine[ϕ]. This function is shown in the 'segment-of-an-orange' shape displayed in the blue-background rectangles. In rejection sampling, one generates a ϕ value from U[$90S$, $90N$], and retains it with probability Cosine[ϕ], i.e. as if a randomly selected location inside the rectangle shown at the bottom right 'landed' in the coloured area rather than the light blue background area. Another possibility is to sample ϕ directly, and without any rejection, from U[$-90S$, $90N$], but to differentially weight observations by Cosine[ϕ]. Yet another is to use the 'inverse-CDF' method. The CDF function is best viewed by first rotating the Figure clockwise by 90 degrees; the inverse function is designed to be read in the 'as is' orientation, by (as shown with the dotted lines) entering the diagram on the horizontal (U) scale, and proceeding upwards and to the right to the vertical, (ϕ, latitude), scale. In effect, the method is equivalent to placing all the latitude lines 'end-to-end' and sampling uniformly from this concatenated 'line.' The sequence of small rectangles in the Figure is a necessarily-coarse version of this, whereas the smooth inverse of the smooth CDF curve (shows as a line) allows one to convert a random fractile value (i.e. $U \sim U[0,1]$) into a random latitude. The dark blue dots in A. in the grid representing the western hemisphere are doubly-systematic location samples – in the southern half, along equi-spaced longitude lines, and in the northern half, along equi-spaced latitude lines. The dark blue dots in the eastern hemisphere are locations whose longitudes were sampled from $U \sim U[-180, 180]$, and whose latitudes were sampled – independently of longitude – from the $[-180, 180]$ distribution shown as pdf(ϕ). [JH will remove the 'on land' locations].

```
NOTES (2010) on sampling the surface of a sphere
########## in fact, the earth is not quite spherical ###########
### but we will ignore that for our exercise

# the 'iso-latitude' circle  at a given latitude
# (or the distance between two longitude lines)
# becomes smaller the further the latitude is from the equator.

# If we treat the earth as a sphere, the ratio (relative to that at the equator)
# is cos(latitude * (pi/180)) : cos(0);

# The diameter from pole to pole is shorter than the diameter at the
# equator (it is squished in a bit from both poles)

## See  http://calgary.rasc.ca/latlong.htm
## for more refined calculations

########## random points on a sphere ############

# think of the sections of a (peeled) orange ####

library(rgl)

  n=10000
  open3d()
  x <- runif(n,-1,1)
  y <- runif(n,-1,1)
  z <- runif(n,-1,1) ;  r=sqrt(x^2+y^2+z^2)
  n.inside=sum(r<=1) ; n.inside
  x=x[r<=1]; y=y[r<=1]; z=z[r<=1]; r=r[r<=1]
  x=x/r; y=y/r; z=z/r
  colours=rep("grey80",n.inside)
  in.wedge= ( abs(x/y) < 0.4 & y < 0  )
  colours[in.wedge] = "red"
  plot3d(x, y, z, size=4, col=colours)
  plot(x,y,col=colours)
```



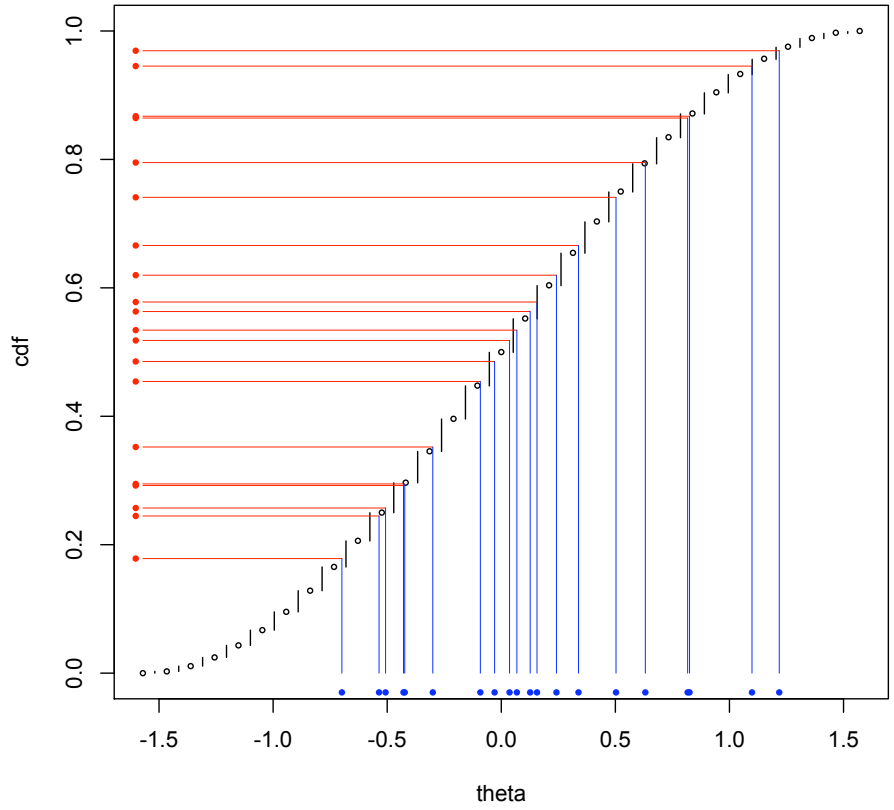



```
# consider a given section,
# width w at lat. theta [ -pi/2 < theta < pi/2 ] is prop. to cos(theta);
# note equator is in middle at theta = 0.

# so pdf(w) prop. to cos(theta)
# normalizing constant [yielding integral of 1] = 0.5, so

# pdf(theta) = 0.5*cos(theta)
```



```
# discrete version... (can make slices smaller and smaller)

n.slices=30; d.theta=pi/n.slices; theta=seq(-pi/2,pi/2, d.theta)

pdf=0.5*cos(theta); plot(theta,pdf, type="l")


# cdf(theta) = 0.5 integral_{-pi/2}^{theta} cos(u) du  = 0.5*(1+sin(theta))

cdf = 0.5*(1+sin(theta)) ; plot(theta, cdf, cex=0.5) # , type="l" )

# longitude lines laid end to end, but in such a way
# that we know which ones are which...

y= cumsum(pdf*d.theta ) # cumulative sum
y=0
for (i in 1:n.slices) {
d.y= 0.5*(pdf[i]+pdf[i+1])*d.theta
    segments(theta[i]+0.5*d.theta, y,  theta[i]+0.5*d.theta, y+d.y)
    y=y+d.y}

# entries at randomly selected vertical locations on (0,1) scale
# find corresponding value on

n.draws=20
for (i in 1:n.draws) {
```

```
random.u = runif(1, 0,1)  # red (input)
# solving 0.5*(1+sin(theta)) = u for theta gives
# theta = inverse.sin (2*u - 1) = arcsin(2*u - 1)
random.theta = asin(2*random.u - 1)  # blue (output)
    points(c(-1.02*pi/2), c(random.u), cex=0.5, pch=19, col="red")
    segments(-pi/2, random.u,  random.theta, random.u, col="red", lwd=0.5)
    segments(random.theta, random.u, random.theta, 0, col="blue", lwd=0.5)
    points( c(random.theta), c(-0.03), cex=0.5, pch=19, col="blue")
}
# [note the Wolfram page uses  acos(2*random.u - 1), but then
# their equator is at pi/2, whereas ours is at 0 ]

# take the theta values where the horizontal lines intersect the longitude lines




# So, to draw from a distribution with a give pdf(.)

# Obtain cdf ...   # draw u ~ Uniform(0,1) ...

# find the . where y intersects cdf

# i.e find ? such that u = cdf(?) # i.e. inverse.cdf(u) = ?

# this works well if inverse.cdf function has closed form

# ANOTHER WAY to remember this way to
# obtain draws from a given distribution ..

# if p is a percentage , then for any p <= 99 ...
# 1 percent of the probability mass lies between the
# p-th and (p+1)-st (per)centiles
# [ can refine this for intervals smaller than 1 percent ]

# there's    1% between  0%-ile and 1%-ile,  1% between  1%-ile and  2%-ile,
#            1% between 10%-ile and 11%-ile, 1% between 11%-ile and 12%-ile,  etc...
# so, if want draws from a distribution, take draws u_1, u_2, u_3, ... from the interval 0-1,
# convert u_i to its counterpart on the x-axis of the cdf...

## jh 2010.09.05 -- corrections/suggestions welcome
```



How to randomly draw Y values (squares) from the (DISCRETE) frequency distribution in this panel?

Each filled circle represents 1% of the probability probability mass. Percentiles are numbered 1-100.

SOLUTION:

Stack the 1% masses vertically, from 1 to 100, so as to trace out a CDF, as shown in next panel.

Then randomly choose which of the 1%'s to draw, and record its Y value.

10 draws shown (and Y values tallied) in blue.

The beige panel shows 1 random draw from a 'CONTINUOUS' distribution. (pdf is shown in lighter colour; its [separate] vertical scale is not shown.)