

Being approximately correct and being precisely wrong

1. Refer to the descriptions of the SMOG index, the Fry method, the Flesch Reading Ease, and the Flesch-Kincaid Grade Level, for measuring **readability** (under Resources for Measurement/Surveys).<sup>1</sup>

For the article or text you have chosen (as per discussion in class), randomly select three separate 100 word passages, and use this *set of three passages* to measure the readability ( $F_1$ ) using the Fry graph. Rather than do so manually, you can use the SMOG calculator to determine the average number of sentences and syllables per hundred words. Repeat the readability measurement ( $F_2$ ) with a second *different* set of three passages. Repeat once more ( $F_3$ ), using a *third set*.

Using these same three sets, calculate the SMOG index, the Flesch Reading Ease, and the Flesch-Kincaid Grade Level.

For each index, use the 3 estimates to calculate the standard error of measurement, and the coefficient of variation. Comment.

2. Propose a method to assess the *validity* of a readability index.
3. [m-s] Derive the link between the standard error of measurement and the (intraclass correlation) reliability coefficient [last line, column 1, p. 7 of notes on “Quantifying Reliability” in Notes on Psychometrics for students in rehabilitation sciences in Resources for Measurement/Surveys. *Hint: it’s simply a matter of using the definition of R.*
4. [m-s] Exercise in section 3: Relationship between test-retest correlation and ICC(X) [In notes on Effect of Errors in X and Y on measured correlation and slope]

5. [m-s] Exercise section 4: Relationship between correlation( $X, X'$ ) and ICC(X) [ibid.]
6. Francis Galton (1822-1911) found that the correlation between (*self-reported*) **parental** and (adult) **offspring heights** was strongest for the one between father and son [ $0.396 \pm 0.024$ ], and weakest for the one between mother and daughter [ $0.284 \pm 0.028$ ]. Given the way he obtained the measurements, can you imagine why this was? <sup>2</sup>  
[It was  $0.302 \pm 0.027$  for mother & son;  $0.360 \pm 0.026$  for father & daughter.]

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5	5.5, 5.5
3	15.0	about 4.0	11.0	8.0
4	15.0	4.0	10.5, 8.5	7.0, 4.5, 3.0
5	15.0	1.5	12.0, 9.0, 8.0	6.5, 2.5, 2.5
6	14.0	8.0		9.5
7	14.0	3.0	16.5, 14.0, 13.0, 13.0	10.5, 4.0
8	14.0	6.5		10.5, 8.0, 6.0
9	14.5	6.0		6.0
10	14.0	5.5		5.5
11	14.0	2.0	14.0, 10.0	8.0, 7.0, 7.0, 6.0, 3.5, 3.0
12	14.0	1.0		5.0
13	13.0	7.0	11.0	2.0
14	13.0	7.0	8.0, 7.0	
15	13.0	6.5	11.0, 10.5	6.7
16	15.0	about 5.0	12.0, 10.5, 10.2, 10.2, 9.2	8.7, 6.5, 4.5, 3.5
17	13.0	4.5	14.0, 13.0, 11.5, 2.5	6.5, 2.3
18	13.6	4.0		6.0, 4.5, 4.0
19	13.2	3.0		2.7
20	12.7	9.0	13.2, 13.0, 12.7	10.0, 9.0, 8.5, 8.0, 6.0
21	12.0	8.0	13.0	3.5, 2.0
22	12.0	abt 7.0	13.0, 11.0	7.0
23	12.0	5.0	14.2, 10.5, 9.5	6.0, 5.5, 5.0, 5.0
24	12.0	5.5		5.5

Family heights: Page 1/8 of notebook in Galton Papers : see “Galton’s family data on human stature” – the link is on the left hand side of JH’s home page.

<sup>2</sup>After you have thought about it for a while, and looked carefully at Galton’s Notebook, you might wish to compare your answer with that given by Karl Pearson: Cf. “Why Galton got different parent-offspring correlations in heights and he (KP) got a larger ones” in the ‘Measurement – Lecture Notes, etc’ section of the bios601 resources page for Measurement.

<sup>1</sup>ToneCheck (<http://tonecheck.com/>) is another interesting tool. See story at <http://www.montrealgazette.com/search/search.html?q=ToneCheck>

7. **Bridging the physical- and the psycho-metric:** The notes on “Increasing Reliability by averaging several measurements” on the right hand column of page 4 of JH’s notes on Quantifying Reliability give the formula for the so-called “Stepped-Up Reliability”. In psychometrics (where the number of items on a test serves as the “several measurements”) this formula serves as the basis for the “Spearman-Brown prediction formula”.<sup>3</sup> [m-s] Invert the formula on p.4 to derive the one on the right hand column of p.1 for Spearman-Brown prediction formula relating the reliability of two versions of a test, one with N times more items than the other.
8. You are trying to estimate, from **imperfect observations** of  $F$  and  $C$ , the values of the two coefficients  $B_0$  and  $B_1$  in the temperature relation  $F = B_0 + B_1 \times C$ .

For each of the following situations, and using the true values of  $B_0 = 32$  and  $B_1 = 9/5 = 1.8$ , simulate<sup>4</sup> 1000 datasets & investigate the behaviour of the 1000 estimates,  $b_0$  and  $b_1$ , of  $B_0$  and  $B_1$ . In each simulation, use samples of size  $n = 4$ , with temperatures of  $C = 14, 16, 18$  and  $20$ .

- (a)  $C$  measured perfectly,  $F$  measured with  $\epsilon_F \sim \text{Gaussian}(\mu = 0, \sigma_{\epsilon_F} = 1)$  errors that are independent of  $F$ . Check – formally, using a test (or CI) based on the mean of the 1000 estimates – for evidence of bias in  $b_1$ . Also check whether the empirical variance of  $b_1$  agrees with that given by the theoretical formula, namely

$$\text{Var}(b_1) = \sigma_{\epsilon_F}^2 / \sum (x - \bar{x})^2.$$

- (b)  $F$  measured perfectly,  $C$  measured with  $\epsilon_C \sim \text{Gaussian}(\mu = 0, \sigma_{\epsilon_C} = 1)$  errors that are independent of  $C$  [*Classical type* error: someone else chose situations when  $C$  was indeed exactly 14, 16, etc, but didn’t tell you what  $C$  was, and instead asked you to independently record  $C$  using your own imperfect instrument, and to use *your* recordings of  $C$  in your estimation of the equation]. Again, formally test for evidence of bias in  $b_1$ .

Do your findings line up with the predictions in the Notes? If the patterns are difficult to see, you might change the number of simulations, the sizes of the errors, the range of  $C$  or the sample size.<sup>5</sup>

<sup>3</sup>[http://en.wikipedia.org/wiki/Spearman-Brown\\_prediction\\_formula](http://en.wikipedia.org/wiki/Spearman-Brown_prediction_formula) .

<sup>4</sup>If new to simulations, see “Computer code to simulate datasets with measurement error” at the bottom of the Resources webpage for measurement/surveys. It gives some ‘starter’ computer code, which you can modify to suit.

<sup>5</sup>The article by Hutcheon et al. “Random measurement error and **regression dilution bias**”, under ‘**r e p r i n t s**’ on JH’s home page, tries to explain these patterns intuitively.

9. Before we study how well we can digitize survival curves, here is an **exercise on communicating what the curves are meant to convey** and the context in which they were generated.

Refer to the article “Associations between C-reactive protein, coronary artery calcium, and cardiovascular events: implications for the JUPITER population from MESA, a population-based cohort study”, available in the Resources link opposite ‘Applications’ in bios601. We digitized the lowermost (green) curve in Figure 2A of that article.

- (a) Read the Abstract and study the Figures in the article. Then, write, *in your own words*, a short news item of 250 words or so (2-3 minutes or so on radio) for your local newspaper and radio station, where you moonlight as a health reporter. In your piece address (i) the rationale for the study (ii) the principal findings and (iii) the implications of these findings. Also suggest a headline for your story. [You might want to study some health reports to see how they are structured.. the order may not be the (i)-(iii) order listed above. An interesting but slightly more highbrow website devoted to science reporting in general is <http://www.sciencedaily.com/>. The websites  
... <http://www.cnn.com/HEALTH/>,  
... <http://www.nytimes.com/pages/health/index.html>,  
... <http://www.bbc.co.uk/news/health/> and  
... <http://www.cbc.ca/news/health/>  
are also worth consulting, and indeed monitoring. ]
- (b) A 65-year old relative of yours reads your story, looks on the internet and finds that a test that measures coronary artery calcium is available in a private clinic in Montreal, and phones you to ask if it would be worth being tested and getting her “score”. What would you say to this relative?

## 10. Errors in digitization

Refer to the duplicate readings you made of the Kaplan-Meier survival curve in the study entitled “Associations between C-reactive protein, coronary artery calcium, and cardiovascular events: implications for the JUPITER population from MESA, a population-based cohort study” available in the Resources link opposite ‘Applications’ in bios601

For now, ignore the point-wise measures of precision, i.e., the standard errors and confidence intervals, that often accompany such curves. These are (decreasing) functions of the numbers of subjects and the numbers of ‘events’; we will cover their calculation later in the term. For now, focus only on the loss of precision as a result of your digitization.

Focus on your two measurements of each of the reported  $y$ -year risks, where  $y = 1, 2, 3, 4, 5, 6, 7$ :

$$y\text{-year CHD risk} = 100 \times (1 - \text{proportion free of CHD at year } y)\%$$

- (a) From your two measurements at each of the 7 timepoints, obtain a 7d.f. estimate of the ‘standard error of measurement’. Do so using a ‘canned’ statistical routine and also ‘from scratch’ in R. Write out the statistical model that you used to obtain this, and list any assumptions it makes.
- (b) The estimate in (a) is an estimate of the ‘within’ observer variation. In order to estimate the ‘*between*’-observer variation, what is the *minimal* information you would need from each of you co-observers? (since JH has access to all of them, he will supply each of them once you email him with your specific request: he can supply the full raw data that could be then put into a canned statistical routine, but he would prefer that you do the calculations ‘from scratch’ in R). Again, write out the statistical model that you used to obtain this, and list any assumptions it makes.
- (c) Here the ‘objects’ to be measured were 7 very specific (fixed) timepoints. Assume for the sake of this exercise that the 7 objects were 7 randomly selected human subjects and that we were interested in calculating an intra-class correlation coefficient to serve as a reliability measure. Carry out the ICC calculation. Restrict your attention to years 1-5 and recalculate the new ICC. Comment on why the ICC becomes smaller.

11. **Bernoulli Error?** A not-discovered-for-almost-300-years error in Bernoulli’s book? Or a not-discovered-for-almost-7-years error by A.W.F. Edwards. *Which is it?*

In his ‘Ars conjectandi three hundred years on’ article in Significance Magazine, Cambridge University Professor Edwards tells us that, a few years ago, he was reviewing Sylla’s English translation of (Jacob) Bernoulli’s book. He worked through one of the expectation problems, and came up with a different answer than Bernoulli. In early June of 2013, a week before the Edwards item was published in Significance, Julian Champkin, the magazine Editor, and a journalist by profession, used this ‘300-year-old error’ in the ‘trailer/teaser’ for the upcoming piece, and his question ‘Can you correct it?’ generated a number of responses on the Significance website.

In the bios601 resources for surveys and measurement, at the bottom of the Webpage, JH has collected together in one .pdf file the item by Champkin, some of the original Bernoulli text in Latin, the full article by Edwards, the Edwards review of the Sylla translation into English, and Sylla’s translation of Bernoulli’s treatment of the problem.

The question arises as to whether it is the probabilities that are incorrect, or the expectation based on them, or whether it is Edwards who is incorrect.

***What is your answer?*** [Remember that Edwards had studied Bernoulli earlier, when writing his book on Pascal’s triangle, and had found an error, that had been reproduced over the centuries in different books, in a table of Bernoulli numbers. So might Bernoulli (or the printers) had been a little bit careless?]

## 12. Imprecision in recording event times

The Introduction to a recent (2013) journal article “Driving under the (Cellular) Influence” by Saurabh Bhargava and Vikram S. Pathania of Carnegie Mellon University begins:

Does talking on a cell phone while driving increase your risk of a crash? The popular belief is that it does – a recent New York Times/CBS News survey found that 80 percent of Americans believe that cell phone use should be banned. This belief is echoed by recent research. Over the last few years, more than 125 published studies have examined the impact of driver cell phone use on vehicular crashes. In an influential paper published in the New England Journal of Medicine, Redelmeier and Tibshirani (1997) – henceforth, RT – concluded that cell phones increase the relative likelihood of a crash by a factor of 4.3. Laboratory and epidemiological studies have further compared the relative crash risk of phone use while driving to that produced by illicit levels of alcohol.

Later, in bios602, you will be introduced to the very clever study design that RT used to arrive at the 4.3.

The 2013 authors then go on to study the topic using a very different but also clever design.

We investigate the causal link between driver cell phone use and crash rates by exploiting a natural experiment induced by the 9pm price discontinuity that characterizes a majority of recent cellular plans. We first document a 7.2 percent jump in driver call likelihood at the 9 pm threshold. Using a prior period as a comparison, we next document no corresponding change in the relative crash rate. Our estimates imply an upper bound in the crash risk odds ratio of 3.0, which rejects the 4.3 asserted by Redelmeier and Tibshirani (1997). Additional panel analyses of cell phone ownership and cellular bans confirm our result.

But while they had very precise data on when cell phones were being used, (see Fig2) the data on crashes were quite messy. To quote the authors:

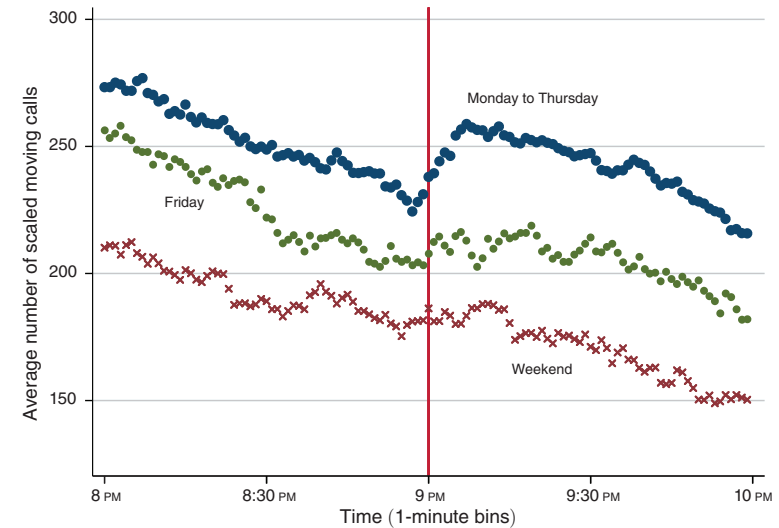


FIGURE 2. CELL PHONE CALL VOLUME FROM MOVING VEHICLES FOR CALIFORNIA FROM 8PM TO 10PM IN 2005

Our analysis principally relies on two sources of crash data. First, the State Data System (SDS) provides data for the universe of reported crashes from 1990 to 2005 for California, Florida, Illinois, Kansas, Maryland, Mississippi, Missouri, Ohio, and Pennsylvania. A well recognized drawback of using a crash database based on self-reports is the presence of substantive periodic *heaping*.

The trajectory of a crash record helps to illuminate the origins of this bias. Once a vehicular crash is reported, police at the scene document various details of the incident, including the minute of the crash occurrence, and submits the paperwork to one of several possible state agencies. While states vary in the specifics that govern data collection and crash qualification criteria, crash records are ultimately centralized and sent once a year to the NHTSA where they are standardized and maintained.

Figure 4 illustrates the nature of the heaping in reports

that characterizes a representative hour in 2005 across the states in our sample. **A close examination indicates that nearly 11 percent of crash reports fall exactly on the hour, 31 percent are on the hour, half hour, or quarter hour, and 61 percent reside in a minute ending in either zero or five.**

VOL. 5 NO. 3 BHARGAVA AND PATHANIA: DRIVING UNDER THE (CELLULAR) INFLUENCE 103

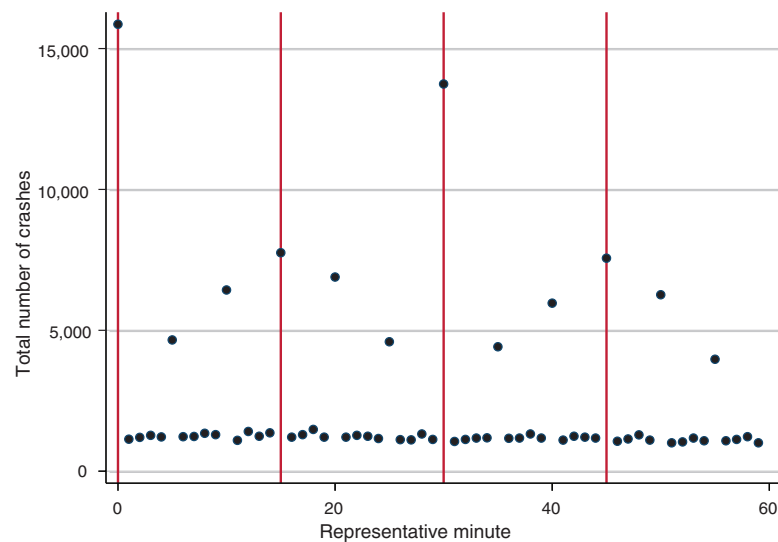


FIGURE 4. PERIODICITY IN SIDS CRASHES ACROSS REPRESENTATIVE HOUR IN 2005 FOR ALL STATES IN SAMPLE

**Exercise:** In this study, the primary contrast involves crash rates in the 1 hour after and the 1 hour before cellphone calls became “free” at 9 pm. Do you think the heaping errors are an insurmountable problem? If you do, why? If not, suggest ways to deal with them.

### 13. Galton’s data more than century later

[See also Questions 3-5 above, and see JH’s notes on Quantifying Reliability under the Measurement Lecture Notes heading in the website]

The 1985 article “Galton’s Data a Century Data” re-analyzes the extensive data collected by Francis Galton at his anthropometric laboratory in the South Kensington Museum in London.

JH has contacted one of the authors (Frank Ahern) who replied that “Despite a great deal of searching, neither I or Jerry McClearn have been able to find the original data that were used back in ’85.”

So, we will start again. But this time, instead of having to go to London and photocopy the records, you can take advantage of the scanned copies provided by the Wellcome Library and the Galton archives. To save you having to find the books (each containing about 500 records) in the large amount of material in the Galton archives, JH has downloaded them and put them on the bios601 website, in the Resources for Sampling/Measurement folder, under the heading (flagged in red) “Data from Galton’s Anthropometric Laboratory.”

For this exercise, which is designed to familiarize you with how to statistically quantify the psychometric (and psychophysical) properties of different measuring instruments, *we will focus on subjects who have been measured more than once*, so that we can assess the *reliability* of the various measures. For now, we will ignore the fact that there is quite a bit of time between some of the measurements, and that some attributes are age-related (we will try later to see at what age the peak is), and so some of the non-repeatability is for legitimate biological reasons.

So as to get a feel for the (small sample) sampling variability of these measures, and also so that it is not too big a data entry burden, you are asked to enter the complete records for 10 such subjects, i.e., subjects who were measured on more than one date. We can pool these student datasets later to get a more – statistically – reliable estimate of the various reliability measures.

In order to standardize the variable names, and provide a small element of quality control, a .csv file (**Spreadsheet for Data Entry**) with several subjects from the first book is provided on the website, immediately after the data books. Add to it the data for the *first ten* eligible ones you find in the range assigned to you (enter *all* of the records per subject, no matter how close or far apart they are in time). After you have added your entries, delete the ones already there — they were merely provided so as to standardize the naming of variables, and to act as a guide to

align the columns correctly, and to make it easier to see any items that are mis-entered.

A few notes at this point (we may discover other oddities that we need to deal with as we go along). JH has noticed that subsequent measurements are some times recorded in metric units rather than Imperial (e.g., cm instead of inches and tenths or inches). We could discuss other ways to enter such mixed units (from JH's past experience, converting as we enter is not an option!) but JH decided that when he met a metric measurement when he had allocated a pair of fields for say inches and tenths, he simply put the metric measurement in the first field and left the second field blank. It should be relatively easy to use programming to harmonize them later.

In the case of blanks, or illegible recordings, please leave the field blank.

JH has noticed some instances where there were several (4 in subject 0001) rows for the first several items (up to the Snellen test) but fewer (e.g. 2 in subject 0001) rows for the later items at the bottom of the page, from sitting height to strength of blow with fist. In such instances, use any indications you can to decide which rows at the bottom of the page go with which ones at the top (in the case cited, JH decided that the first and fourth rows were complete, as were both of the bottom ones, so he put these with the first and fourth). In such cases, use the remarks column to flag the case.

Here are the books assigned to the different students. Contact JH if your ID number is not in the list.

ID	Subjects
JH	0001-0491
26xxxxx21	0511-1028
26xxxxx19	1029-1530
26xxxxx57	1531-2020
26xxxxx99	2021-2520
26xxxxx78	2521-3021
26xxxxx65	3022-3521
26xxxxx58	3522-4000
26xxxxx90	4001-4500
26xxxxx94	4501-5000
	5001-5500
	5501-6000
	6001-6500
	7001-7459

Once you have entered the data, adopt the supplied R code to calculate the ICC for each of the measures shown in Table 1 of the 1985 article. Do not worry about timing or segregation by sex, or age-correction – you will not have enough data to do so; we will do this later when we pool the data. It appears (but JH is not entirely certain) that the 1985 authors used a simple Pearson product moment correlation with paired measurements. The advantage of the ICC is that while it is still connected mathematically with the Pearson correlation (see exercises above), it is more general and it uses whatever number of measurements per person there are. It is less cumbersome than using all possible pairwise correlations, or selecting just two.

Compare the ICCs with the test-retest correlations in Table 1 of the 1985 'a century later' paper, and comment on any substantial differences.



14. Physical Activity: JH 2010-2013

Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	A Faire
1 9593 3844	2 5157 2494	3 3202 3202 Battery	4 2605	5 4856 6445	6 7487 363	7 6779 956	
8 6690 571	9 10052 4484	10 7347 7899	11 7146 299	12 7529 649	13 4628 729	14 6750 1500 Battery	
15 8230 1576	16 6759 5057	17 12926 8804	18 15278 15393	19 10538 11645	20 11255 11447	21 15954 8348	
22 9595 6971	23 12432 10475	24 7421 5562	25 14000 12649	26 8280 7586	27 7274 2791	28 18378 15690	
29 8960 5819	30 7044 10864	31 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 two files (2010-2011 and 2012-2013) 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 file has scanned images (see above right) of the pages of paired recordings from the log-book.

The exercise in sampling from these data raised the issue of how many days one needs to sample in order to ensure that the estimate one gets is close to what one would obtain with a census, i.e., a 100% sample of days. Similar issues occur in dietary recall surveys. The least costly method is the food frequency questionnaire (Google for more info); a much more costly one is the x-day 24-Hour dietary recall method. How large x should be for different sub-populations (e.g., children, young adults, the elderly) has been studied. In measuring physical activity, it is common to use quite expensive accelerometers, and so they are usually given to research subjects for just one randomly chosen week.

The Omron model shown costs a lot less, and unlike the accelerometers –

which store minute by minute activity – just records the number of steps for each of the last 7 days. JH’s data help us answer the question of how many weeks are needed to get a good estimate of his yearly activity.

(a) divide the 2010-2011 data into weeks, and derive a (somewhat oversimplified) 1-way analysis of variance table, with week as the factor.

in this greatly oversimplified model, the numbers of steps ( $y$ ) on any day ( $j$ ) within week  $w$  ( $i=1 \dots 104$ ) can be written as

$$y_{w,j} = \mu + b_w + \epsilon_{w,j}$$

(b) For didactic purposes, treat the model as a random-effects one, i.e., with week as the random factor. Thus, the 104  $b_w$ ’s are assumed to be a random sample drawn from a  $N(0, \sigma_w^2)$  distribution.<sup>6</sup> Even though they may have a lot of structure, treat the variations across days within a week as uncorrelated ‘disturbances’ or ‘errors’ ( $\epsilon_{yr,w,y,j}$ ) with variance  $\sigma^2$  but no structure (i.e. treat all  $\epsilon$ ’s as exchangeable, so that order of observations within the same week is irrelevant – in the file, you only need to know which week it is, not which day of the week. Clearly, there may be strong intra-week patterns, but for now assume that you are not even told which observation corresponds to which day of the week.

From the Expected Mean Squares (EMS) for this model<sup>7</sup>

Source	Sum of Squares	df	Mean Square	EMS
Weeks	$SS_w$	103	$SS_w/df$	$\sigma^2 + 7\sigma_w^2$
Error	$SS_e$	$104 \times 6$	$SS_e/df$	$\sigma^2$

use the method of moments to estimate the  $\sigma_w^2$  and  $\sigma^2$  components.

(c) Using the results from (b), and the same overly simplified model, work out the expected variance of estimators that average recordings from (i) 3 random days in 1 random week (ii) 1 random day in each of 3 random weeks (iii) 3 random days in each of 3 random weeks.

(d) Could you have arrived at the results in (c) using the ‘Stepped-Up’ Reliability formula referred to in page 4 of the Quantifying Reliability notes?

<sup>6</sup>Using Roman b’s and Greek  $\beta$ ’s to distinguish random effects from fixed effects is a recent convention: it was not used when JH learned linear models.

<sup>7</sup>See also pages 4 and 5 of Notes on Introduction to Measurement Statistics, and pages 3 and 4 of the Notes on Quantifying Reliability (on the Resources website, under the heading ‘Measurement – Lecture Notes, etc’). ‘Weeks’ in the current example correspond to ‘persons’ or ‘subjects’ or ‘families’ in those examples.

**15. Repeatability of a Test – and of the statistical analysis itself!**

Refer to the report ‘A Novel Test of Endurance Running Performance’ in the Resources website [under the tab ‘Data from various repeatability studies’].

- (a) Redo the 2-way ANOVA ‘with participant and trial as main effects’ to see if you can reproduce the reported coefficient of variation.
- (b) Use a 1-way ANOVA, with subjects as a random effect, and the 3 trials as replicates (i.e. ignoring the order) and calculate an overall coefficient of variation. [A very similar 1-way ANOVA is shown in the 1st column of page 5 of the ‘Introduction to Measurement Statistics’ Notes on the Resources website. Page 3 of the Notes ‘Quantifying Reliability’ has an example with 2 measurements per family, but the principle is the same.]  
Which makes more sense to you, the CV based on their 2-way ANOVA, or yours based on a 1-way ANOVA?
- (c) Calculate subject-specific coefficients of variation (just as was reported in Table 1 in the article on breath alcohol – the link to this article can be found just above the one for the endurance test). Summarize the 10 CVs using say the median and the range. Would you report the ‘overall’ CV the authors did, or some summary of the 10 subject-specific ones? Give a reason for your choice.
- (d) Use the results of the 1-way ANOVA<sup>8</sup> to calculate an intra-class correlation (ICC).
- (e) In this setting, which makes more sense, a CV or an ICC? Why?
- (f) Rerun the ICC code several times on random subsets of the subjects. As you reduce the sample size to just 2 or 3, does the ICC stay stable? Use the example to say what the ICC tells us that the CV can not, and what the CV tells us that the ICC can not.
- (g) How could one ‘rig’ (i.e., manipulate) the sample of subjects in the breath alcohol study to (i) maximize (ii) minimize the ICC?

**16. How reproducible and accurate are free smartphone apps to track your steps, calories burned, distance and active time?**

The letter ‘Accuracy of Smartphone Applications and Wearable Devices for Tracking Physical Activity Data’ in JAMA in February 2015 [under the tab ‘Data from various repeatability studies’] reports

This prospective study recruited healthy adults aged 18 years or older through direct verbal outreach at a university. Participants gave verbal informed consent to walk on a treadmill set at 3.0 mph for 500 and 1500 steps, each twice, for no compensation. An observer (M.A.C.) counted steps using a tally counter in August 2014. This study was approved by the University of Pennsylvania institutional review board.

A convenience sample of 10 applications and devices was selected from among the top sellers in the United States. On the waistband, each participant wore the Digi-Walker SW-200 pedometer (Yamax), which has been well validated for research,6 and 2 accelerometers: the Zip and One (Fitbit). On the wrist, each wore 3 wearable devices: the Flex (Fitbit), the UP24 (Jawbone), and the Fuelband (Nike). In one pants pocket, each carried an iPhone 5s (Apple) simultaneously running 3 iOS applications: Fitbit (Fitbit), Health Mate (Withings), and Moves (ProtoGeo Oy). In the other pants pocket, each carried the Galaxy S4 (Samsung Electronics) running 1 Android application: Moves (ProtoGeo Oy).

...

Across all devices, 552 step count observations were recorded from 14 participants in 56 walking trials. Participants were 71.4% female, had a mean (SD) age of 28.1 (6.2) years, and had a mean (SD) self-reported body mass index (calculated as weight in kilograms divided by height in meters squared) of 22.7 (1.5).

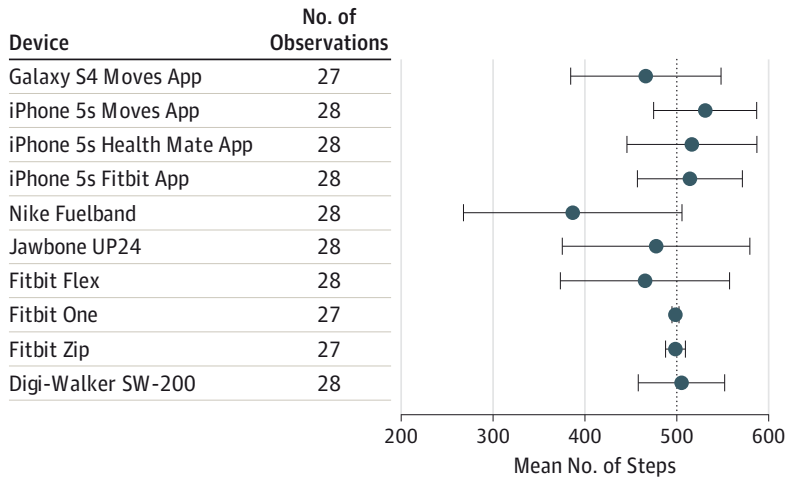
...

Figure 1 shows the results for the 500 step trials by device and Figure 2 shows the results for the 1500 step trials. Compared with direct observation, the relative difference in mean step count ranged from -0.3% to 1.0% for the pedometer and accelerometers, -22.7% to -1.5% for the wearable devices, and -6.7% to 6.2% for smartphone applications. Findings were mostly consistent between the 500 and 1500 step trials.

<sup>8</sup>The R code supplied makes use of an ICC package, but it is always safer to check with a worked example that a package you don’t know is doing what you want it to do.

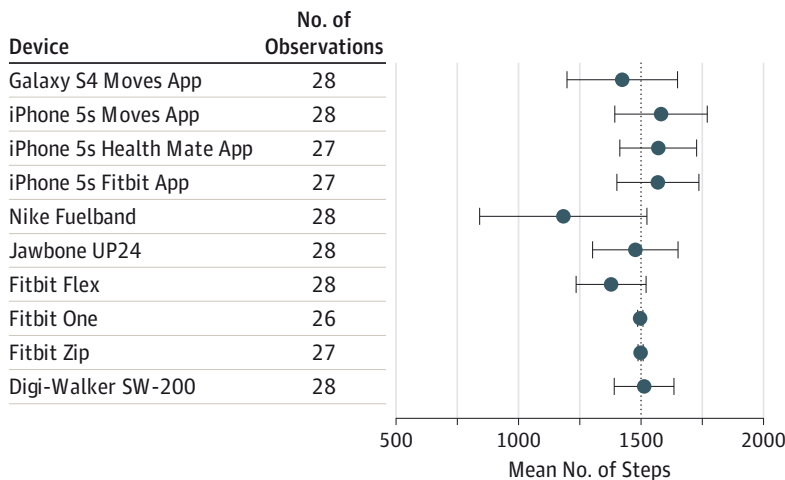


**Figure 1. Device Outcomes for the 500 Step Trials**



The vertical dotted line depicts the observed step count. The error bars indicate  $\pm 1$  SD.

**Figure 2. Device Outcomes for the 1500 Step Trials**



The vertical dotted line depicts the observed step count. The error bars indicate  $\pm 1$  SD.

- (a) Rewrite the authors findings using the words ‘under-’ and ‘over-counted.’
- (b) For which instruments is there evidence that this ‘bias’ is non-zero? *You can use your eye to determine the means and SDs, or use the ones in the .pdf file shared by senior author (‘I’m attaching the raw data that we have to share’) and available on the course website.*
- (c) The data summaries were in response to an email from JH to the author, asking if there was ‘any chance you would be able to share the Excel file of raw data, so we should see if the deviations from the target were all over the place, or peculiar to a few people or a few devices. I can imagine the pockets on some people being a bit deep and wide.. and that the machines in them slosh around – I sometimes keep my \$20 dollar step counter in my pocket instead of on my belt.’

Imagine that the author had shared these data as 552 separate lines, each one containing a step count, a participant ID (1-14), the target (500 or 1500), the occasion (1st or 2nd) and the name of the device.<sup>9</sup> Write out a plan for analyzing them, including the model you would use, the meaning of each component (parameter) in the statistical model, how you would estimate each component, a table of results (use made up, but realistic numbers), and a sketch of one or more graphs that would quickly tell the same story.

- (d) This Fall, the EPIB601 and the BIOS601 classes will carry out their own investigations. The Epidemiology teacher is planning to test an app called Pacer - Pedometer plus Weight Loss and BMI Tracker By Pacer Health, Inc that is available for free for both the iPhone and Android devices. Dr Patel (senior author of the letter) ‘particularly like[s] Withings HealthMate because it has a good user interface and works with both iPhones and Androids. Fitbit is also good but works with a limited set of Androids.’

For the BIOS601 class of -- -- , come prepared to participate in a planning session, where together we will design (and subsequently carry out) our investigation into the reproducibility and validity of a few smartphone apps with respect to steps, distance, calories, etc.

<sup>9</sup>At the end of each trial, step counts from each device were recorded. In rare instances that a device was not properly set to record steps (8 of 560 observations), these data were not included. The mean step count and standard deviation for each device was estimated using Excel (Microsoft). Across all devices, 552 step count observations were recorded from 14 participants in 56 walking trials.

17. 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.

XXI. *Experiments to determine the Density of the Earth.* By Henry Cavendish, Esq. F.R.S. and A.S.

Read June 21, 1798.

MANY years ago, the late Rev. JOHN MICHELL, of this Society, contrived a method of determining the density of the earth, by rendering sensible the attraction of small quantities of matter; but, as he was engaged in other pursuits, he did not complete the apparatus till a short time before his death, and did not live to make any experiments with it. After his death, the apparatus came to the Rev. FRANCIS JOHN HYDE WOLLASTON, Jacksonian Professor at Cambridge, who, not having conveniences for making experiments with it, in the manner he could wish, was so good as to give it to me.

*The following Table contains the Result of the Experiments.*

Exper.	Mot. weight	Mot. arm	Do. corr.	Time vib.	Do. corr.	Density.
1	m. to +	14,32	13,42	"	-	5,5
	+ to m.	14,1	13,17	14,55	-	5,61
2	m. to +	15,87	14,69	-	-	4,88
	+ to m.	15,45	14,14	14,42	-	5,07
3	+ to m.	15,22	13,56	14,39	-	5,26
	m. to +	14,5	13,28	14,54	-	5,55
4	m. to +	3,1	2,95	-	6,54	5,36
	+ to -	6,18	-	7,1	-	5,29
5	- to +	5,92	-	7,3	-	5,58
	+ to -	5,9	-	7,5	-	5,65
6	- to +	5,98	-	7,5	-	5,57
	m. to -	3,03	2,9	-	-	5,53
7	- to +	5,9	5,71	-	-	5,62
	m. to -	3,15	3,03	7,4	6,57	5,29
8	- to +	6,1	5,9	by mean.	-	5,44
	m. to -	3,13	3,00	-	-	5,34
9	- to +	5,72	5,54	-	-	5,79
	+ to -	6,32	-	6,58	-	5,1
10	+ to -	6,15	-	6,59	-	5,27
11	+ to -	6,07	-	7,1	-	5,39
12	- to +	6,09	-	7,3	-	5,42
13	- to +	6,12	-	7,6	-	5,47
	+ to -	5,97	-	7,7	-	5,63
14	- to +	6,27	-	7,6	-	5,34
	+ to -	6,13	-	7,6	-	5,46
15	- to +	6,34	-	7,7	-	5,3
16	- to +	6,1	-	7,16	-	5,75
17	- to +	5,78	-	7,2	-	5,68
	+ to -	5,64	-	7,3	-	5,85

[http://en.wikipedia.org/wiki/Cavendish\\_experiment](http://en.wikipedia.org/wiki/Cavendish_experiment): in 1798 Cavendish found that the Earth's density was  $5.448 \pm 0.033$  times that of water (due to a **simple arithmetic error, found in 1821**, the erroneous value  $5.48 \pm 0.038$  appears in his paper).