

Being approximately correct and being precisely wrong

MEASURING THE QUALITY OF A MEASURING INSTRUMENT

Much of our terminology comes from psychology, where the characteristics are more psychophysical than physical, and where there is no physical scale with centimetres, or Kilograms or minutes. Indeed there is a field of psychology devoted to *psychometrics*. *Sadly, the consequences of poor measurements are skipped over in most statistics courses and textbooks.*

There are two aspects, **Reliability** (‘reproducibility’ / ‘precision’, the degree to which values of the same (unchanged) object would stay the same, and thus the relative positions of different objects would remain unchanged, if the measurement were repeated) and **Validity** (the extent to which it measures what it is intended to measure).

When asked which they would study first, most jump immediately to validity. Thus, in the following case, described in a letter by Dr Fowler to the Editor of JAMA [248;831, 1982] they want to know how the predictions turned out.

The “Drano Test”: [Drano is a gel, sold by Johnson and Johnson, “formulated thick enough to easily pour through water straight to the clog[ged drain], dissolving it fast.”]

During the past several years, we have been asked frequently to do the “Drano test” to determine the sex of an unborn baby. It has been published in the lay press that this is a reliable means of sex determination. A Medline search failed to reveal anything in the medical literature concerning the Drano test. As a result, we performed the test in 100 consecutive pregnant women, checking monthly during the last trimester. The test was done by adding a small amount of crystal Drano to approximately 2 mL of urine, agitating, and interpreting results in one minute’s time. Reportedly, the color green indicates a male baby, and yellow to amber indicates a female.

But Dr Fowler knew that *consistency (repeatability) is a prerequisite for accuracy (validity)*

Of the 100 patients, 21 failed to have the same color change consistently.

Of the babies born to these 21, eleven were girls and ten were boys. Of the remaining 79, we were right in sex determination of 37; of these, there were 20 girls and 17 boys. We were wrong in 42 predictions; of these, there were 22 girls and 20 boys.

		P r e d i c t i o n			
		Inconsistent	Girl	Boy	TOTAL
“TRUTH”	GIRL	11	20	22	53
	BOY	<u>10</u> 21	<u>20</u> 40	<u>17</u> 39	<u>47</u> 100

From this brief study, it would appear that the Drano test for ante-natal sex determination is roughly equivalent to flipping a coin.

JH has also watched engineers test a copying machine. They first test was to see if the machine gave the same copy of a single object. The next was how faithful it was to the original. **So we will start with reliability.**

The notes on page 2-18 are from lectures JH gave in a course on measurement for the rehabilitation sciences. Since few of these students had had experience with Anova’s (widely used to estimate reliability coefficients) it began with a general introduction to these (pp 2-6), showing how the Anova table – usually used for F tests - can also be used to *estimate* variance components. This use of them (and the Method of Moments) may be new to you too.

Page 7 is a nice introduction to reliability & validity: the ideas are general, not confined to examinations, but to any measurement method/instrument. Note the 2 reliability measures, one (an SD) measured in the same scale (scores, cm, Kg, ...) that the instrument uses, the other as a (more abstract) fraction of overall variance. Psychometricians prefer the latter because – unlike in physical measurement – their scales are arbitrary and the object being measured is not a phenomenon, but more a noumenon (abstract).

Pp. 8-16 used Anova tables to estimate the variance components used as inputs to the reliability coefficients (**Intra-Class Correlations, p.10**). Pages 17-18 are ways to assess validity. [NB the different ways terms like accuracy, reliability, validity are used in statistics, and outside statistics].

Pp. 19-21: the **Effects of Measurement Error & role of the ICC.**

RELIABILITY (Reproducibility, Precision): The extent to which obtain the same answer/value/score if object/subject is measured repeatedly under similar situations

Some ways to quantify Reliability

- For one subject: average variation of individual measurements around their mean... either the square root of the average of squared deviations, i.e. standard deviation (SD); or the average absolute deviation, which will usually be quite close to the SD. Could also use range or other measures such as Inter Quartile Range.
- For one subject: average variation or SD as % of the mean of the measurements for that subject... called the within-subject Coefficient of Variation (CV) if calculate it as $[SD/Mean] \times 100$.
- For several subjects: : average the the CV's calculated for the different subjects; if CV's are highly variable, may want to give some sense of this using the range or other measure of spread of the CV's.

Unfortunately, the CV gives no sense of how well the measurements of different subjects (ss) segregate from each other

How about • SD of within-ss measurements or • SD of between-ss measurements (see last item below)?

- Correlation (Pearson or Spearman) if 2 assessments of each ss. ??
- Using correlation between scores on random halves of a test, can estimate how 'reproducible' the full test is (helpful if cannot repeat the test)
- If the measurement in question concerns a population (.e.g., the percentage of smokers among Canadian adults) and if it is measured (estimated) using a *statistic*: e.g. the proportion in a random sample of 1000 adults, it is possible from statistical laws concerning averages to quantify the reliability of the statistic without having to actually perform repeated measurements (samples). For simple random sampling, the formula

$$SE[mean] = \frac{SD[individuals]}{\text{number of individuals measured}}$$

allows us to quantify the reliability indirectly. If we didn't know this formula, we could also arrive at an answer by various re-sampling methods applied to the individuals in the sample at hand – again without resorting to observing any additional individuals.

- Some function of Variance of Within-ss measurements and Variance of Between-ss values? ? Estimate these **Components of Variance** using Analysis of Variance (**Anova**)

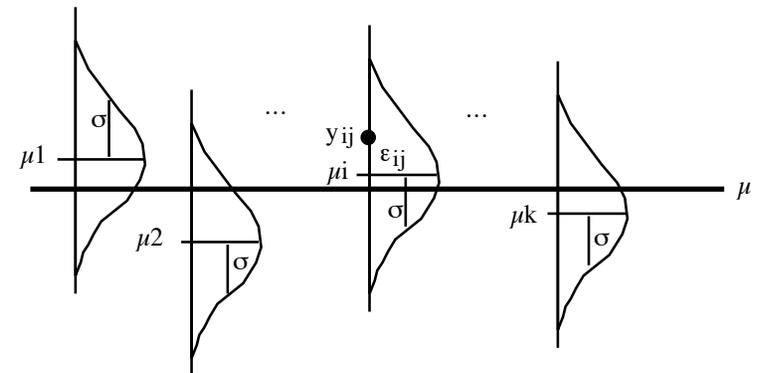
First, a **General Orientation to Anova** and its **primary use** use, testing differences between μ 's of k (2) different groups.

E.g. 1-way ANOVA:

DATA:

	Group					
	1	2	.	i	.	k
Subject						
1	y ₁₁
2
.
j	.	.	.	y _{ij}	.	.
.
n	y _{kn}
Mean	\bar{y}_1	\bar{y}_2	.	\bar{y}_i	.	\bar{y}_k
Variance	s^2_1	s^2_2	.	.	.	s^2_k

MODEL



σ refers to the variation (SD) of all possible individuals in a group; It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij}$$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\sum \sum (\bar{y}_{ij} - \bar{y})^2 = \sum \sum (\bar{y}_i - \bar{y})^2 + \sum \sum (\bar{y}_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Groups Sum of Squares + WITHIN Group Sum of Squares

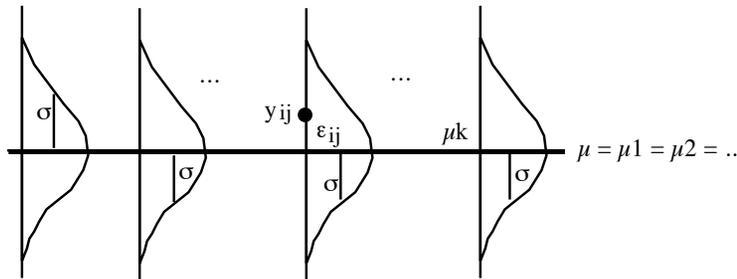
ANOVA TABLE

SOURCE	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	P-Value
BETWEEN	xx.x	k-1	xx.x	$\frac{MS_{BETWEEN}}{MS_{WITHIN}}$ (= SS /df)	0.xx
WITHIN	xx.x	k(n-1)	xx.x		

LOGIC FOR F-TEST (Ratio of variances) as a test of

$H_0: \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_k$

UNDER H0



Means, based on samples of n, should vary around μ with a variance of $\frac{\sigma^2}{n}$

Thus, if H_0 is true, and we calculate the empirical variance of the k different \bar{y}_i 's, it should give us an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{\sum [\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of $\frac{\sigma^2}{n}$

i.e. $\frac{n \sum [\bar{y}_i - \bar{y}]^2}{k-1}$ is an unbiased estimate of σ^2

i.e. $\frac{\sum \sum [\bar{y}_{ij} - \bar{y}]^2}{k-1} = MS_{BETWEEN}$ is an unbiased estimate of σ^2

Whether or not H_0 is true, the empirical variance of the n (within-group) values

y_{i1} to y_{in} i.e. $\frac{\sum [\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ should give us an unbiased estimate of σ^2

i.e. $s^2_i = \frac{\sum [\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$ is an unbiased estimate of σ^2

so the average of the k different estimates,

$$\frac{1}{k} \sum s^2_i = \frac{1}{k} \sum \frac{\sum [\bar{y}_{ij} - \bar{y}_i]^2}{n-1}$$

is also an unbiased estimate of σ^2

i.e. $\frac{\sum \sum [\bar{y}_{ij} - \bar{y}_i]^2}{k[n-1]} = MS_{WITHIN}$ is an unbiased estimate of σ^2

THUS, under H_0 , both $MS_{BETWEEN}$ and MS_{WITHIN} are unbiased estimates of estimates of σ^2 and so their ratio should, apart from sampling variability, be 1. IF however, H_0 is not true, $MS_{BETWEEN}$ will tend to be larger than MS_{WITHIN} , since it contains an extra contribution that is proportional to how far the μ 's are from each other.

In this "non-null" case, the $MS_{BETWEEN}$ is an unbiased estimate of

$$\sigma^2 + \frac{\sum n[\mu_i - \bar{\mu}]^2}{k-1}$$

and so we expect that, apart from sampling variability, the ratio $\frac{MS_{BETWEEN}}{MS_{WITHIN}}$ should be greater than 1. The tabulated values of the F distribution (tabulated under the assumption that the numerator and denominator of the ratio are both estimates of the same quantity) can thus be used to assess how extreme the observed F ratio is and to assess the evidence against the H_0 that the μ 's are equal.

How ANOVA can be used to estimate Components of Variance used in quantifying Reliability.

The basic ANOVA calculations are the same, but the MODEL underlying them is different. First, in the more common use of ANOVA just described, the groups can be thought of as all the levels of the factor of interest. The number of levels is necessarily finite. The groups might be the two genders, all of the age groups, the 4 blood groups, etc. Moreover, when you publish the results, you explicitly identify the groups.

When we come to study subjects, and ask "How big is the intra-subject variation compared with the inter-subject variation, we will for budget reasons only study a sample of all the possible subjects of interest. We can still number them 1 to k, and we can make n measurements on each subject, so the basic layout of the data doesn't change. All we do is replace the word 'Group' by 'Subject' and speak of BETWEEN-SUBJECT and WITHIN-SUBJECT variation. So the data layout is...

DATA:

Measurement	Subject				
	1	2	i	...	k
1	y ₁₁
2
...
j	.	.	y _{ij}	.	.
...
n	y _{kn}
Mean	\bar{y}_1	\bar{y}_2	\bar{y}_i	...	\bar{y}_k
Variance	s ² ₁	s ² ₂	s ² _k

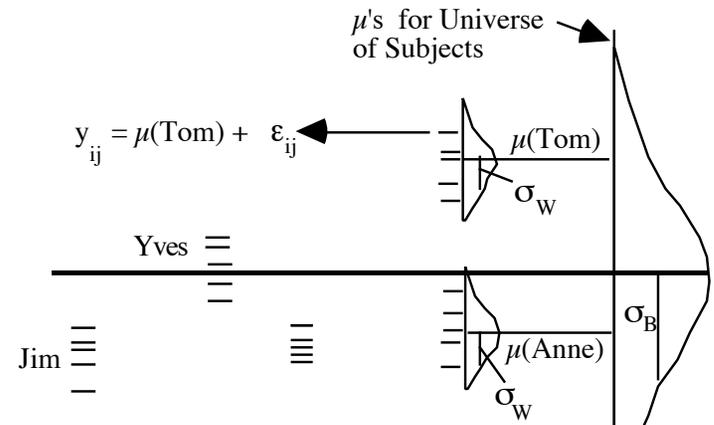
MODEL

The model is different. There is no interest in the specific subjects. Unlike the critical labels "male" and "female", or "smokers", "nonsmokers" and "exsmokers" to identify groups of interest, we certainly are not going to identify subjects as Yves, Claire, Jean, Anne, Tom, Jim, and Harry in the publication, and nobody would be fussed if in the dataset we used arbitrary subject identifiers to keep track of which measurements were made on whom. we wouldn't even care if the research assistant lost the identities of the subjects -- as long as we know that the correct measurements go with the correct subject!

The "Random Effects" Model uses 2 stages:

- (1) random sample of subjects, each with his/her own μ
- (2) For each subject, series of random variations around his/her μ

Notice the diagram has considerable 'segregation' of the measurements on different individuals. There is no point in TESTING for (inter-subject) differences in the μ 's. The task is rather to estimate the relative magnitudes of the two variance components σ^2_B and σ^2_W .



σ_B refers to the SD of the universe of μ 's ; It is an unknowable parameter and can only be ESTIMATED

σ_W refers to the variation (SD) of all possible measurements on a subject It is an (unknowable) parameter; it can only be ESTIMATED.

Or, in symbols...

$$y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$\alpha_i \sim N(0, \sigma^2_B)$$

$$\epsilon_i \sim N(0, \sigma^2_W)$$

DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION

$$\sum\sum(\bar{y}_{ij} - \bar{y})^2 = \sum\sum(\bar{y}_i - \bar{y})^2 + \sum\sum(\bar{y}_{ij} - \bar{y}_i)^2$$

TOTAL Sum of Squares = BETWEEN Subjects + Sum of Squares + WITHIN Subjects + Sum of Squares

ANOVA TABLE (Note absence of F and P-value Columns)

SOURCE	Sum of Squares	Degrees of Freedom	Mean Square	What the Mean Square is an estimate of*
SOURCE	SS	df	MS (= SS /df)	
BETWEEN Subjects	xx.x	k-1	xx.x	$\sigma^2_W + n \sigma^2_B$
WITHIN Subjects	xx.x	k(n-1)	xx.x	σ^2_W

ACTUAL ESTIMATION OF 2 Variance Components

MS_{BETWEEN} is an unbiased estimate of $\sigma^2_W + n \sigma^2_B$

MS_{WITHIN} is an unbiased estimate of σ^2_W

By subtraction...

MS_{BETWEEN} - MS_{WITHIN} is an unbiased estimate of $n \sigma^2_B$

$\frac{\mathbf{MS_{BETWEEN} - MS_{WITHIN}}}{\mathbf{n}}$ is an unbiased estimate of σ^2_B

This is the **definitional** formula; the **computational** formula may be different.

* Pardon my ending with a preposition, but I find it difficult to say otherwise. These parameter combinations are also called the "Expected Mean Squares". They are the long-run expectations of the MS statistics. As Winston Churchill would say, "For the sake of clarity, this one time this wording is something up which you would put".

Example....

DATA:

Measurement	Subject					Variance
	Tom	Anne	Yves	Jean	Claire	
1	4.8	5.5	5.1	6.4	5.8	4.5
2	4.7	5.2	4.9	6.2	6.3	4.1
3	4.9	5.2	5.3	6.6	5.6	4.0
Mean	4.8	5.3	5.1	6.4	5.9	4.2
Variance	0.01	0.03	0.04	0.04	0.13	0.07

Variance = 0.614

ANOVA TABLE (Check... I did it by hand!)

SOURCE	Sum of Squares	Degrees of Freedom	Mean Square	What the Mean Square is an estimate of... *
SOURCE	SS	df	MS (= SS /df)	
BETWEEN Subjects	9.205	5	1.841	$\sigma^2_W + n \sigma^2_B$
WITHIN Subjects	0.640	12	0.053	σ^2_W
TOTAL	9.845	17		

ESTIMATES OF VARIANCE COMPONENTS

MS_{WITHIN} = **0.053** is an unbiased estimate of σ^2_W

$\frac{\mathbf{1.841 - 0.053}}{\mathbf{3}}$ = **0.596** is an unbiased estimate of σ^2_B

1-Way ANOVA Calculations performed by SAS; Components estimated manually

```
PROC GLM in SAS ==> estimating components 'by hand'
DATA a; INPUT Subject Value; LINES;
1 4.8
1 4.7
...
6 4.5
proc glm; class subject; model value=subject / ss3;
random subject ;
```

See worked example using earsize data.

If unequal numbers of measurements per subject, see formula in A&B or Fleiss

Estimating Components of Variance using "Black Box"

PROC VARCOMP; class subject ; model Value = Subject ;
See worked example following...

2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64

subject	5				6				7				8			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

INTRA-OBSERVER VARIATION (e.g. observer #1)

e.g. observer #1

PROC GLM in SAS ==> estimating components 'by hand'

INPUT subject rater occasion earsize; if observer=1;
The data set has 16 obsns & 4 variables.

proc glm; class subject; model earsize=subject / ss3;
random subject ;

General Linear Models Procedure: Class Level Information

Class Levels Values
SUBJECT 8 1 2 3 4 5 6 7 8 ; # of obsns. in data set = 16

Dependent Variable: EARSIZE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	341.00	48.71	35.43	0.0001
Error	8	11.00	1.38		
Corrected Total	15	352.00			

R-Square C.V. Root MSE EARSIZE Mean
0.968750 1.80 1.17260 65.0

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SUBJECT	7	341.00	48.71	35.43	0.0001

Source Type III Expected Mean Square
SUBJECT Var(Error) + 2 Var(SUBJECT)

$$\begin{aligned} \text{Var(Error)} + 2 \text{Var(SUBJECT)} &= 48.71 \\ \text{Var(Error)} &= 1.38 \\ 2 \text{Var(SUBJECT)} &= 47.33 \\ \text{Var(SUBJECT)} &= 47.33 / 2 = 23.67 \end{aligned}$$

Estimating Variance components using PROC VARCOMP in SAS

proc varcomp; class subject ; model earsize = subject ;

Variance Components Estimation Procedure: Class Level Information

Class Levels Values

SUBJECT 8 1 2 3 4 5 6 7 8 ; # obsns in data set = 16

MIVQUE(0) Variance Component Estimation Procedure

Variance Component	Estimate
EARSIZE	
Var(SUBJECT)	23.67
Var(Error)	1.38

• ICC (Fleiss § 1.3)

$$\text{ICC} = \frac{\text{Var(SUBJECT)}}{\text{Var(SUBJECT)} + \text{Var(Error)}} = \frac{23.67}{23.67 + 1.38} = 0.94$$

1-sided 95% Confidence Interval (see Fleiss p 12)

df for F in CI: (8-1)= 7 and 8

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{35.43 - 3.5}{35.43 + (2 - 1) \cdot 3.5} = 0.82$$

EXERCISE: Carry out the estimation procedure for one of the other 3 observers.

INTERPRETING YOUR GRE SCORES

(Blurb from Educational Testing Service)

Your test score is an estimate, not a complete and perfect measure, of your knowledge and ability in the area tested. In fact, if you had taken a different edition of the test that contained different questions but covered the same content, it is likely that your score would have been slightly different. The only way to obtain perfect assessment of your knowledge and ability in the area tested would be for you to take all possible test editions that could ever be constructed. Then assuming that your ability and knowledge did not change, the average score on all those editions, referred to as your "true score," would be a perfect measure of your knowledge and ability in the content areas covered by the test. Therefore, scores are estimates and not perfect measures of a person's knowledge and ability. Statistical indices that address the imprecision of scores in terms of standard error of measurement and reliability are discussed in the next two sections.

STANDARD ERROR OF MEASUREMENT

The difference between a person's true and obtained scores is referred to as "error of measurement."* The error of measurement for an individual person cannot be known because a person's true score can never be known. The average size of these errors, however, can be estimated for a group of examinees by the statistic called the "standard error of measurement for individual scores." The standard error of measurement for individual scores is expressed in score points. About 95 percent of examinees will have test scores that fall within two standard errors of measurement of their true scores. For example, the standard error of measurement of the GRE Psychology Test is about 23 points. Therefore, about 95 percent of examinees obtain scores in Psychology that are within 46 points of their true scores. About 5 percent of examinees, however, obtain scores that are more than 46 points higher or lower than their true scores.

Errors of measurement also affect any comparison of the scores of two examinees. Small differences in scores may be due to measurement error and not to true differences in the abilities of the examinees. The statistic "standard error of measurement of score differences" incorporates the error of measurement in each examinee's score being compared. This statistic is about 1.4 times as large as the standard error of measurement for the individual scores themselves. Approximately 95 percent of the differences between the obtained scores of examinees who have the same true score will be less than two times the standard error of measurement of score differences. Fine distinctions should not be made when comparing the scores of two or more examinees.

RELIABILITY

The reliability of a test is an estimate of the degree to which the relative position of examinees' scores would change if the test had been administered under somewhat different conditions (for example, examinees were tested with a different test edition).

Reliability is represented by a statistical coefficient that is affected by errors of measurement. Generally, the smaller the errors of measurement in a test, the higher the reliability. Reliability coefficients may range from 0 to 1, with 1 indicating a perfectly reliable test (i.e., no measurement error) and zero reliability indicating a test that yields completely inconsistent scores. Statistical methods are used to estimate the reliability of the test from the data provided by a single test administration. Average reliabilities of the three scores on the General Test and of the total scores on the Subject Tests range from .88 to .96 on recent editions. Average reliabilities of subscores on recent editions of the Subject Test range from .82 to .90.

Data regarding standard errors of measurement and reliability of individual GRE tests may be found in the leaflet *Interpreting Your GRE General and Subject Test Scores*, which will be sent to you with your GRE Report of Scores.

VALIDITY

The validity of a test—the extent to which it measures what it is intended to measure—can be assessed in several ways. One way of addressing validity is to delineate the relevant skills and areas of knowledge for a test, and then, when building each edition of the test, make sure items are included for each area. This is usually referred to as content validity. A committee of ETS specialists defines the content of the General Test, which measures the content skills needed for graduate study. For Subject Tests, ETS specialists work with professors in that subject to define test content. In the assessment of content validity, content representativeness studies are performed to ensure that relevant content is covered by items in the test edition.

Another way to evaluate the validity of a test is to assess how well test scores forecast some criterion, such as success in grade school. This is referred to as predictive validity. Indicators of success in graduate school may include measures such as graduate school grades, attainment of a graduate degree, faculty ratings, and departmental examinations. The most commonly used measure of success in assessing the predictive validity of the GRE tests is graduate first year grade point average. Reports on content representativeness and predictive validity studies of GRE tests may be obtained through the GRE Program office.

* The term "error of measurement" does not mean that someone has made a mistake in constructing or scoring the test. It means only that a test is an imperfect measure of the ability or knowledge being tested.

From the days when GRE's were taken on paper.

Outline

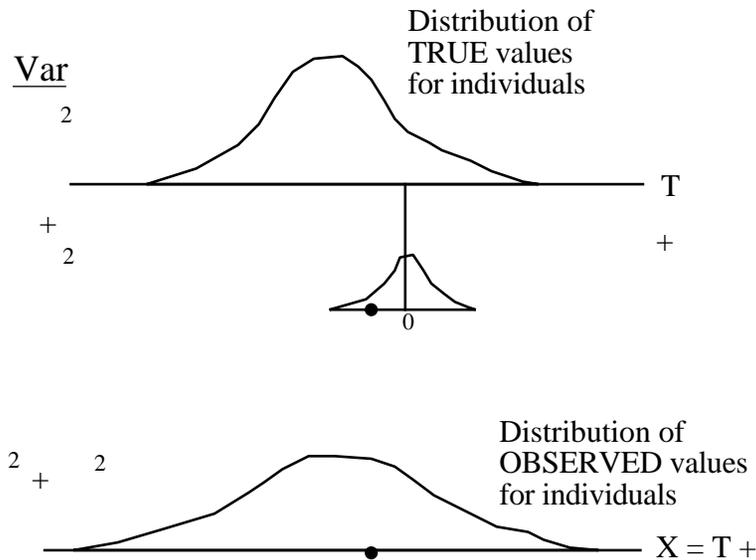
Some ways to quantify Reliability:

- Reliability Coefficient
- Internal Consistency (Cronbach's α)

Implications:

-
.....

Model for Reliability



"True" scores / values not knowable;

Variance calculation assumes that the distribution of errors is independent of T

Reliability Coefficient

$$r_{xx} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

i.e. the fraction of observed variation that is 'real'

Note that one can 'manipulate' r by choosing a large or small σ_T^2

Effect of # of Items on Reliability Coefficient
(if all items have same variance and same intercorrelations)

SCALE 2 N Times more items than SCALE 1

$$r_{SCALE 2} = \frac{N \times r_{SCALE 1}}{1 + [N-1] \times r_{SCALE 1}}$$

e.g.

Scale	# Items	r
1	10	0.4
2	20 (× 2)	0.57
3	30 (× 3)	0.67

Cronbach's α

k items

$$\alpha = \frac{k \times \bar{r}}{1 + [k-1] \times \bar{r}}, \text{ where } \bar{r} = \text{average of inter-item correlations}$$

α is an estimate of the expected correlation of one test with an alternative form with the same number of items.

α is a lower bound for r_{XX} i.e. $r_{XX} \geq \alpha$

$$r_{XX} = \alpha \quad \text{if items are parallel .}$$

parallel

Average [item 1] = Average [item 2] = Average [item 3] = ...

Variance[item 1] = Variance[item 2] = Variance [item 3] = ...

Correlation[item 1, item 2] = Correlation[item 1, item 3] = ...

=Correlation[item 2, item 3] = ...

INTRAClass CORRELATIONS (ICC's)

- Various versions**

TEST-RETEST

INTRA-RATER

INTER-RATER...

- Formed as Ratios of various Variances**

$$\text{e.g. } \frac{\sigma_{\text{TRUE}}^2}{\sigma_{\text{TRUE}}^2 + \sigma_{\text{ERROR}}^2}$$

with estimates of various σ^2 's substituted for the σ^2 's .

Estimates of various components typically derived from ANOVA.

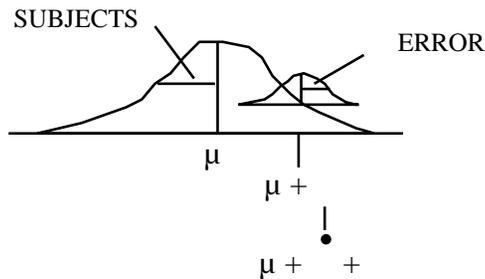
- Note the distinction between DEFINITIONAL FORM (involving PARAMETERS) and COMPUTATIONAL FORM (involving STATISTICS)**

Fleiss Chapter 1 good here; Norman & Streiner not so good!!)

ICC's (Portnoy and Wilkins)

- (1) multiple (unlabeled) measurements of each subject
- (2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.
- (3) as in (2), but these raters studied are the only raters of interest

.....
 (1) multiple (unlabeled) measurements of each subject



$$ICC = \frac{2 \text{ SUBJECTS}}{2 \text{ SUBJECTS} + 2 \text{ ERROR}}$$

Model for observed data:

$$y[\text{subject } i, \text{ measurement } j] = \mu + \epsilon_{ij}$$

EXAMPLE 1

This example is in the spirit of the way the ICC was first used, as a measure of the greater similarity within families than between families: Study by Bouchard (NEJM) on weight gains of 2 members from each of 12 families: It is thought that there will be more variation between members of different families than between members of the same family: family (genes) is thought to be a large source of variation; the two twins per family are thought of as 'replicates' from the family and closer to each other (than to others) in their responses. Here the "between" factor is family i.e. families are the subjects and the two twins in the family are just replicates and they don't need to be labeled (if we did label them 1 and 2, the labels would be arbitrary, since the two twins are thought to be 'interchangeable'. (weight gain in Kg over a summer)

model: weight gain for person j in family i = $\mu + \mu + \epsilon_{ij}$

1-way Anova and Expected Mean Square (EMS)

Source	Sum of Sq	d.f	Mean Square	Expected Mean Square
Between (families)	99	11	9.0	$2 \text{ "error"} + k_0 \text{ }^2_{\text{between}}$
Error(Within families)	30	12	2.5	2 "error"

Total	129	23		

In our example, we measure k=2 members from each family, so k_0 is simply 2

[if the k's are unequal, k_0 is somewhat less than the average k... $k_0 = \text{average } k - (\text{variance of } k\text{'s}) / (n \text{ times average } k)$...see Fleiss page 10]

Estimation of parameters that go to make up ICC

2.5 is an estimate of 2 "error"

9.0 is an estimate of $2 \text{ "error"} + 2 \text{ }^2_{\text{between}}$

6.5 is an estimate of $2 \text{ }^2_{\text{between}}$

$\frac{6.5}{2}$ is an estimate of $^2_{\text{between}}$

$$\frac{\frac{6.5}{2}}{\frac{6.5}{2} + 2.5} = \frac{3.25}{3.25 + 2.5} = 0.57$$

is an estimate of ICC = $\frac{^2_{\text{between}}}{^2_{\text{between}} + ^2_{\text{error}}}$

COMPUTATIONAL Formula for "1-way" ICC

$$\frac{\frac{MS_{\text{between}} - MS_{\text{within}}}{k_0}}{\frac{MS_{\text{between}} - MS_{\text{within}}}{k_0} + MS_{\text{within}}}$$

$$= \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}} + (k_0 - 1)MS_{\text{within}}} \quad [\text{shortcut}]$$

is an estimate of the ICC

Notes:

- Streiner and Norman start on page 109 with the 2-way anova for inter-observer variation. There are mistakes in their depiction of the SSerror on p 110 [it should be $(6-6)^2 + (4-4)^2 + (2-1)^2 + \dots + (8-)^2 = 10$. If one were to do the calculations by hand, one usually calculates the SS_{total} and then obtains the SS_{error} by subtraction]
- They then mention the 1-way case, which we have discussed above, as "the observer nested within subject" on page 112
- Fleiss gives methods for calculating CI's for ICC's.

EXAMPLE 2: INTRA-OBSERVER VARIATION FOR 1 OBSERVER**Computations performed on earlier handout...**

$$\text{Var}(\text{SUBJECT}) = 23.67 \quad \text{Var}(\text{ERROR}) = 1.38$$

$$\hat{ICC} = 23.67 / (23.67 + 1.38) = 0.94$$

An estimated 94% of observed variation in earsize measurements by this observer is 'real' .. i.e. reflects true between-subject variability.

Note that I say 'an estimated 94% ...'. I do this because the 94% is a statistic that is subject to sampling variability (94% is just a point estimate or a 0% Confidence Interval). An interval estimate is given by say a 95% confidence interval for the true ICC (lower bound of a 1-sided CI is 82% ... see previous handout)

Increasing Reliability by averaging several measurements

In 1-way model: $y_{ij} = \mu + \alpha_i + e_{ij}$

where $\text{var}[\alpha_i] = \sigma^2_{\text{between subjects}}$; $\text{var}[e_{ij}] = \sigma^2_{\text{error}}$

Then if we average k measurements, i.e.,

$$\bar{y}_i = \mu + \alpha_i + \bar{e}_i$$

then

$$\text{Var}_i[\bar{y}_i] = \sigma^2_{\text{between}} + \frac{\sigma^2_{\text{error}}}{k}$$

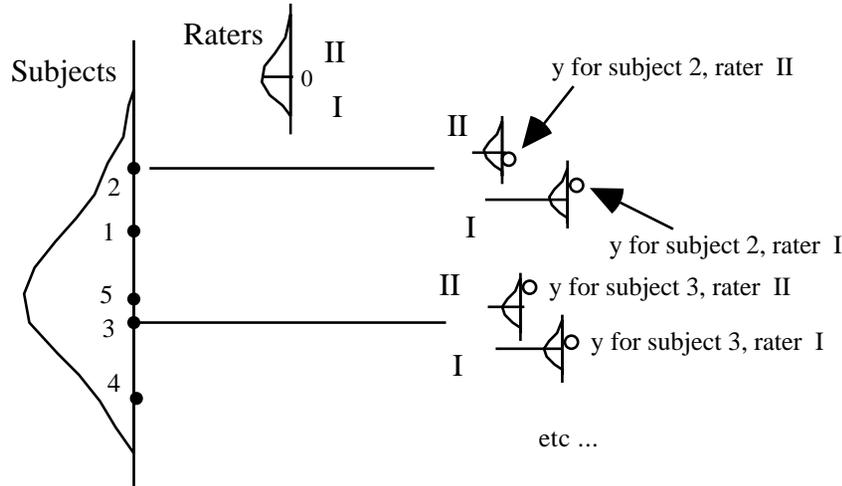
$$\text{So ICC}[k] = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \frac{\sigma^2_{\text{error}}}{k}}$$

This is called **"Stepped-Up" Reliability**.

ICC's (Portnoy and Wilkins).

(2) same set of raters measure each subject; raters thought of as a random sample of all possible raters.

• Model



$$\mu + [subject] + [rater] + error$$

2 subjects 2 raters 2 error

• From 2- way data layout (subjects x Raters)

estimate 2_{subjects} , 2_{raters} and 2_{error} by 2-way ANOVA

• Substitute variance estimates in appropriate ICC form

e.g. 2 measurements (in mm) of earsize of 8 subjects by each of 4 observers

subject	1				2				3				4			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	67	65	65	64	74	74	74	72	67	68	66	65	65	65	65	65
2nd	67	66	66	66	74	73	71	73	68	67	68	67	64	65	65	64
subject	5				6				7				8			
obsr	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1st	65	62	62	61	59	56	55	53	60	62	60	59	66	65	65	63
2nd	61	62	60	61	57	57	57	53	60	65	60	58	66	65	65	65

ESTIMATING INTER-OBSERVER VARIATION from occasion=1;

```
PROC GLM in SAS ==> estimating components 'by hand'
INPUT subject rater occasion earsize; if occasion=1; (32 obsns)
```

```
proc glm; class subject rater; model earsize=subject rater / ss3;
random subject rater;
```

General Linear Models Procedure: Class Level Information

Class	Levels	Values
SUBJECT	8	1 2 3 4 5 6 7 8
RATER	4	1 2 3 4

Number of observations in data set = 32

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	764.500	76.45	78.80	0.0001
Error	21	20.375	0.97		
Corrected Total	31	784.875			

R-Square	C.V.	Root MSE	EARSIZE Mean
0.974040	1.534577	0.98501	64.1875

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SUBJECT	7	734.875000	104.98	108.20	0.0001
RATER	3	29.625000	9.87	10.18	0.0002

Source Type III Expected Mean Square

SUBJECT	Var(Error) + 4 Var(SUBJECT)
RATER	Var(Error) + 8 Var(RATER)

So... solving 'by hand' for the 3 components...

$$\begin{aligned} \text{Var(Error)} + 4 \text{Var(SUBJECT)} &= 104.98 \\ \text{Var(Error)} &= 0.97 \\ \implies 4 \text{Var(SUBJECT)} &= 104.01 \\ \implies \text{Var(SUBJECT)} &= 104.01 / 4 = 26.00 \end{aligned}$$

$$\begin{aligned} \text{Var(Error)} + 8 \text{Var(RATER)} &= 9.87 \\ \text{Var(Error)} &= 0.97 \\ \implies 8 \text{Var(RATER)} &= 8.90 \\ \implies \text{Var(RATER)} &= 8.90 / 8 = 1.11 \end{aligned}$$

$$\text{Var(Error)} = 0.97$$

Estimating Variance components using PROC VARCOMP in SAS

```
proc varcomp; class subject rater; model earsize = subject rater;
```

Variance Component	Estimate EARSIZE
Var(SUBJECT)	26.00
Var(RATER)	1.11
Var(Error)	0.97

• **ICC: "Raters Random"** (Fleiss § 1.5.2)

$$\text{ICC} = \frac{\text{Var}(\text{SUBJECT})}{\text{Var}(\text{SUBJECT}) + \text{Var}(\text{RATER}) + \text{Var}(\text{Error})} = \frac{26.00}{26.00 + 1.11 + 0.97} = 0.93$$

1-sided 95% Confidence Interval (see Fleiss p 27)

df for F in CI: (8-1)= 7 and v* , where

$$v^* = \frac{(8-1)(4-1)(4 \cdot 0.93 \cdot 10.18 + 8[1+(4-1) \cdot 0.93] - 4 \cdot 0.93)^2}{(8-1) \cdot 4^2 \cdot 0.93^2 \cdot 10.18^2 + (8[1+(4-1) \cdot 0.93] - 4 \cdot 0.93)^2} = 8.12$$

so from Tables of F distribution with 7 & 8 df, F = 3.5

So lower limit of CI for ICC is

$$= \frac{8(104.98 - 3.5 \cdot 0.97)}{8 \cdot 104.98 + 3.5 \cdot [4 \cdot 9.87 + (8 \cdot 4 - 8 - 4) \cdot 0.97]} = 0.78$$

• **ICC: if use one "fixed" observer** (see Fleiss p 23, strategy 3)

$$\text{ICC} = \frac{\text{Var}(\text{SUBJECT})}{\text{Var}(\text{SUBJECT}) + \text{Var}(\text{Error})} = \frac{26.00}{26.00 + 0.97} = 0.96$$

lower limit of 95% 1-sided CI (eqn 1.49: F = 2.5 ; 7 & 7x3=21 df)

$$\text{ICC} = \frac{104.98 - 2.5}{104.98 + (4-1) \cdot 2.5} = 0.91$$

USING ALL THE DATA SIMULTANEOUSLY

(can now estimate subject x Rater interaction .. i.e extent to which raters 'reverse themselves' with different subjects)

Components of variance when use **both** measurements (all 64 obsns)

```
proc varcomp;                                proc varcomp;
  class subject rater;                        class subject rater;
  model earsize = subject rater;             model earsize = subject rater
                                              subject*rater;
```

Variance Component	Estimate EARSIZE	Variance Component	EARSIZE
Var(SUBJECT)	25.52	Var(SUBJECT)	25.47
Var(RATER)	0.70	Var(RATER)	0.67
Var(Error)	1.37	Var(SUBJECT*RATER)	0.31
		Var(Error)	1.13

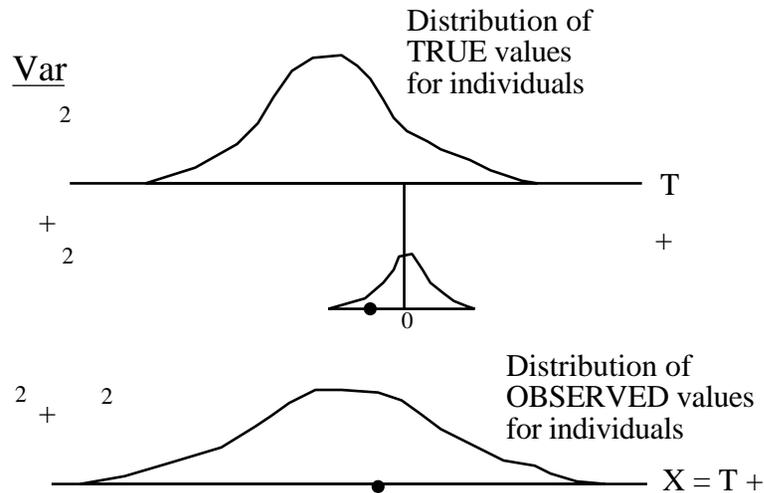
LINK between STANDARD ERROR OF MEASUREMENT and RELIABILITY COEFFICIENT

Example: GRE Tests (cf blurb from Educational Testing Service)

Standard Error of Measurement = 23 points

Reliability Coefficient: R = 0.93

recall...



$$\sigma_e = 23 \implies \sigma_e^2 = 529 ;$$

$$R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} = 0.93 \implies \sigma_T^2 = \frac{R \times \sigma_e^2}{1 - R} = \frac{0.93 \times 529}{1 - 0.93} = 7028$$

$$\sigma_T^2 + \sigma_e^2 = 7028 + 529 = 7557 \implies \sqrt{\sigma_T^2 + \sigma_e^2} = \sqrt{7557} = 87$$

So if 3 SD's on either side of the mean of 500 covers most of the observed scores, this would give a range of observed scores of $500 - 261 = 239$ to $500 + 261 = 761$.

Another way to say it (see Streiner and Norman, bottom of page 119) :-

$$\sigma_e = \sqrt{\sigma_T^2 + \sigma_e^2} \times \sqrt{1 - R} = \text{SD}[\text{observed scores}] \times \sqrt{1 - R}$$

Confidence Intervals / Sample Sizes for ICC's

see Fleiss...

CI's based on F distribution tables;

CI's not symmetric;

More interested in 1-sided CI's i.e. (lower bound, 1) i.e. ICC 0.xx;

See also Donner and Eliasziw.

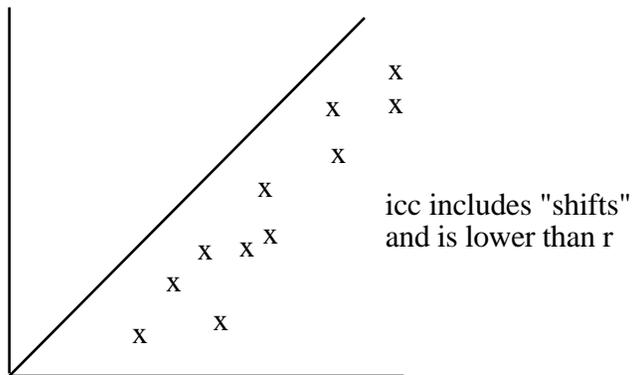
NOTE: If interested in ICC that incorporates random raters, then sample size must involve both # of raters and # of subjects

CI will be very wide if use only 2 or 3 raters

Approach sample size as "n's or raters and subjects needed for a sufficiently narrow CI.

Why Pearson's r is not always a good [or practical] measure of reproducibility

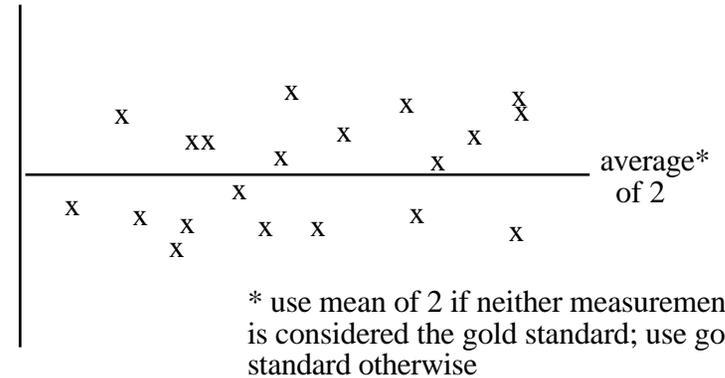
1. It does not pick up "shifts"



2. not practical if > 2 measurements or variable # of measurements per subject
ICC 'made for' such situations

Method of Bland & Altman [Lancet]

Difference of 2 measurements



- +++ see biases quickly
- can explain to your in-laws (can you explain ICC to them?)
- emphasises errors in measurements scale itself (like ± 23 in GRE score)
- if don't know real range, magnitudes of standard error of measurement not helpful (see Norman & Streiner)
- cannot use with > 2 measurements
- doesn't generalize to raters

Assessing reproducibility of measurements made on a CATEGORICAL scale

Categorizations of n subjects by RATER 2

		C1	C2	C3
Categorizations of subjects by RATER 1	C1			
	C2			
	C3			
				n

See chapter 13 in Fleiss's book on Rates and Proportions
or pp 516-523 of Chapter 26 of Portnoy and Wilkins

- Simple Measure

$$\% \text{ agreement} = \frac{\# \text{ in diagonal cells}}{n} \times 100$$

- Chance-Corrected Measure

$$= \frac{\% \text{ agreement} - \% \text{ agreement expected by chance}^*}{100\% \text{ agreement} - \% \text{ agreement expected by chance}}$$

* expected proportion = $p[\text{row}] * p[\text{col}]$ --- over the diagonals

(see Aickin's arguments against 'logic' of chance-correction:
Biometrics 199)

can give weights for 'partial' agreement

if > 2 raters, use range or average of pairwise kappas

with quadratic weights, weighted kappa = icc

Validity Statistics

<u>INSTRUMENT</u>	<u>CRITERION</u>	<u>STATISTIC(S)</u>
numerical	numerical - same scale	<ul style="list-style-type: none"> • calculate discrepancies (Altman & Bland) • describe distribution of discrepancies
numerical	numerical - different scale	<ul style="list-style-type: none"> • correlation [Pearson or Spearman]
numerical	ordered categories (e.g. known groups)	<ul style="list-style-type: none"> • correlation [ranks] • ANOVA [groups]
ordered categories	ordered categories	<ul style="list-style-type: none"> • Kendall's "tau"
numerical or ordered categories	binary	<ul style="list-style-type: none"> • difference in means [parametric/ nonparametric test] • ROC curve
binary	binary	<ul style="list-style-type: none"> • sensitivity (Probability + on instr. if Criterion +) & specificity (Probability – on instr. if Criterion –) • predictive values • (phi) statistic (see Streiner & Norman)

Effect of Measurement Errors in X and Y on measured correlation and slope

Denote the “true” (but unobservable) values by X and Y and the observed (error-containing) measurements by X' and Y' . We quantify the degree to which the errors in X' and Y' distort the correlation $\rho_{X,Y}$ and the slope $\beta_{Y/X}$

From the general formulae

$$\rho_{X,Y} = \frac{E[XY] - E[X] \times E[Y]}{SD[X] \times SD[Y]} = \frac{Covar[X,Y]}{\sqrt{Var[X] \times Var[Y]}} \quad (1)$$

and

$$\beta_{Y/X} = \frac{E[XY] - E[X]E[Y]}{VAR[X]} = \rho_{X,Y} \frac{SD[Y]}{SD[X]}, \quad (2)$$

we can derive the consequences of the errors in X and in Y .

Let $X' = X + \epsilon_X$ where ϵ_X has mean 0 and variance $Var[\epsilon_X]$, and is independent of X , so that

$$E[X'] = E[X + \epsilon_X] = E[X] + E[\epsilon_X] = E[X] + 0 = E[X] \quad (3)$$

$$Var[X'] = Var[X + \epsilon_X] = Var[X] + Var[\epsilon_X] \quad (4)$$

Let $Y' = Y + \epsilon_Y$, where ϵ_Y has mean 0 and variance $Var[\epsilon_Y]$, and is independent of Y , so that

$$E[Y'] = E[Y + \epsilon_Y] = E[Y] + E[\epsilon_Y] = E[Y] + 0 = E[Y] \quad (5)$$

$$Var[Y'] = Var[Y + \epsilon_Y] = Var[Y] + Var[\epsilon_Y] \quad (6)$$

$$E[X'Y'] = E[\{X + \epsilon_X\}\{Y + \epsilon_Y\}] = E[XY + \epsilon_X Y + \epsilon_Y X + \epsilon_X \epsilon_Y] = E[XY]. \quad (7)$$

By definition

$$ICC[X] = \frac{Var[X]}{Var[X] + Var[\epsilon_X]} \quad \& \quad ICC[Y] = \frac{Var[Y]}{Var[Y] + Var[\epsilon_Y]}, \quad (8)$$

with

$$0 \leq ICC[X] \leq 1 \quad \& \quad 0 \leq ICC[Y] \leq 1. \quad (9)$$

1 $\rho_{X',Y'}$: expected correlation of two error-containing variables

$$\begin{aligned}\rho_{X',Y'} &= \frac{E[X'Y'] - E[X'] \times E[Y']}{SD[X'] \times SD[Y']} \\ &= \frac{E[XY] - E[X] \times E[Y]}{\sqrt{Var[X'] \times Var[Y']}} \\ &= \frac{Covar[X, Y]}{\sqrt{\{Var[X'] + Var[\epsilon_X]\} \times \{Var[Y'] + Var[\epsilon_Y]\}}}\end{aligned}$$

dividing above and below by $\sqrt{Var[X] \times Var[Y]}$

$$\begin{aligned}&= \frac{\frac{Cov[X, Y]}{\sqrt{Var[X] \times Var[Y]}}}{\sqrt{\frac{Var[X'] + Var[\epsilon_X]}{Var[X]} \times \frac{Var[Y'] + Var[\epsilon_Y]}{Var[Y]}}} \\ &= \frac{\rho_{X, Y}}{\sqrt{\frac{1}{ICC[X]} \times \frac{1}{ICC[Y]}}}\end{aligned}$$

so that...

$$\rho_{X',Y'} = \sqrt{ICC[X]} \times \sqrt{ICC[Y]} \times \rho_{X, Y} \leq \rho_{X, Y}.$$

Thus, the correlation is **attenuated*** (dampened/weakened) by the imperfections (random errors) in the X and Y measurements.

One can reverse the equation to get a “**de-attenuated**” correlation:

$$\rho_{X, Y} = \frac{\beta_{X', Y'}}{\sqrt{ICC[X]} \times \sqrt{ICC[Y]}}$$

<http://www.m-w.com/dictionary/attenuate>

*Main Entry: *attenuate* ; Function: adjective

Etymology: Middle English *attenuat*, from Latin *attenuatus*, past participle of *attenuare* to make thin, from *ad-* + *tenuis* thin

1 : reduced especially in thickness, density, or force

2 : tapering gradually usually to a long slender point

2 $\beta_{Y'/X'}$: expected slope of error-containing Y on error-containing X

$$\begin{aligned}\beta_{Y'/X'} &= \frac{E[X'Y'] - E[X'] \times E[Y']}{Var[X']} \\ &= \frac{E[XY] - E[X] \times E[Y]}{Var[X'] + Var[\epsilon_X]} \\ &= \frac{Covar[X, Y]}{Var[X'] + Var[\epsilon_X]} \\ &= \frac{Covar[X, Y]}{Var[X]} \times \frac{Var[X]}{Var[X'] + Var[\epsilon_X]} \\ &= \beta_{Y/X} \times ICC[X].\end{aligned}$$

so that...

$$\beta_{Y'/X'} = \beta_{Y/X} \times ICC[X] \leq \beta_{Y/X}$$

i.e., the slope is **attenuated** (dampened / weakened / flattened / moved towards 0) by the imperfections in the X measurements. Random errors in Y add to the residual variation, and thus increase the instability of the estimated slope, but do not (on average) attenuate the slope.

One can reverse the equation to get a “**de-attenuated**” slope:

$$\beta_{Y'/X'} = \frac{\beta_{Y'/X'}}{ICC[X]}$$

3 Relationship between test-retest (X' , X'') and $ICC[X]$

X' and X'' denote 2 independent measurements of the X on a randomly selected individual, e.g., measuring one's cholesterol / height/ IQ twice in a short period of time, where X has not changed, and where ϵ_1 and ϵ_2 are independent.

In psychometrics, the term “*test-retest*” is reserved for a self-administered test, such as a questionnaire that is completed by the subject rather than an observer or test-administrator. Otherwise (e.g., if one wishes to study intra-observer or inter-observer variation) psychometricians speak of observer variation, rather than test-retest, studies.

Exercise: Show that

$$\rho_{X',X''} = ICC[X]$$

4 Relationship between $\rho_{X',X}$ and $ICC[X]$

This applies when we can think of X as the ‘gold standard’.

Exercise: Show that

$$\rho_{X',X} = \sqrt{ICC[X]}$$

EXERCISES

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).¹

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, p14 in the notes.] *Hint: it's simply a matter of using the definition of R.*
4. [m-s] Exercise in section 3 (p. 21) of 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 ± 0.024], and weakest for the one between mother and daughter [0.284 ± 0.028]. Given the way he obtained the measurements, can you imagine why this was? ² [It was 0.302 ± 0.027 for mother & son; 0.360 ± 0.026 for father & daughter.]

	Father	Mother	Sons in order of height	Daughters in order of height
1	12.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5	8.5, 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	8.0	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	8.5		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	13.0	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	13.0	14.0		6.0, 4.5, 4.0
19	13.2	3.0		2.7
20	12.7	4.0	13.2, 13.0, 12.7	10.0, 9.0, 8.5, 8.0, 6.0
21	12.0	8.0	13.0	8.5, 8.0
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 "Galton's family data on human stature"

links: left side of JH's home page; direct <http://www.biostat.mcgill.ca/hanley/galton/>

7. **Bridging the physical- and the psycho-metric:** The notes on "Increasing Reliability by averaging several measurements" on the right hand column of page 11 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".³

[m-s] Invert the formula on p.11 to derive the one on the right hand column of p.8 for the Spearman-Brown prediction formula relating the reliability of two versions of a test, one with N times more items than the other.

³ Wikipedia has an entry called 'Spearman Brown prediction formula'.

¹ToneCheck (<https://techcrunch.com/2010/07/20/toncheck/>) 'sounded' like an interesting tool; it's not clear if 'make it' commercially, or was bought by another company!

²After you have thought about it for a while, and looked carefully at Galton's Notebook, you might wish to compare your answer with Karl Pearson's explanation: "Why Galton got different parent-offspring correlations in heights" <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/pearson1930vol13ach14p17-18.pdf> and also look at "why he (KP) got larger ones" <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/PearsonBka1902pp377-378.pdf> using this protocol (p358-) <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/PearsonLee1903.pdf> in the 'Measurement - Lecture Notes, etc' section of the bios601 resources page for Measurement.

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 $B_0 = 32$ and $B_1 = 9/5 = 1.8$, simulate⁴ 1000 datasets and 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.⁵

9. **Attenuation of fitted 'F on C' slopes when progressively greater amounts of error are added to the C measurements**

Run the R code provided under the heading 'Animation (in R) of effects of errors in X on slope of Y on X'. <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/ErrorsInXAnimation.R.txt> It uses the 'animation' package to add progressively greater amounts of error to the C measurements and show how effects they affect the fitted slopes. Include the plot with your answers. Examine the trace of the fitted slopes, and try to mathematically link the

⁴If new to simulations, see "Computer code to simulate datasets with measurement error" <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/FandC.R.txt> at the bottom of the Resources webpage for measurement/surveys. It gives some 'starter' computer code, which you can modify to suit.

⁵The article by Hutcheon et al. "Random measurement error and regression dilution bias", <http://www.biostat.mcgill.ca/hanley/Reprints/RegressionDilutionBMJ.pdf> in the Resources for Measurement page tries to explain these patterns intuitively.

pattern of the 'decay' with the amount of error. *Hint*: as we saw earlier, the attenuation should be a function of (actually, proportional to) the ICC_C ; so use the various amounts of error in C (ranging from $\sigma_{\epsilon_C} = 0$ to $\sigma_{\epsilon_C} = 22$) to calculate the various ICC_C 's and see if the predicted attenuations line up with the trace.

10. 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?

11. 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 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 your 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.

12. **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?]

13. 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:

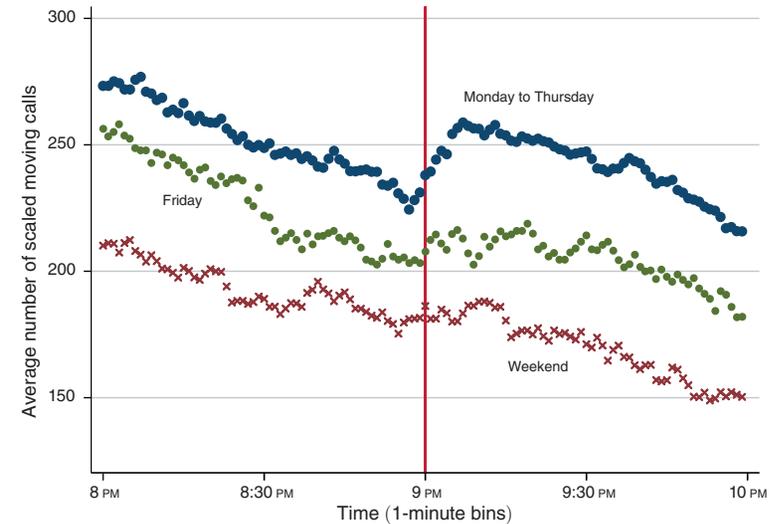


FIGURE 2. CELL PHONE CALL VOLUME FROM MOVING VEHICLES FOR CALIFORNIA FROM 8PM TO 10PM IN 2005

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

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.

15. Physical Activity: JH 2010-2017



Time: Break, Sat, Family

Janvier 2012							2012	
Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	A Faire	
1 9593 3844	2 5157 2494	3 3202 3202 Batting	4 2605 2605	5 4856 6495	6 7487 363	7 6779 956		
8 6640	9 10052 571	10 7347 7899	11 7146 299	12 7529 699	13 4628 729	14 6750 1500 Batting		
15 8230	16 6759	17 12926	18 15218	19 10538	20 11255	21 15954		
	1576	5057	8804	15393	11645	11447	8248	
22 9595	23 13432	24 7421	25 14000	26 8280	27 7274	28 18378		
	6971	10475	5562	12649	7586	2791	15690	
29 5819	30 7044	31 9337						
	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 files (2010-2011, 2012-2013, 2014-2015, 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 pdf files have 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 2010-2111 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.⁶ 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 σ^2 but no structure (i.e. treat all ϵ ’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 model⁷

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

use the method of moments to estimate the σ_w^2 and σ^2 components.

⁶Using Roman b ’s and Greek β ’s to distinguish random effects from fixed effects is a recent convention: it was not used when JH learned linear models.

⁷See 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.

(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?

16. 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’].

- Redo the 2-way ANOVA ‘with participant and trial as main effects’ to see if you can reproduce the reported coefficient of variation.
- 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?
- 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.
- Use the results of the 1-way ANOVA⁸ to calculate an intra-class correlation (ICC).
- In this setting, which makes more sense, a CV or an ICC? Why?
- 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.

⁸The 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.

(g) How could one ‘rig’ (i.e., manipulate) the sample of subjects in the breath alcohol study to (i) maximize (ii) minimize the ICC?

17. 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).

...

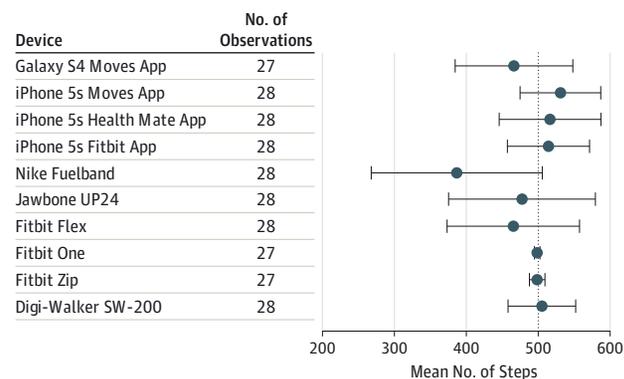
Across all devices, 552 step count observations were recorded from 14 participants in 56 walking trials. Participants were 71.4% female, had a mean (SD) age of 28.1 (6.2) years, and had a mean (SD) self-reported body mass index (calculated as weight in kilograms divided by height in meters squared) of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

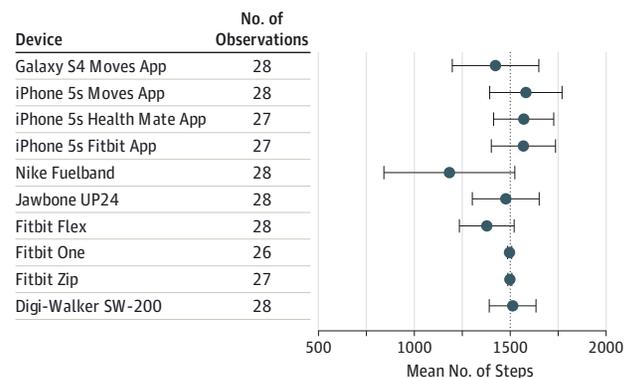
count ranged from -0.3% to 1.0% for the pedometer and accelerometers, -22.7% to -1.5% for the wearable devices, and -6.7% to 6.2% for smartphone applications. Findings were mostly consistent between the 500 and 1500 step trials.

Figure 1. Device Outcomes for the 500 Step Trials



The vertical dotted line depicts the observed step count. The error bars indicate ±1 SD.

Figure 2. Device Outcomes for the 1500 Step Trials



The vertical dotted line depicts the observed step count. The error bars indicate ±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.⁹ 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) In the Fall of 2016, the EPIB601 class carried out its own investigations. The Epidemiology teacher tested 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[d] 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.’

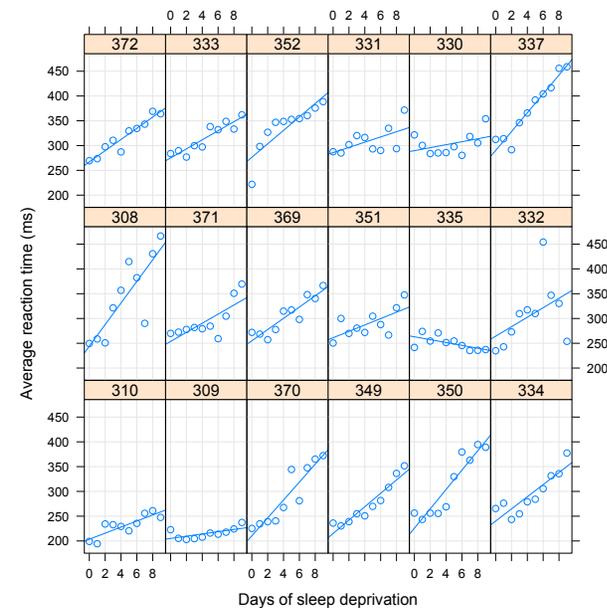
For the BIOS601 of 2016, students were asked to prepare to participate in a planning session, where together they would design (and subsequently carry out) their own investigation into the reproducibility and validity of a few smartphone apps with respect to steps, distance, calories, etc

⁹At 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.

18. Reaction times

The orientational material below is from the `sleepstudy` data re-analyzed in Ch. 3 of the excellent (online) book ‘`lme4`: Mixed-effects modeling with R’, dated June 25 2010, by Douglas M. Bates. The data are included in the `lme4` package – and were used again in the 2017 Epidemiology (teaching) article by Weichenthal, Baumgartner and Hanley.

Belenky et al. [2003] report on a study of the effects of sleep deprivation on reaction time for a number of subjects chosen from a population of long- distance truck drivers. These subjects were divided into groups that were allowed only a limited amount of sleep each night. We consider here the group of 18 subjects who were restricted to three hours of sleep per night for the first ten days of the trial. Each subject’s reaction time was measured several times on each day of the trial.



‘Average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject’s data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right along rows starting at the bottom row, by increasing intercept of these per-subject linear regression lines. The subject number is given in the strip above the panel.’

The 2003 article [European Sleep Research Society, *J. Sleep Res.*, 12, 1-12] that Bates cites is more specific about the Psychomotor vigilance test (PVT), and the number of trials (JH estimates 100 or so) that went into each datapoint shown in the graph [note that Bates used the average response latency whereas Belenky used its reciprocal.]

The PVT measures simple reaction time to a visual stimulus, presented approximately 10 times/minute (interstimulus interval varied from 2 to 10 s in 2-s increments) for 10 min and implemented in a thumb-operated, hand-held device (Dinges and Powell 1985). Subjects attended to the LED timer display on the device and pressed the response button with the preferred thumb as quickly as possible after the appearance of the visual stimulus. The visual stimulus was the LED timer turning on and incrementing from 0 at 1-ms intervals. In response to the subject's button press, the LED timer display stopped incrementing and displayed the subject's response latency for 0.5 s, providing trial-by-trial performance feedback. At the end of this 0.5-s interval the display turned off for the remainder of the foreperiod preceding the next stimulus. Foreperiods varied randomly from 2 to 10 s. Dependent measures, averaged or summed across the 10-min PVT session, included mean speed (reciprocal of average response latency), number of lapses (lapse = response latency exceeding 500 ms), and mean speed for the fastest 10% of all responses.

In bios601 in 2017, each of you will make some rough ('amateur') reaction time measurements, so as to learn what your reaction times are like, and to plan a study into whether they are faster when using your dominant rather than your non-dominant hand.

The 2003 measurements relied on a thumb-operated, hand-held device and a microcomputer program described in 1985.¹⁰

To make your own measurements, you can choose this quite intuitive web tool¹¹ – and use either the keyboard or the mouse/trackpad. It only performs and shows the results of 5 trials at a time. So – since you will need to calculate the mean and SD of 10 individual times – you will need to copy the individual times into R, 5 at a time.

¹⁰Dinges, D. F. and Powell, J. W. Microcomputer analyses of performance on a portable, simple, visual RT task during sustained operations. *Behav. Res. Meth. Instrum. Comput.*, 1985, 17: 652-655.

¹¹<https://faculty.washington.edu/chudler/java/redgreen.html>

[To get around this, JH wrote a simple R program that may not be as accurate or fancy but that stores the individual times from however many you do into a vector. The R code (and links to web-based tools, and to some scholarly and newspaper articles on reaction times) is available under **Online Tools** on the webpage for the Resources for measurement.]

The main objective is to gain experience with 'hands on' data, and with sample size planning, so try both tools and choose between them.

[If you have energy to spare, you can try to empirically determine how closely this R-based instrument and the web-based instrument agree.]

Before running the measurements, be sure to practice first.

- (a) Run 10 trials using your dominant hand, and calculate the mean reaction time, the SD, and the SE of the mean (SEM).

Convert the SEM into a coefficient of variation (CV¹²). How does *this* CV (which measures the 'instability' of the *mean*) relate to the CV for *individual* measurements?

Use the SEM to calculate a 95% confidence interval to accompany your point estimate of the true mean. Why use a larger-than-1.96 multiplier to calculate the margin of error?

- (b) Suppose you wished to perform enough trials that the margin of error would be less than 5% of the mean. Using the SD (or SEM, or CV) you already obtained¹³, calculate how many trials you would need.

Guidance on such sample size *considerations* (JH prefers this term over sample size *requirements*) can be found in section 4 of his bios601 Notes on Mean/quartile of a quantitative variable:- models / inference / planning

- (c) Suppose you wished to (i) test whether, or (ii) measure how much, the mean of reaction times (*r.t.*) obtained with your dominant hand (D) differs from the mean of reaction times obtained with your non-dominant hand (ND).

¹²When reporting a CV, it is customary to do it so as a *percentage*

¹³Of course, if you were to run that many trials, there is no guarantee that the SD would be the same as the SD you got for the 10 – it could be higher or it could be lower. But use the SD of the 10 as the best guess for planning purposes

You will make n measurements with each hand. Assume that there is no ‘fatigue factor’ or ‘order-of-testing’ effect, so that it doesn’t matter whether you first do the n with one hand and then the n with the other. [If there were a fatigue factor, or order effect, then we would want to think of other designs, possibly involving pairing/blocking].

The 2 n ’s may be large enough that the relevant sampling distribution of the difference of two independent sample means (Student’s t) is close to a Z distribution; otherwise, use trial and error. Also assume that the variability is about the same in both r.t. series.

For (i) you will use a 95% confidence interval for the difference of two unknown means, $\mu_D - \mu_{ND}$.

For (ii) you will use the test statistic $\frac{\overline{r.t.D} - \overline{r.t.ND}}{SE \text{ of this difference}}$, and $\alpha = 0.05$ (2-sided).

For the *estimated difference* determine the n per hand that would yield a margin of error of at most: 10 milliseconds; 5 milliseconds.

For the *statistical test* determine the n per hand that would give you an 80% chance of obtaining a ‘statistically significant’ test result if the true difference in milliseconds were: 5, 10, 25.

For the *statistical test*, also determine the chance of obtaining a ‘statistically significant’ test result (the statistical ‘power’, or $1-\beta$) if each n is fixed at 25, but the true difference in milliseconds was: 1, 5, 10, 25.

What if the SD you used for planning was too large? too large?

- (d) Do a few trials using the tool <https://www.justpark.com/creative/reaction-time-test/> that was featured in the newspaper story ‘Brain test judges how old you are based on your reaction time.’

Consider their reaction-time vs. age curve, and how it was fitted. The website don’t say (i) how they selected the 2,000 people aged 18 and above that they surveyed, or (ii) how many trials they asked

each of them to do.

As for (i), describe one scenario where the curve they obtained would be ‘flatter’ than the one that would be obtained if representative population-based samples were recruited at each age.

Suppose¹⁴ that each of the very large number of subjects in each 1-year-wide age-bin was tested a very large number of times. Suppose then that within each age-bin we sorted the persons from slowest to fastest and selected the ‘median’ (middlemost) person. Suppose further¹⁵ that from age 25 to age 64, these medians made an almost perfect straight line with slope 2 ms per year of age, or 0.5 years of age per ms of response latency if we plot age on the vertical (y) axis and response latency on the horizontal (x) axis.

For now, we will retain these 40 people from this ‘ideal’ world. As for (ii), we will ask them to make *just 1 trial each*, and (like the website) use these 40 values to fit the LS line of age(y) upon latency(x).

Assuming within-person variation of the same magnitude as in your own set of measurements, what is your best estimate of what the fitted slope will be? *Hint*: remember some earlier exercises.

The above scenario selected the *median* person in each bin. If you picked one *random* person from each bin, what is your best estimate of what the fitted slope will be? (State your assumptions).

Write a few sentences summarizing why (even if their sample of subjects is representative) the age-latency graph in the website may be inaccurate, and in what respect.

- (e) What if each median-person’s latency was measured perfectly (large n), but ages were in bins (intervals) 5 years wide (so that, e.g., the persons aged 25, 26, 27, 28 and 29 are put at age 27), and we fitted the LS line of latency(y) upon the *midpoint* (x) of each age bin?

¹⁴This ideal universe where subjects are easily recruited, and have lots of patience and can maintain their attention over a very large number of trials, is just for didactic purposes.

¹⁵Now we are really dreaming! While we are at it, we will assume symmetric age-specific distributions.

Note re terminology:

In the situation where $x = \text{latency}$, the errors in measuring the true X values are uncorrelated with these true values of X . This is called the *classical* ‘errors in X ’ situation. It is the nastier case.

$$X = \text{true value}; \quad x = X + \epsilon_X, \text{ with } \epsilon_X \perp X$$

In the situation where $x = \text{the mid-age of the bin}$, the errors in measuring the true X values (ages) are correlated with the true values of X , but uncorrelated with the observed x 's. This is called the *Berkson* ‘errors in X ’ situation. It is less nasty, but it does increase the (sampling) variability of the estimated slope.

$$X = \text{true value}; \quad x = X + \epsilon_X, \text{ with } \epsilon_X \perp x$$

JH's favourite example of Berkson error (one he adapted for the earlier exercise on F v.s C temperatures) is one that may have come from Berkson himself: An investigator wished to measure temperatures in an oven at various times.

- An unreliable thermometer, i.e., one that gives readings that fall equally on both sides of the truth, would generate classical errors.
- The temperatures shown on the thermostat are as likely to be above/below the true temperature at any given moment of interest; as you can check, these would be Berkson errors.

For more on these, consult JH's Ch. 4 notes in his Applied Linear Models course 679, or the books or presentation by the (measurement-expert) statistician Raymond Carroll https://www.stat.tamu.edu/~carroll/talks/NCI_MEM_Call.pdf

19. **Instead of measuring heights with a tape, how about using a smartphone app?**

Q. prompted by revisiting Pearson's protocol, and by this piece, <https://lifehacker.com/which-ar-measuring-app-is-more-accurate-1827242756>, found when searching 'measuring heights smartphone'.

20. **Who seems to age faster?** The following are the reported ages of the 40 students in JH's course 513-607 (Inferential Statistics) in 1986.

```
AGE.1986 = c( rep(22,4), 23, 25, rep(26,3), rep(27,4), rep(28,3),
             rep(29,2), rep(30,5), rep(31,4), rep(32,2),
             rep(33,2), rep(34,2), 35, 36, 37, rep(38,2), 39,42 )
```

- Make a new variate `AGE.1999` from the `AGE.1986` variate.
- use the `lm` function to estimate, from the regression of `AGE.1999` on `AGE.1986`, how much these students aged in the intervening 13 years.
- Notice the use of uppercase `AGE` to denote the true age. What if these 40 students had reported their 1986 ages as their true 1986 ages ± 5 years (with the - or + determined at random, without regard to the person's true age)? i.e. as (say) `age.1986 = AGE.1986 + 5*sample(c(-1,1), 40, replace=TRUE)`. Note the use of lowercase `age` to denote the ‘error-containing’ value [In the measurement error literature, and in JH's notes, it is common to use X and Y for the true values and X^* and Y^* , or X' and Y' , for the error-containing values.]

Now, again, use the `lm` function to estimate, from the regression of `AGE.1999` on `age.1986`, how much these students aged in the intervening 13 years, and who aged the most and who the least. Comment on your findings, and give a non-technical explanation that your engineer-sibling would understand.

You might want to simulate several `age.1986` vectors to convince yourself that the effects are reproducible. Or – if keen on algebra – work out how much, on average, the attenuation is.

- Apply the `lm` function again, but this time with `AGE.1999 - age.1986` as the ‘ y ’ variate, and `age.1986` as the ‘ x ’ variate. Comment.

21. **Some ‘big-ticket’ epidemiology examples.** Refer to the 1997 article The INTERSALT study: background, methods, findings, and implications. *Am. J. Clin. Nutr.* 65, 626S-642S. by J Stamler¹⁶ It is available here: <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/interSALT.pdf>

The last 2 columns of Table 2 on p630S are entitled ‘Observed coefficient as percentage of true coefficient’ when there is one measurement and when there are four. The footnote explains how it is calculated from the ratio of the intra- to inter-individual variance and the number of measurements.

- Verify the calculations for 24-h Urinary Na excretion.

¹⁶cited on p. 72. of Chapter 4 (Principles of measurement) of Cox and Donnelly's 2011 book *Principles of Applied Statistics*, available as an eBook from McGill. <https://www-cambridge-org.proxy3.library.mcgill.ca/core/books/principles-of-applied-statistics/E7225E64F86B2C8193CA3C57621B6338>.

- (b) Suppose you had the ICC (rather than the Intra- to inter- variance) as the column header. Alter the wording of the footnote accordingly.
- (c) What is the relation between these and the ‘stepped-ip reliability’ measures addressed in question 7?
- (d) Show how the numbers in row 2 of Table 3 were derived from those in row 1.
- (e) Use JH’s daily steps in 2010 (see Q. 15) to work out an ‘intra-’ variance. We will assume JH’s intra- is typical of the ‘intra-’ variance of other people of JH’s age.
- (f) We don’t have average steps per day for that year for many other people his age, but assume we did have it for a large number of individuals, and that the mean of this large number of person-specific yearly averages is 6,000 steps/day, and the (inter-individual) SD is 2,000 steps/day. Using this SD, and the results from (c), to add a row, entitled say ‘activity, measured as steps/day,’ to Table 2.
- (g) Comment on the slope for the leftmost portions of the relationship in the top left panel (Total physical activity (cpm)) of Figure 2 of the article “Dose-response associations between accelerometer measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis.” The file is here: <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/PhysActivityAllCauseMortality.pdf> Since the y axis is on a log scale, it shows $\log(\text{Hazard ratio})$ v.s physical activity as being approximately linear for the early part.
 Try to de-attenuate the slope, by assuming (a bit unrealistically) that the measurements shown on the X axis are based on 7 random days over a year, and that the intra- to inter- variance ratio you calculated in part (d) applies to the Total physical activity (cpm) measurements in this study.
- (h) Refer to the article “Association of Office and Ambulatory Blood Pressure With Mortality and Cardiovascular Outcomes.” The file is here <http://www.biostat.mcgill.ca/hanley/bios601/Surveys/BP-1-time-24-hr.pdf> Which blood pressure index would you expect to have the strongest relationship with mortality rates, and why? Are your expectations borne out?

Type IV error

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
	- to +	6,12	-	7,6	-	5,47
13	+ to -	5,97	-	7,7	-	5,63
	- to +	6,27	-	7,6	-	5,34
14	+ to -	6,13	-	7,6	-	5,46
	- to +	6,34	-	7,7	-	5,3
15	- to +	6,1	-	7,16	-	5,75
	- to +	5,78	-	7,2	-	5,68
17	+ to -	5,64	-	7,3	-	5,85

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

22. 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.