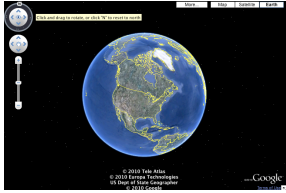


How Deep is the Ocean?

(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 were asked to draw a simple random sample of 200 locations on the Earth's surface,¹ and obtain from the SRTM30_PLUS database the land elevation or ocean depth at each of these. This year, to save some time, both the drawing of the sample, and the 200 database lookups have already been done for you – there is a .csv file with your name on it at the bottom of the 'Oceanography Data' webpage page. From these 'readings', calculate a point estimate of the percentage.² 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 or table that produces an 'exact' confidence interval. Compare the 'exact' interval with the 'approximate' one above.

¹Gelman & Nolan's Teaching Statistics: A Bag of Tricks – McGill Library eBook: see Ch 9 – have 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 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!

- iv. Using what you were able to find online or from your textbooks, explain to a relative who is an engineer how exactly this 'exact' confidence interval is calculated. [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.³ 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.⁴ 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 n' ? Calculate its 10-th and 90-th percentiles (pretend you *know* the value of the parameter that determines its distribution).⁵

³Some 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!

⁴In doing so, make sure to reduce your ' n ' accordingly. Some students in previous years continued to use an n of 200!

⁵R 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.

4 More efficient (or more practical) sampling strategies

(*Very briefly*) describe the circumstances⁶ 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) 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. 7) was considered too technical for the Significance article. It is included on p. 7 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., $\phi = 45.5\text{N}$), 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.

(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) the pdf, (ii) the CDF has a closed form.

⁶The Cross-Canada Survey of Radon Concentrations in Homes [Resources] might help.

(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. 5) tries to explain the 'inverse CDF' method in pictures rather than via calculus.

Pages 7-8 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 10 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.⁷

This article <http://www.biostar.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.

⁷In 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 2010-2017



Janvier 2012							
Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	A Faire
9593 3844	5157 2494	3202 3202 Battery	2605 2605	4856 6495	7487 363	6779 956	
6640	10052	7347	7146	7529	4628	6750	
571	4487	7899	299	649	729	1500	
8230	6759	12926	15278	10538	11255	15954	
1576	5057	8804	15393	11645	11447	8348	
9595	12432	7421	14000	8280	7274	18378	
6971	10475	5567	12649	7586	2791	15690	
8760	7044	9337					
5819	10864	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 four files (2010-2011, 2012-2013, 2014-2015, and 2016-2017) 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, 2014-2015.pdf and 2016-2017.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 8 years. Those for the first 2 years are easy to obtain, but to obtain exact values for 2012 to 2017 would take more data entry work than is reasonable for a single assignment.⁸

* 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 2017.

* 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 2010-2017. 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.⁹

⁸Unfortunately, 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.

⁹JH 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 2017.

7 Gasoline consumption of family minivan

Date	Amount	Price
April 2006	214 L	\$29.4
April 19	259.3 L	\$20.10
April 26	325 L	\$26.50
May 2	262 L	\$20.28
May 8	842 L	\$35.00
May 15	326.2 L	\$26.47
May 18	1334 L	\$13.85
May 23	1431 L	\$20.45
June 1	1647 L	\$20.45

Date	Amount	Price
July 29	87330 L	\$48.9
July 30	87613 L	\$33.61
Aug 2	88290 L	\$75.00
Aug 6	88817 L	\$39.00
Aug 7	89243 L	\$35.03
Aug 9	89423 L	\$66.3
Aug 10	89863 L	\$45.05
Aug 11	90346 L	\$38.40

Under the heading “Automobile Fuel Purchases: Toyota Sienna April 2006 - June 2017” near the top of the Resources webpage you will find the documentation of the gasoline purchases for a Toyota Sienna (cf. extracts above). All of the details up until earlier this year can be found in the pdf file of 68 pages.

In order to analyze the fuel consumption over the 11 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 litres¹⁰ – **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, \dots, 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.

¹⁰litres = LitresPurchased + 3.78541 × GallonsPurchased

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, \dots, 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’¹¹ standard error.

For the first, which uses the *page* as the sampling unit, estimate $\text{Var}[\bar{l}_5]$ as $(1/5) \times$ the s^2 of the 5 l ’s \times the finite population correction, $(40-5)/(40-1)$.¹²

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

Once you have computed these, you are asked to form a 50% confidence interval¹³ 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.

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 blog¹⁴, [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?

¹¹An expert in sampling might make a somewhat more refined estimate of the $L/100\text{Km}$, and a more refined standard error calculation

¹²Textbooks vary as for the expression to be used.

¹³No, the 50% is not a typo.

¹⁴Horton also blogged about this in 2011: <https://www.r-bloggers.com/example-8-36-quadratic-equation-with-real-roots/>

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 100 times, record on a sheet of paper the sequence of heads and tails, and hand it in. Many students took a ‘shortcut’ and made up the sequence ‘out of their head.’

- Out of your head, ‘make up’¹⁵ a random sequence of the results (Head=1, Tail=0) of 100 tosses. enter this sequence of one hundred 0’s and 1’s into an R vector named ‘sequence’.¹⁶ 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 Berkeley teacher might judge whether the student ‘made up’ the sequence, or actually took the time and tossed the coin 100 times.

10 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.

¹⁵Making it up is, of course, much faster than actually tossing the coin 100 times, and recording the sequence.

¹⁶`sequence=c(, ,)`

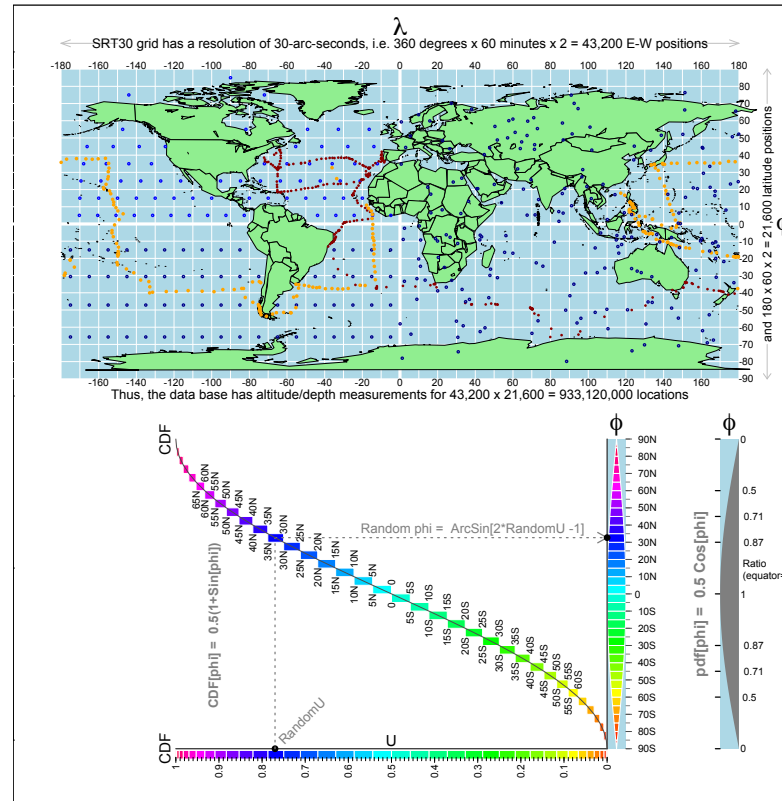


Figure 4: **A.** The resolution in the modern-day SRTM30PLUS database. Schematic representation of the rectangular grid of 933 million recordings in the SRTM30PLUS database, along with the locations of the soundings taken by the outward (red) and return (orange) portions of the 1872-76 Challenger Expedition. The soundings ranged from 4 to 4,475 fathoms: mean approx. 1400 (2700 metres, 1.6 miles). The locations, and the recorded 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 $\text{Cosine}[\phi]$. 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 $\text{Cosine}[\phi]$, 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 $\text{Cosine}[\phi]$. 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 $\text{pdf}(\phi)$. [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

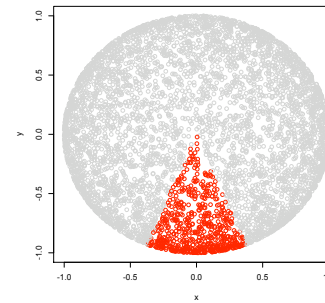
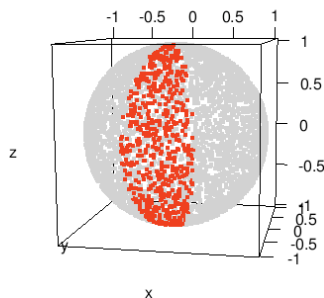
##### random points on a sphere #####

# 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)

```



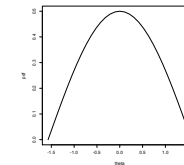
```

# 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

```

