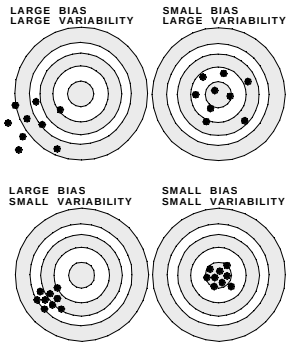



Being approximately correct and being precisely wrong

## PREFACE

Few general statistics textbooks[1] deal with the topic of measurement error. However, it an important topic. Measurement error is present in all scientific work; while it can sometimes be lessened with careful planning and extra effort, it cannot be completely avoided. It some contexts it just adds noise and makes signals harder to detect. *But in others, its effects are both more subtle and more serious*: it systematically shifts the parameter estimates produced by models.

Its effects become more complicated and unpredictable in larger statistical models with two or more measured-with-error variables. We will restrict attention here to simple comparative situations involving one X and one Y. We will focus on the '*classical error model*' where the errors in the measurements are uncorrelated with the true values being measured. We briefly mention the '*Berkson error model*,' where the errors in the measurement are uncorrelated with the error-containing measurement – and thus correlated with the true value being measured.

More complicated cases (for example, human under-/over-reporting of values) require some knowledge of the subject matter area to know when such are likely to be present.

In the following home-grown material, JH will rely in part on notes he prepared when teaching measurement concepts/principles to graduate students in the physical and occupational therapy sciences. There, since the qualities/quantities being 'measured' were often **psychophysical rather than physical**, he drew on the very long history of '**psychometrics**' in the psychology and education literature. This was a learning experience for him, since, up until then, he had mainly been concerned with the physical quantities measured in medical research.[2] He was greatly helped by reading parts of the textbook *Psychometric Theory* by Nunnally, and found that Nunnally's examples of constructing (and measuring something by) exams were concrete and understandable illustrations of the concepts and principles. For those interested, his 1975 'lookback' is a very nice introduction to the subject. There he tells us that

> Since Spearman wrote on the topic in 1904, the theory of measurement reliability, or the converse of that, the theory of measurement error, has been a special interest of psychologists. I have always thought that this was partly out of necessity and partly because reliability theory is so neatly mathematized. In his 1904 article. Spearman developed most of the basic statistics pertaining to reliability that are still with us, including corrections for attenuation, the standard error of measurement, the correction of the split-half reliability coefficient for test length, and other statistics that are identified with test reliability. This deductive model, and the attendant mathematical developments, stood as a comprehensive theory of reliability for over 30 years.

## QUANTIFYING THE QUALITY OF A MEASURING INSTRUMENT

Not surprisingly, much of our terminology comes from psychology, where measurements are seldom in centimetres, or kilogram. For the latter we have established standards established/regulated by various governmental agencies, such as those in the US, Canada , the EU and the UK

There are two aspects, ***Reliability*** ('reproducibility' / 'precision', the degree to which measurements 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 the measurements measure what they are intended to measure).

### *But which to consider first?*

When asked which they would study first, most students jump immediately to validity. Thus, in the following case, described in a Dr Fowler's 1982 letter to the Editor of JAMA they want to know how the predictions turned out.

[1] Examples of specialized textbooks focusing just on measurement error are *Measurement error in nonlinear models* by Carroll, Ruppert, and Stefanski 1998 and the 2006 *Measurement error in nonlinear models : a modern perspective.* by Carroll, the 1987 *Measurement error models* by Fuller, the 2004 *Measurement error and misclassification in statistics and epidemiology : impacts and Bayesian adjustments* by Gustafson. See also the excellent **Chapter** 1 '***Reliability of Measurent***' in the (older) book ***The design and analysis of clinical experiments*** by **Fleiss**. These, and others, are available from the McGill Library. A Google search will bring up more, as well as a large number of journal articles and other literature on this now-more-widely-studied topic.

[2] JH was aware of measurement errors when as part of his PhD work, he made low-tech exhaled-breath measurements of the carbon monoxide in the blood of cigarette smokers, and used the amounts of various substances found in their cigarette butts to estimate their mouth-level exposure to nicotine. But he did nothing about it. Later, when he began working in clinical trials in oncology, he got a rude awakening. Fortunately, this time, after seeing just how much measurement error there was, and heeding the warning in the report, the Eastern Cooperative Oncology Group changed their cutoff for a partial response from a 25% to a 50% reduction in the cross-section area of the measured tumour.

> 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* [3] i.e. **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.

| | | Prediction | | | |
|---|---|---|---|---|---|
| | | Inconsistent | Girl | Boy | TOTAL |
| "TRUTH" | GIRL | 11 | 20 | 22 | 53 |
| | BOY | 10 | 20 | 17 | 47 |
| | | 21 | 40 | 39 | 100 |

> *From this brief study, it would appear that the Drano test for antenatal sex determination is roughly equivalent to flipping a coin.*

JH has also watched technicians test a photocopying 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***.

[3] Not to be confused with its meaning in mathematical statistics, when at issue is the behaviour of statistical estimators.

**Roadmap for the remainder of these Notes**

The notes on pages 3-19 are from lectures JH gave in a course on measurement for the rehabilitation sciences.

- Pages 3-7: since few of these students had had experience with ANoVA's (widely used to estimate reliability coefficients) the lectures began with a general introduction to these, 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 8 is a **nice introduction to reliability and validity**: the ideas are general, not confined to examinations, but to any measurement method/instrument.

  Note the **two reliability measures**:

  – a standard deviation (SD) measured in the **same scale** (scores, cm, Kg, ... ) **that the instrument uses**, [smaller is better]
  – 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). [bigger is better]

  These are the 'converses' that Spearman mentioned
- Pages 9-17 use ANoVA tables to estimate the variance components used as inputs to the reliability coefficients (**Intra-Class Correlations, p.10**).
- Pages 18-19 are ways to assess **validity**. [NB the different ways terms like accuracy, reliability, validity are used in statistics, and outside statistics].
- Pages 20-22: the **Effects of Measurement Error** & **role of the ICC**.
- Pages 22-23: **Berkson** Error

**A passing comment**: If JH had his way, we would be using the term reproducibility rather than reliability. One possible problem with the word 'reliable' is that it has many meanings in everyday English. What does one mean by a 'reliable' friend, or car, or clock, or watch? (and if you look up the history of the word, it was not without controversy grammar-wise). To JH at least, the word reproducible is less ambiguous. There is also the joke about a stopped clock: it is correct 2 times a day! Does that make its result reproducible? reliable? ? ?

**RELIABILITY** (Reproducibility, Precision): The extent to which one obtains the same (or very similar) answers/values/scores if an object/subject is measured repeatedly under similar situations

Some ways to quantify Reliability

- For ***one*** subject (s) or object:
  - 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.
  - SD of the measurements as a percentage of the mean of the measurements ... the [within-subject] Coefficient of Variation (CV).
- For ***several*** subjects (ss.) or objects:
  - average 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. In biochemical determinations, CV may differ at different concentrations.
  - The (within-ss) CV gives no sense of how well the measurements of different subjects (ss) segregate from each other. We might compare the SD of the within-ss measurements with the SD of the between-ss measurements, but, as we will see below, it makes more sense to focus on variances.
  - Can we use correlation (Pearson or Spearman) if we have 2 assessments of each ss.? What if we have a triplets of measurements for each subject? (cf. study of otitis media in twins and triplets)
  - 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 object is a population*** (.e.g., the percentage of smokers, or showing immunity, 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

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

  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.
- ***When the scale is arbitrary***, it is common to use a function of the (theoretical) Variance of the Within-ss measurements and the Variance of the Between-ss values. Classically, these **Components of Variance** were estimated using **Analysis of Variance** (**ANoVA**). We will also use a Bayerian approach.

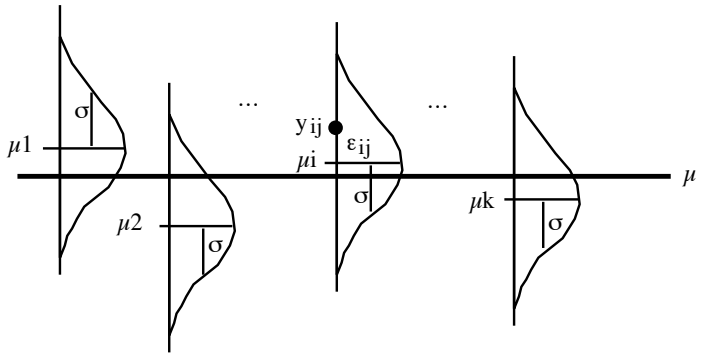
## First, a General Orientation to Anova and its primary/classical use: testing differences between $\mu$'s of $k$ ( 2 ) different groups. [4]

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...**

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

[4] This and the next several pages contain images of pages made with MS Word decades ago. JH hasn't had time to redo them in LateX. **2020 additions in red**

**DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION**

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

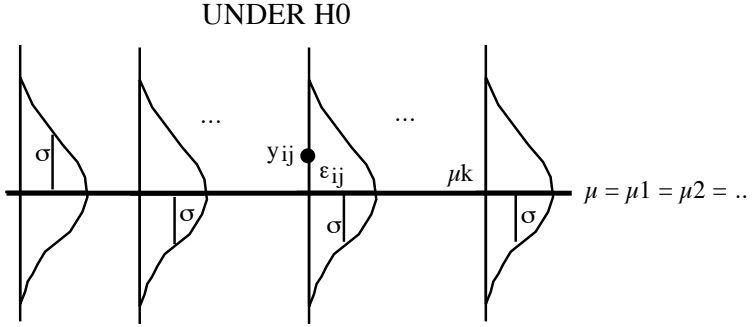
| TOTAL Sum of Squares | = | BETWEEN Groups Sum of Squares | + | WITHIN Group Sum of Squares |
|---|---|---|---|---|

**ANOVA TABLE**

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

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

**$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$**




Means, based on samples of n, should vary around $\mu$ 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 $\sigma^2$

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

Whether or not $H_0$ is true, the empirical variance of the n (within-group) values $y_{i1}$ to $y_{in}$ i.e. $\frac{\sum[y_{ij} - \bar{y}_i]^2}{n-1}$ should give us an unbiased estimate of $\sigma^2$

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

so the average of the k diferent estimates,

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

is also an unbiased estimate of $\sigma^2$

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

THUS, under $H_0$, both $MS_{BETWEEN}$ and $MS_{WITHIN}$ are unbiased estimates of estimates of $\sigma^2$ 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 $\mu$'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 estimaes 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 $\mu$'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 though 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 varaition, 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'y 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:**

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

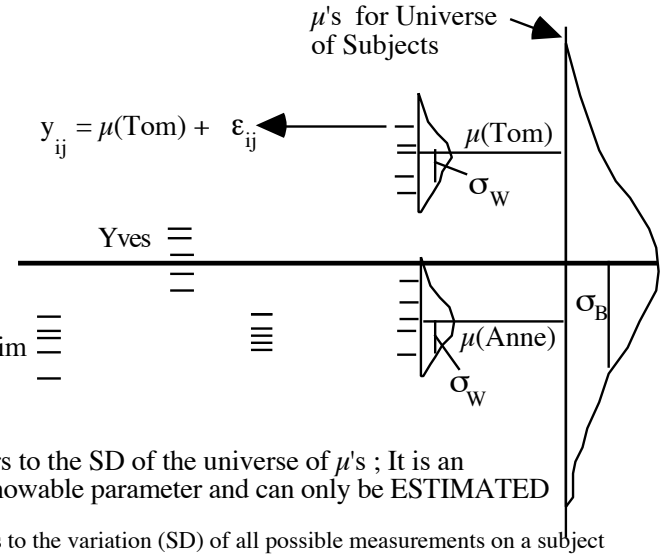
**MODEL**

The model is different. There is no interest in the specific subjects. Unlike the critical labels "male" anf "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 measurents go with the correct subject!

The "Random Effects" Model uses 2 stages:

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

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




$\sigma_B$ refers to the SD of the universe of $\mu$'s ; It is an unknowable parameter and can only be ESTIMATED

$\sigma_W$ 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...**

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

$$= \mu + \alpha_i + \varepsilon_{ij}$$

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

$\varepsilon_{ij} \sim N(0, \sigma^2_W)$

**DE-COMPOSITION OF OBSERVED (EMPIRICAL) VARIATION**

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

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

**ACTUAL ESTIMATION OF 2 Variance Components**

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

$\mathbf{MS_{WITHIN}}$ is an unbiased estimate of $\sigma^2_W$

By subtraction...

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

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

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:**

| | Subject Tom | Anne | Yves | Jean | Claire | Emma |
|---|---|---|---|---|---|---|
| Measurement 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 SS | Degrees of Freedom df | **Mean Square** MS (= SS /df) | **What the Mean Square is an estimate of...** * |
|---|---|---|---|---|
| BETWEEN Subjects | 9.205 | 5 | 1.841 | $\sigma^2_W + n\,\sigma^2_B$ |
| WITHIN Subjects | 0.640 | 12 | 0.053 | $\sigma^2_W$ |
| TOTAL | 9.845 | 17 | | |

**ESTIMATES OF VARIANCE COMPONENTS**

$\mathbf{MS_{WITHIN}}$ **= 0.053** is an unbiased estimate of $\sigma^2_W$

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

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

**This is an example of the method of moments.**

**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 | | | | 6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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
                      Sum of     Mean
Source             DF  Squares   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)

Var(Error) + 2 Var(SUBJECT) = 48.71
Var(Error)                  =  1.38
             2 Var(SUBJECT) = 47.33
               Var(SUBJECT) = 47.33 / 2 = 23.67
```

The measurements were made by 4 physical/occupational therapy graduate students (obsr 1-4), on the remaining n = 8 subjects – 6 fellow students and 2 instructors, JH and SD. After a short meeting to standardize themselves, and using clear plastic rulers, each of the 4 measured each of the 8 twice (1st and 2nd) The final subjcet whould be numbered 8, not 6.

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

                          Estimate
Variance Component        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

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

**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.

This excellent blurb is from the days when Graduate Records Examinations (GRE's) were taken on paper and took 3 hours or more. Today, with the '**adaptive testing**' that is possible with online tests, the exams can be personalized and speeded up, by starting with questions that are used to get a general sense of one's ability/knowledge, and them using questions of known difficulty to move up or down. So the sequence (and number) of questions one candidates gets differ from those another candidate gets. This leads into a later topic: the method of split halves used to measure reliability.

### 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.

This last comment about the meaning of the word '**error**' also applies in statistics more generally. When communicating with non-statisticians, we should carefully explain our terminology.

**Outline**

**Some ways to quantify Reliability:**

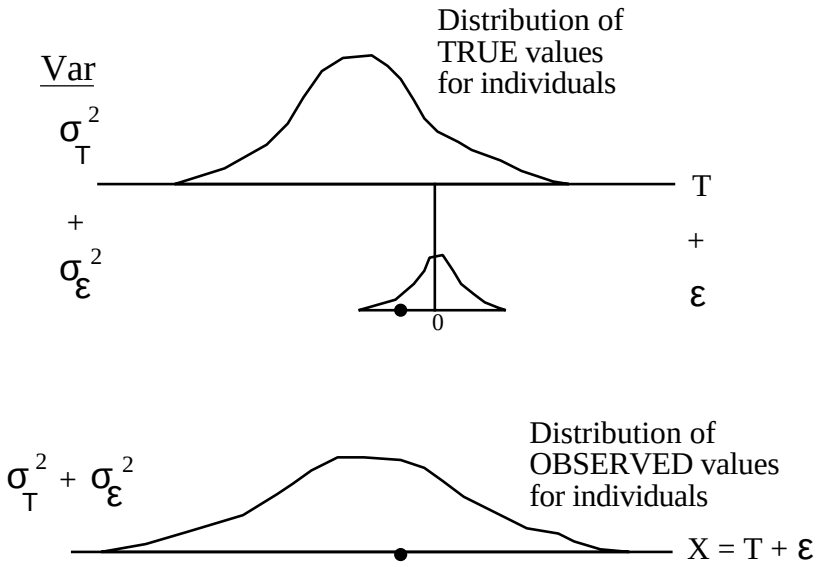
- Reliability Coefficient
- Internal Consistency (Cronbach's α)

**Implications:**

-

.....

**Model for Reliability**




"True" scores / values not knowable;

~~ε ⊥ T~~ Variance calculation assumes that the distribution of errors is independent of T

**Reliability Coefficient**

$r_{XX} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_\varepsilon}$ i.e. the fraction of observed variation that is 'real'

Note that one can 'manipulate' r by choosing a large or small $\sigma^2_T$

**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 |

The above 'stepped-up reliability' formula is often called the Spearman–Brown prediction formula. See Brown 1910: correlation of mental abilities.
See Wainer et al. on why unnecessarily long tests may be impeding the progress of Western civilisation.

**Cronbach's $\alpha$**

k items

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

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

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

$r_{XX} = \alpha$ 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**

  e.g. $$\frac{\sigma^2_{TRUE}}{\sigma^2_{TRUE} + \sigma^2_{ERROR}}$$

  with estimates of various $\sigma^2$ 's substituted for the $\sigma^2$ '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!!)

**To prevent cheating, in some multiple choice exams students get versions with the order of questions permuted – or even different questions. How reproducible are students' scores? Do students rank the same if we use the subtotals for the even- and odd-numbered questions? We could correlate a random-half versus the reminder. We could average the correlation over many such random splits. Cronbach's alpha is seen as a way of not having to run a second exam, but to work out the reliability '*internally*' from the one exam The ETS has a large enough question bank to ensure a high correlation between scores on different versions**

**ICC's (Portnoy and Wilkins)**

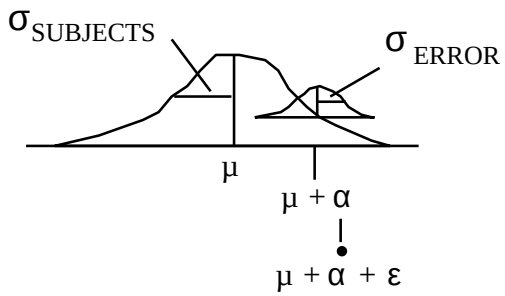
*(1) multiple (unlabeled) measurements of each subject*

*(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*




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

Model for observed data:

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

**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 though 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 = $\mu + \mu + \alpha_i + e_{ij}$

**... backstory of example 1 here; data & R Code* here**

**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 | $\sigma^2_{\text{"error"}} + k_0\ \sigma^2_{\text{between}}$ |
| Error(Within families) | 30 | 12 | 2.5 | $\sigma^2_{\text{"error"}}$ |
| 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... $k_0$ = average k – (variance of k's) / (n times average k) ...see Fleiss page 10]

**Estimation of parameters that go to make up ICC**

2.5 is an estimate of $\sigma^2_{\text{"error"}}$

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

---

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

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

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

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

*R code uses MCMC to obtain a credible interval for ICC

**Origin of w'in- ('intra-') vs. b'ween- family ('class') terminology**

**COMPUTATIONAL Formula for "1-way" ICC**

$$\frac{\frac{MSbetween - MSwithin}{k_0}}{\frac{MSbetween - MSwithin}{k_0} + MSwithin}$$

$$= \frac{MSbetween - MSwithin}{MSbetween + (k_0 - 1)MSwithin} \quad \text{[shortcut]}$$

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 $(6\text{-}6)^2+(4\text{-}4)^2+(2\text{-}1)^2 +...(8\text{-})^2 =10$. If one were to do the calculations by hand, one usually calculates the SStotal and then obtains the SSerror 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...**

Var(SUBJECT) = 23.67 Var(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: $y_{i,j} = \mu + \alpha_i + e_{ij}$

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

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

$$ybar_i = \mu + \alpha_i + ebar_i$$

then

$$\underset{i}{Var}[ybar_i] = \sigma^2_{between} + \frac{\sigma^2_{\text{"error"}}}{k}$$

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

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

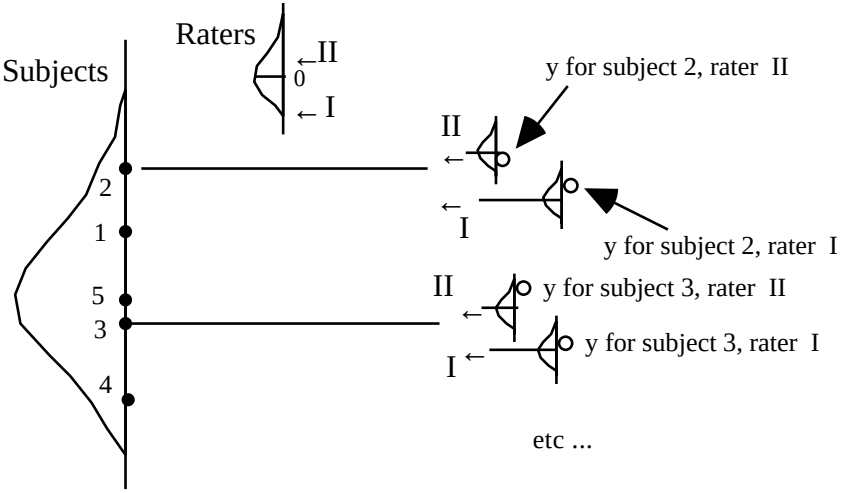
**Computational versions, used to save steps when computing statistics by hand, hide the concept (definition) behind these statistics.**

**The MCMC method yields a more realistic credible interval than Fleiss' CI.**

**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**




$$\mu + \alpha_{[subject]} + \beta_{[rater]} + \varepsilon$$

$$\sigma^2_{subjects} \quad \sigma^2_{raters} \quad \sigma^2_{error}$$

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

estimate $\sigma^2_{"subjects"}$, $\sigma^2_{"raters"}$ and $\sigma^2_{"error"}$ 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 | | | | 6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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

                          Sum of      Mean
Source                DF  Squares    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...

```
Var(Error) + 4 Var(SUBJECT) = 104.98
Var(Error)                  =   0.97
        ==>  4 Var(SUBJECT) = 104.01
        ==>    Var(SUBJECT) = 104.01 / 4 = 26.00

Var(Error) + 8 Var(RATER)   =   9.87
Var(Error)                  =   0.97
        ==>  8 Var(RATER)   =   8.90
        ==>    Var(RATER)   =   8.90 / 8 =  1.11

               Var(Error)                =  0.97
```

Estimating Variance components using PROC VARCOMP in SAS

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

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

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

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

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

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

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

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

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

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

## 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 |

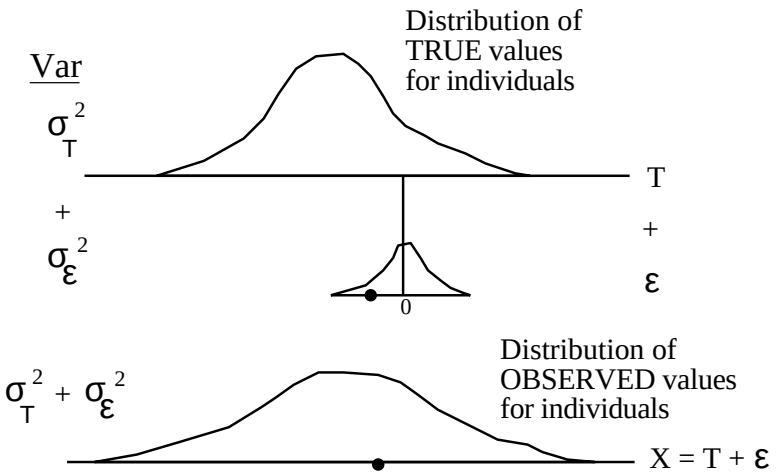
## LINK between STANDARD ERROR OF MEASUREMENT and RELIABILITY COEFFICIENT

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

Standard Error of Measurement = 23 points

Reliability Coefficient: R = 0.93

recall...




$\sigma^2_e = 23 \quad ==> \sigma^2_e = 529$ ;

$$R = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_e} = 0.93 \quad ==> \sigma^2_T = \frac{R \times \sigma^2_e}{1 - R} = \frac{0.93 \times 529}{1 - 0.93} = 7028$$

$$\sigma^2_T + \sigma^2_e = 7028 + 529 = 7557 \quad ==> \quad \sqrt{\sigma^2_T + \sigma^2_e} = \sqrt{7557} = 87$$

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) :-

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

**See section 12.7.2 (Measurement error, here for a colour-based depiction of the mixing of 'true' and 'error' distributions.**

## 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.

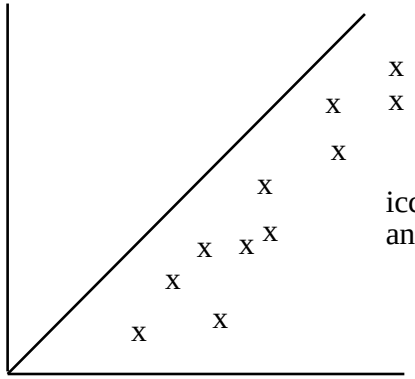
**NOTE**: If interested in ICC that incorporates random raters, then sample size must involve both # of raters and # of raters

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"

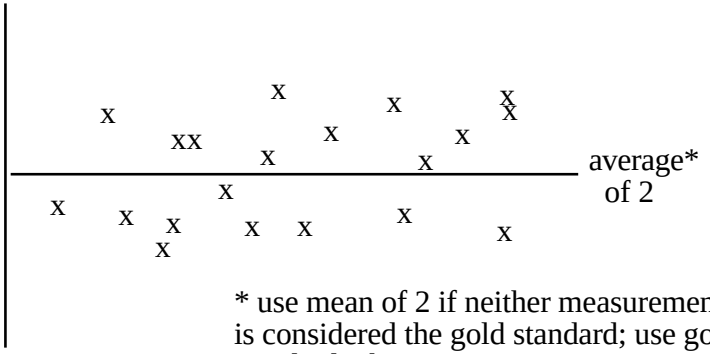


icc includes "shifts" and is lower than r

2. not practical if > 2 measurements or variable # of measurements per subject

   ICC 'made for' such situations

## Method of Bland & Altman [Lancet ]




* use mean of 2 if neither measurements is considered the gold standard; use gold standard otherwise

+++ see biases quickly

can explain to your in-laws
(can you explain ICC to them?)

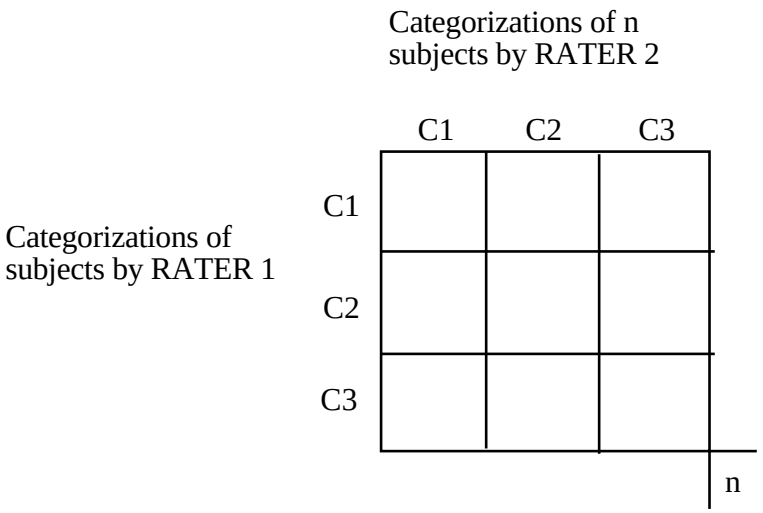
emphasises errors in measurements scale itself
(like ±23 in GRE score)

– – – 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 <u>raters</u>

## Assessing reproducibility of measurements made on a CATEGORICAL scale




*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

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

* expected proportion = $\sum$ p[row]*p[col] --- $\sum$ 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

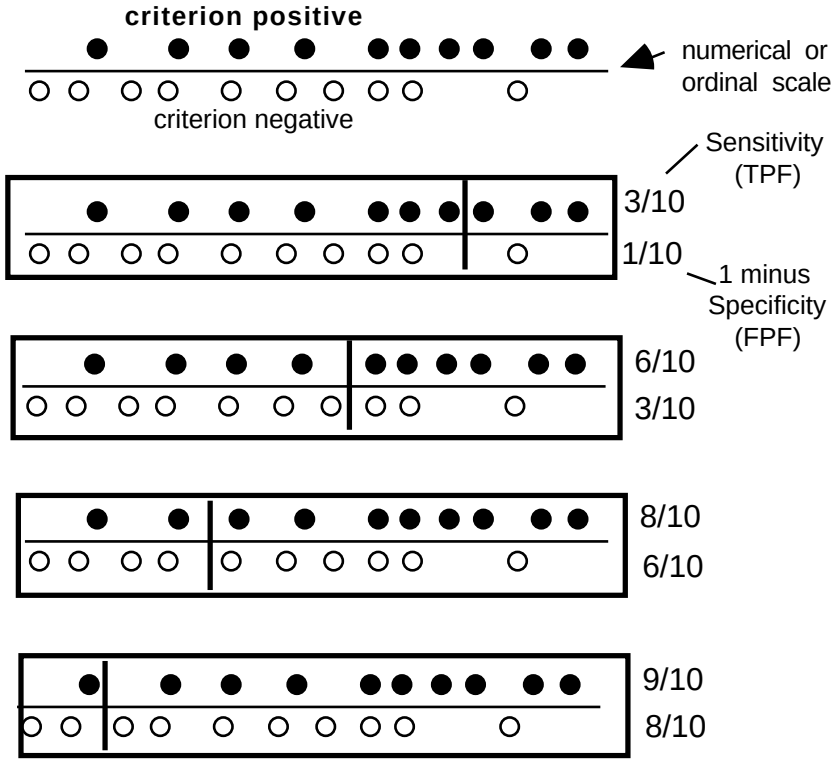
| INSTRUMENT | CRITERION | STATISTIC(S) |
|---|---|---|
| numerical | numerical - same scale | • calculate discrepancies (Altman & Bland)<br>• describe distribution of discrepancies |
| numerical | numerical - different scale | • correlation<br>[Pearson or Spearman] |
| numerical | ordered categories (e.g. known groups) | • correlation [ranks]<br>• ANOVA [groups] |
| ordered categories | ordered categories | • Kendall's "tau" |
| numerical or ordered categories | binary | • difference in means<br>[ parametric/ nonparametric test ]<br>• ROC curve |
| binary | binary | • sensitivity<br>(Probability + on instr. if Criterion +)<br>&<br>specificity<br>(Probability – on instr. if Criterion –)<br><br>• predictive values<br>• ϕ (phi) statistic (see Streiner & Norman) |

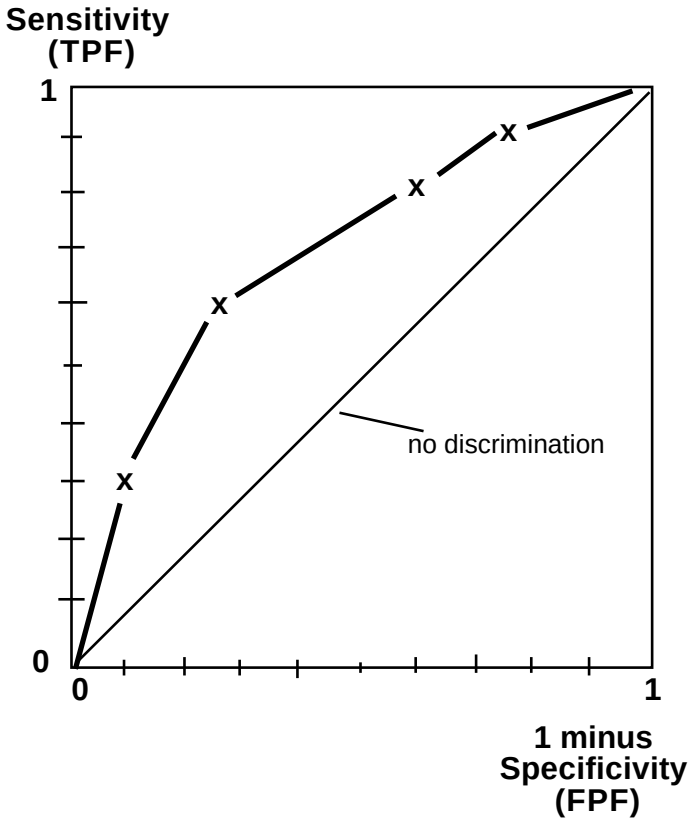
**For many reasons, this treatment of Validity is short: if the quality is physical, we usually have a gold standard, so assessment is direct (example: Tracking Physical Activity Data). If it is psycho-physical or conceptual the challenge is less in the statistical methods, and more in finding indirect ways to assess it: see the paragraph on VALIDITY in the blurb about the GRE, or the q. on readability in Exercises 1 and 2 below.**

**Receiver Operating Characteristic Curve**

| instrument | criterion |
|---|---|
| numerical or ordered scale | binary ( ● or o) |

- SERIES of {sensitivity, specificity} statistics, each based on a different cut-off
- usually plotted on a graph, showing tradeoff between sensitivity and specificity
- Summary statistics (performance)
  - sensitivity at a given (specified) specificity
  - area under the ROC curve




TPF: True Positive Fraction
FPF: False Positive Fraction

**Reference**: section 5 chapter 13 in 2nd edition of Basic & Clinical Biostatistics by Beth Dawson-Saunders and Robert Trapp, Appleton & Lange, Norwalk (CT)

.

# CONSEQUENCES

# of

# MEASUREMENT ERROR

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

.

typeset with LaTeX

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

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

and

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

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

Let $X' = X + \epsilon_X$ where $\epsilon_X$ has mean 0 and variance $Var[\epsilon_X]$, **and is independent of X**[5], so that

$$E[X'] = E[X + \epsilon_X] = E[X] + E[\epsilon_X] = E[X] + 0 = E[X] \tag{3}$$

$$Var[X'] = Var[X + \epsilon_X] = Var[X] + Var[\epsilon_X] \tag{4}$$

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

$$E[Y'] = E[Y + \epsilon_Y] = E[Y] + E[\epsilon_Y] = E[Y] + 0 = E[Y] \tag{5}$$

$$Var[Y'] = Var[Y + \epsilon_Y] = Var[Y] + Var[\epsilon_Y] \tag{6}$$

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

By definition

$$ICC[X] = \frac{Var[X]}{Var[X] + Var[\epsilon_X]} \;\&\; ICC[Y] = \frac{Var[Y]}{Var[Y] + Var[\epsilon_Y]}, \tag{8}$$

with

$$0 \leq ICC[X] \leq 1 \;\&\; 0 \leq ICC[Y] \leq 1. \tag{9}$$

[5] This is known as the '***Classical***' error model.

# 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...

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

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**:

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

```
http://www.m-w.com/dictionary/attenuate
*Main Entry:  1atátenáuáate ; 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...

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

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**:

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

As Spearman told us, this method goes back to 1904. In the last 40 years, there have been major developments in the statistical handling of measurement error (see the textbooks mentioned at the beginning). One of the more creative is **Simulation Extrapolation Estimation** ('**SIMEX**') in Parametric Measurement Error Models, Cook and Stefanski, (1994) JASA , 89, 1314-1328. See also the article Simulation-Extrapolation: The Measurement Error Jackknife by the same authors, as well as the `simex` Package for `R`.

## 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 $\epsilon_1$ and $\epsilon_2$ 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

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

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

This applies when we can think of $X$ as the 'gold standard'.

***Exercise:*** Show that

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

# **Berkson** error model

Results 1-4 are based on the **Classical** error model.

There is another, less common, error model, where the consequences are different. This is the **'Berkson'** error model, named after this Mayo Clinic statistician/epidemiologist, Joseph Berkson.

- Again, we denote the "true" (but unobservable) values by $X$ and the observed (error-containing) measurements by $X'$
- Let $X' = X + \epsilon_X$ where $\epsilon_X$ has mean 0 and variance $Var[\epsilon_X]$, **and is independent of $X'$**

This type of error is present in some environmental epidemiology studies, as nicely illustrated in the folllowing excerpt from this teaching piece.[6]

> This distinction [between classical and Berkson error models] is not well known and a little tricky to understand, but it has major implications for the effects of the error.
>
> • Classical: The average of many replicate measurements of same true exposure would equal the true exposure.
>
> • Berkson: The same approximate exposure (proxy) is used for many subjects; the true exposures vary randomly about this proxy, with mean equal to it.
>
> **Example**:
>
> A study investigates the relation of mean exposure to lead up to age 10 with intelligence quotient (IQ) in 10 year old children living in the vicinity of a lead smelter. The IQ is measured by a test administered at age 10. Consider two study designs for assessing exposure:
>
> • *Design 1*: Each child has one measurement made of blood lead, at a random time during their life. The blood lead measurement will be an approximate measure of mean blood lead over life. However, if we were able to make many replicate measurements (at different random time points), the mean would be a good indicator of lifetime exposure. This measurement error is thus classical.

[6] Its author, Ben Armstrong, worked in Occupational Health at McGill for some years, and (as you many gather from his piece) was known as an outstanding teacher. When here, Ben worked with Gilles Thériault on studies of electromagnetic fields and cancer in electric utility workers in Québec and France.

• *Design 2*: The children's place of residence at age 10 (assumed known exactly) are classified into three groups by proximity to the smelter - close, medium, far. Random blood lead samples, collected as described in design 1, are averaged for each group, and this group mean used as a proxy for lifetime exposure for each child in the group. Here the same approximate exposure (proxy) is used for all subjects in the same group, and true exposures, although unknown, may be assumed to vary randomly about the proxy. This measurement error is thus Berkson type error.

Another situation giving rise to Berkson error is when exposures are estimated from observed determinants of exposure with an exposure prediction model. Often error has both classic and random components, although one usually predominates.

On page 6 of these notes JH gives an example of Berkson error, which he thinks he got from Ben Armstrong.[7]

It would apply if you recorded the temperature $X'$ to which a stove or thermostat was 'set', but the true temperature $X$ at the time the response ($Y$) was observed/measured could be somewhat above or below $X'$. The red dots in JH's diagram are what one might observe, and thus used to fit the regression model.

One way to think of the effect of Berkson error is that it moves the $Y[X']$ value *vertically*, but keeps it directly above the $X'$ value.

By contrast, the effect of Classical error is that it moves the $Y[X]$ value *horizontally* to $X'$, but keeps it at the same vertical position.

Panel B (opposite) shows temperature fluctuations about a 'set' value.[8]

You could also 'see' why the Berkson error does not 'flatten the line' if (with $X'$ – and thus $X$ – centered) you examine the structure of $\hat{\beta} = \frac{\sum X' \times Y_{X'}}{\sum (X')^2}$.

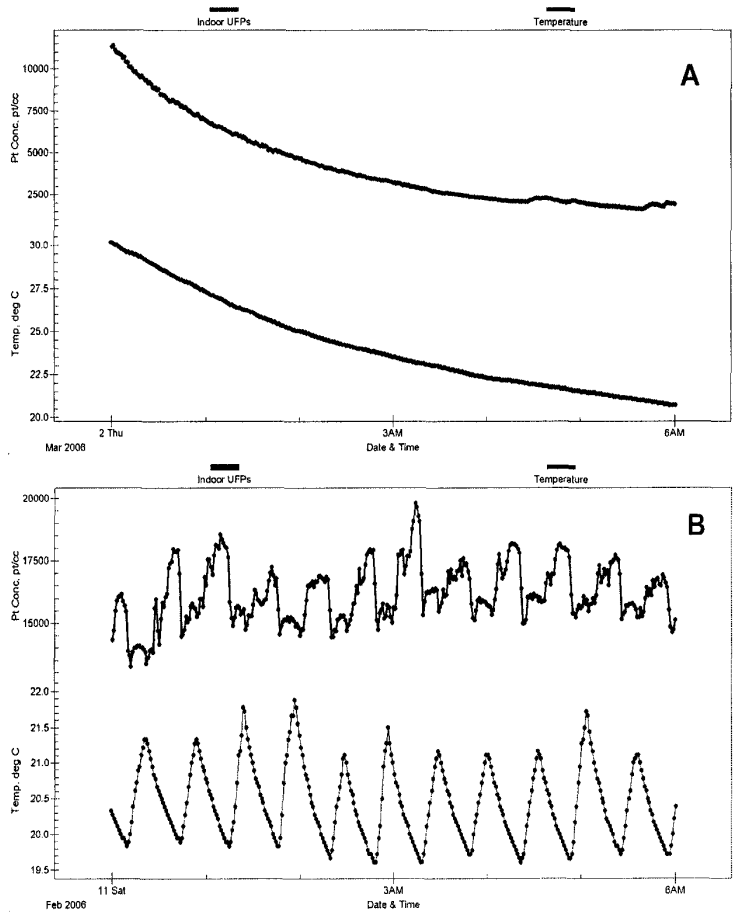



Figure 6a and 6b. Real-time overnight indoor UFP concentrations ($cm^{-3}$ (top line) and temperature (°C) (bottom line) as a sign of home heating system activation. (A) Home with a stand-alone wood stove, (B) Home with electric baseboard heaters.

[7] This situation is alluded to in this article.
[8] From the thesis work of Scott Weichenthal

.

# 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).[9]

   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, p15 in the notes above.] *Hint: it's simply a matter of using the definition of R.*

4. **[m-s]** Exercise in section 3 (p. 22 of notes) 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 [$0.396 \pm 0.024$], and weakest for the one between mother and daughter [$0.284 \pm 0.028$]. Given the way he obtained the measurements, can you imagine why this was? [10] [It was $0.302 \pm 0.027$ for mother & son; $0.360 \pm 0.026$ for father & daughter.]

7. ***Bridging the physical- and the psycho-metric:*** The notes on "Increasing Reliability by averaging several measurements" on the right hand column of page 12 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".[11]

   **[m-s]** Invert the formula on p.12 to derive the one on the right hand column of p.9 for the Spearman-Brown prediction formula relating the reliability of two versions of a test, one with N times more items than the other.

8. You are trying to estimate, from **imperfect observations** of $F$ and $C$, the values of the two coefficients $B_0$ and $B_1$ in the temperature relation $F = B_0 + B_1 \times C$.

   For each of the following situations, and using the true values $B_0 = 32$ and $B_1 = 9/5 = 1.8$, simulate[12] 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 $\epsilon_F \sim$ 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

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

   (b) $F$ measured perfectly, $C$ measured with $\epsilon_C \sim$ Gaussian$(\mu = 0, \sigma_{\epsilon_C} = 1)$ errors that are independent of $C$ [*Classical type* error: someone

[9] ToneCheck ( `https://techcrunch.com/2010/07/20/tonecheck/`) 'sounded' like an interesting tool; it's not clear if 'make it' commercially, or was bought by another company!

[10] After 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/pearson1930vol3ach14p17-18.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.

[11] Wikipedia has an entry called 'Spearman Brown prediction formula'.

[12] If 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.

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 $b_1$.

(c) $F$ measured perfectly, $C$ recorded from a thermostat that was 'set' to 14, 16, etc, where $\epsilon_C \sim$ Gaussian($\mu = 0$, $\sigma_{\epsilon_C} = 1$) around the set value[13] [*Berkson Error*]: you use the set values 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.[14]

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 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 $\text{ICC}_C$; so use the various amounts of error in C (ranging from $\sigma_{\epsilon_C} = 0$ to $\sigma_{\epsilon_C} = 22$) to calculate the various $\text{ICC}_C$'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
urlhttp://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:

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

(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`

[13] Figure 6b on page 23 suggests it would not be exactly Gaussian.

[14] The 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.

Write out the statistical model that you used to obtain this, and list any assumptions it makes.

(b) The estimate in (a) is an estimate of the 'within' observer variation.

In order to estimate the '*between*'-observer variation, what is the *minimal* information you would need from each of you co-observers? (since JH has access to all of them, he will supply each of them once you email him with your specific request: he can supply the full raw data that could be then put into a canned statistical routine, but he would prefer that you do the calculations 'from scratch' in `R`).

Again, write out the statistical model that you used to obtain this, and list any assumptions it makes.

(c) Here the 'objects' to be measured were 7 very specific (fixed) time-points. 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 you 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.

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 Berrnoulli'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?]

Here is the `R` code JH used to count the cases, and here is Edward's reply.

Wikipedia entry for Edwards.

One wonders what he thinks of the removal of the Latin Square from the dining hall window at Gonville and Caius College, Cambridge.

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:



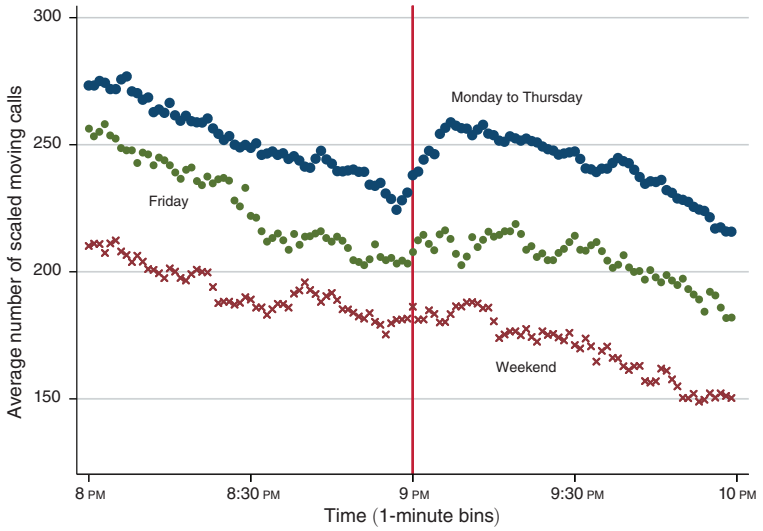



FIGURE 2. CELL PHONE CALL VOLUME FROM MOVING VEHICLES FOR CALIFORNIA FROM 8PM TO 10PM IN 2005

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

.



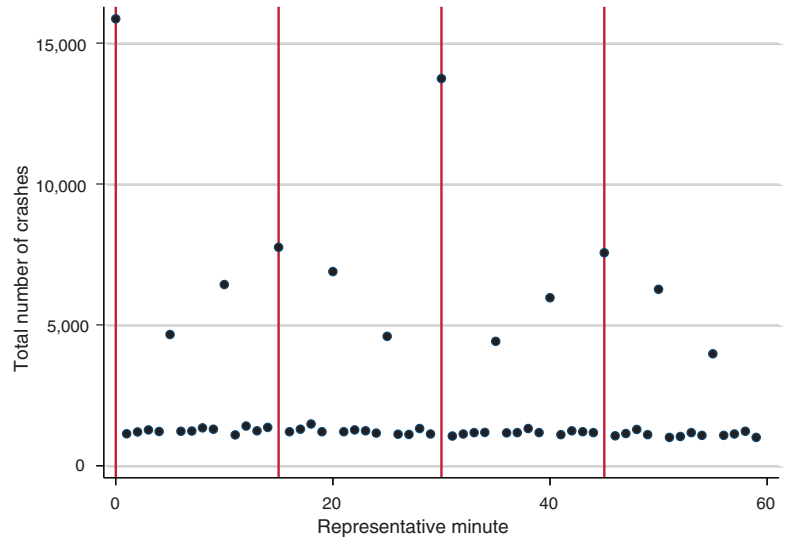



FIGURE 4. PERIODICITY IN SIDS CRASHES ACROSS REPRESENTATIVE HOUR IN 2005 FOR ALL STATES IN SAMPLE

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

14. **Galton's data more than century later**

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

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

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

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

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

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

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

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

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

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

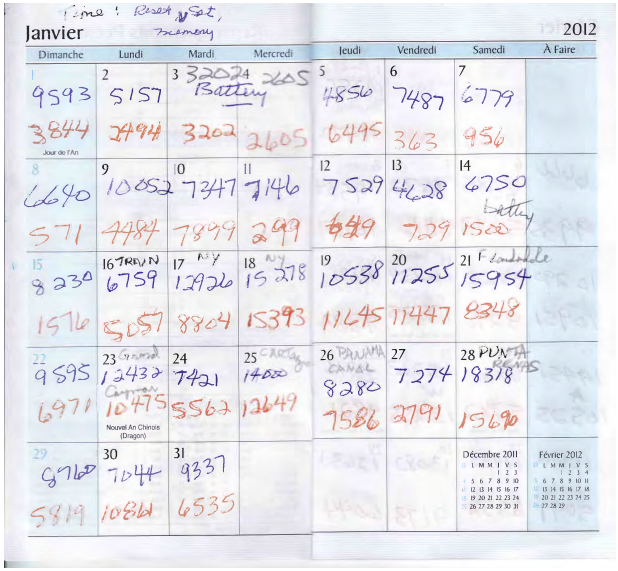
Here are the books assigned to the different students. Contact JH if your ID number is not in the list.

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

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

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

15. **Physical Activity: JH 2010-2017**


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 six files (2010-2011, 2012-2013 , 2014-2015, 2016-2017, 2018, 2019-2020) under the heading "Physical Activity: How many steps a day has JH being doing since 2010?" near the top of the Resources webpage. The 2010-2011 `.csv` file has the paired recordings for 2010, as well as JH's ones for 2011. The 2012-2013, 2014-2015, 2016-2017, 2018 and 2019-2020 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 over-simplified) 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...104) can be written as

$$y_{w,j} = \mu + \mathrm{b}_w + \epsilon_{w,j}$$

(b) For didactic purposes, treat the model as a random-effects one, i.e., with week as the random factor. Thus, the 104 $\mathrm{b}_w$'s are assumed to be a random sample drawn from a $N(0, \sigma^2_w)$ distribution.[15] Even though they may have a lot of structure, treat the variations across days within a week as uncorrelated 'disturbances' or 'errors' ($\epsilon_{yr,w.y,j}$) with variance $\sigma^2$ but no structure (i.e. treat all $\epsilon$'s as exchangeable, so that order of observations within the same week is irrelevant – in the file, you only need to know which week it is, not which day of the week. Clearly, there may be strong intra-week patterns, but for now assume that you are not even told which observation corresponds to which day of the week.

From the Expected Mean Squares (EMS) for this model[16]

[15] Using Roman b's and Greek $\beta$'s to distinguish random effects from fixed effects is a recent convention: it was not used when JH learned linear models.

[16] See also pages 4 and 5 of Notes on Introduction to Measurement Statistics, and pages 3 and 4 of the Notes on Quantifying Reliability (on the Resources website, under the heading 'Measurement – Lecture Notes, etc'). 'Weeks' in the current example correspond to 'persons' or 'subjects' or 'families' in those examples.

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

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

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

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

---

See p7-8, week to week variability in JH's average steps per day, 2010-11.

---

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']. as well as the `data and R code`.

(a) Redo the 2-way ANOVA 'with participant and trial as main effects' to see if you can reproduce the reported coefficient of variation.

(b) Use a 1-way ANOVA, with subjects as a random effect, and the 3 trials as replicates (i.e. ignoring the order) and calculate an overall coefficient of variation. [A very similar 1-way ANOVA is shown in the 1st column of page 6 of the 'Introduction to Measurement Statistics' Notes on the Resources website. Page 11 of the Notes 'Quantifying Reliability' has an example with 2 measurements per family, but the principle is the same.]
Which makes more sense to you, the CV based on their 2-way ANOVA, or yours based on a 1-way ANOVA?

(c) Calculate subject-specific coefficients of variation (just as was reported in Table 1 in the article on breath alcohol – the link to this article can be found just above the one for the endurance test). Summarize the 10 CVs using say the median and the range. Would you report the 'overall' CV the authors did, or some summary of the 10 subject-specific ones? Give a reason for your choice.

(d) Use the results of the 1-way ANOVA[17] to calculate an intra-class correlation (ICC).

(e) In this setting, which makes more sense, a CV or an ICC? Why?

(f) Rerun the ICC code several times on random subsets of the subjects. As you reduce the sample size to just 2 or 3, does the ICC stay stable? Use the example to say what the ICC tells us that the CV can not, and what the CV tells us that the ICC can not.

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

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

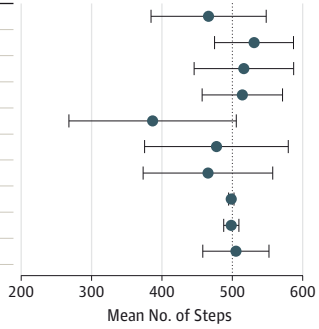
> This prospective study recruited healthy adults aged 18 years or older through direct verbal outreach at a university. Participants gave verbal informed consent to walk on a treadmill set at 3.0 mph for 500 and 1500 steps, each twice, for no compensation. An observer (M.A.C.) counted steps using a 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

[17] The R code supplied makes use of an ICC package, but it is always safer to check with a worked example that a package you don't know is doing what you want it to do.

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. Device Outcomes for the 500 Step Trials**

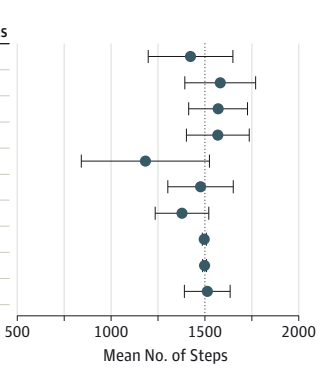
| Device | No. of Observations |
|---|---|
| Galaxy S4 Moves App | 27 |
| iPhone 5s Moves App | 28 |
| iPhone 5s Health Mate App | 28 |
| iPhone 5s Fitbit App | 28 |
| Nike Fuelband | 28 |
| Jawbone UP24 | 28 |
| Fitbit Flex | 28 |
| Fitbit One | 27 |
| Fitbit Zip | 27 |
| Digi-Walker SW-200 | 28 |




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

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

| Device | No. of Observations |
|---|---|
| Galaxy S4 Moves App | 28 |
| iPhone 5s Moves App | 28 |
| iPhone 5s Health Mate App | 27 |
| iPhone 5s Fitbit App | 27 |
| Nike Fuelband | 28 |
| Jawbone UP24 | 28 |
| Fitbit Flex | 28 |
| Fitbit One | 26 |
| Fitbit Zip | 27 |
| Digi-Walker SW-200 | 28 |




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



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.

(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 devise.[18] 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

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

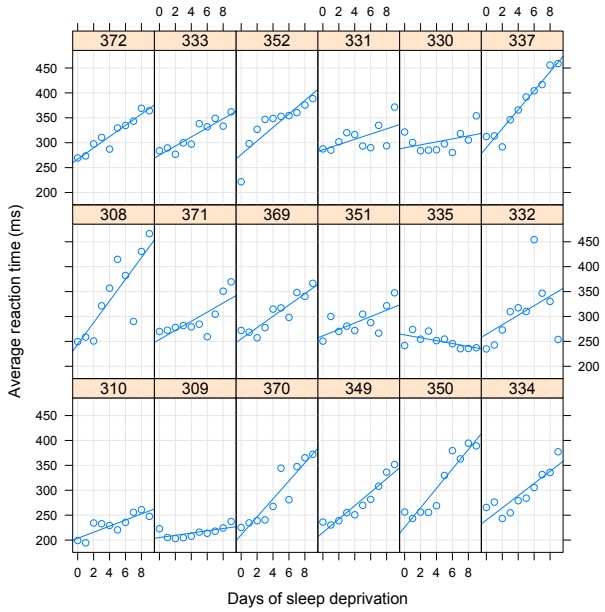
both iPhones and Androids. Fitbit is also good but works with a limited set of Androids.'

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

## 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 Sample Size Estimation for Random-effects Models: Balancing Precision and Feasibility in Panel Studies 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, 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.

To make your own measurements, you can choose this quite intuitive web tool `https://faculty.washington.edu/chudler/java/redgreen.html` – 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.

[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. Links to web-based tools, and to some scholarly and newspaper articles on reaction times) are 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 (CV[19]). 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 to be less than 5% of the mean. Using the SD (or SEM, or CV) you already obtained[20], 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).

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 $\frac{\overline{r.t._D} - \overline{r.t._{ND}}}{SE\ of\ this\ difference}$, and $\alpha = 0.05$ (2-sided).

[19] When reporting a CV, it is customary to do it so as a *percentage*

[20] Of 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

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.

Suppose[21] 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 further[22] 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.

[21] This 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.

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

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?

**Note re terminology**:

In the situation where $x$ = 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.

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

In the situation where $x$ = 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.

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

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 topics, you might consult JH's Ch. 4 notes, 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 in 2019 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'. As of 2020, one can find this for the iPhone, iPad, or iPod touch. Think about ways to test its validity and reproducibility.

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 )
```

    (a) Make a new variate `AGE.1999` from the `AGE.1986` variate.

    (b) 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.

    (c) 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.

    (d) 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 Stamler[23] 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.

    (a) Verify the calculations for 24-h Urinary Na excretion.

    (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 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.

[23] cited 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`.

(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 accelerometry 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(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?

22. **Measuring COVID-19 Antibody Seroprevalence in Santa Clara County, California in early April 2020**:

The following is adapted from the full report of April 11.

**Methods**: In early April, 2020, California investigators tested Santa Clara county residents for antibodies to SARS-CoV-2 using an immunoassay. Participants were recruited using Facebook ads targeting a representative sample of the county by demographic and geographic characteristics. They reported the prevalence of antibodies to SARS-CoV-2 in a sample of 3,330 people [2,718 adults and 612 children], adjusting for zip code, sex, and race/ethnicity. They also adjusted for test performance characteristics using 3 different estimates: (i) the test manufacturer's data, (ii) a sample of 37 positive and 30 negative controls tested at Stanford University, and (iii) a combination of both.
**Results**: The unadjusted (crude) prevalence of antibodies to SARS-CoV-2 in Santa Clara County was [50/3,330 =] 1.50% (exact binomial 95CI 1.11-1.97%) [A], and the population-weighted prevalence was 2.81% (95CI 2.24-3.37%) [B]. Under the three scenarios for test performance characteristics, the population prevalence of COVID-19 in Santa Clara ranged from 2.49% (95CI 1.80-3.17%) to 4.16% (2.58-5.70%) [C]. These prevalence estimates represent a range between 48,000 and 81,000 people infected in Santa Clara County by early April, 50- 85-fold more than the number of confirmed cases. [taken from April 11 report]

*The questions below are taken directly from the Part A (`bios700`) exam of August 4, 2020.*

*For questions a-f, ignore the misleading wording in A and B, which refer to the proportion of positive tests (there were 50 positive tests), not the prevalence of antibodies. (In the statistical analyses section, they use the term 'frequencies of positive tests as a proportion of the sample size.')*

(a) Explain how the exact binomial CI in [A] is calculated. How different is it in this instance from the usual 'non-exact' test?

(b) Give a reason why neither the 'exact' nor the 'inexact' CI around the 1.5% is relevant. (You might delay answering until you have been through the remaining questions)

(c) Explain in more detail how you think they arrived at [B], i.e., the prevalence of 2.81% and the 95% CI 2.24-3.37.

(d) Given that there are 57 zip codes, do you see any issues in calculating a SE? Explain.

(e) Suggest one other way of arriving at a point and interval estimate of the county percentage [incidentally, besides the non-representativeness of the sample with respect to zip code, sex, and race/ethnicity mismatches, the age distribution of the sample did not match that of the county either.]

(f) They say in the main text that their 2.24-3.37 CI around the 2.81% was computed 'without clustering the standard errors for members of the same household' and that it was 1.45-4.16 when this clustering was taken into account. Explain why the CI that takes account of the household clustering is the more appropriate of the two, why it is wider, and one way you would calculate it.

(g) In A and B, by using the terms 'test positivity' and 'prevalence' interchangeably, the authors are implicitly assuming that the test is perfect, i.e., always positive when antibodies are present (100% *se*nsitive), and always negative when antibodies are absent (100% *sp*ecific). In C, they "adjusted the prevalence for test sensitivity and specificity. Because the SARS-CoV-2 assays are new, [they] applied three scenarios of test kit sensitivity and specificity." The first scenario used the manufacturer's validation data (*se*nsitivity 91.8%; *sp*ecificity 99.5%).'

i. (Unrealistically) for now, take the given *se* and *sp* values as having negligible sampling error, and ignore the fact that the

sample of 3,330 is not representative of the county. Derive the estimating equation whose solution gives the Maximum Likelihood estimate of the population prevalence, $\theta$. (You don't need to solve the estimating equation.)

ii. You can get to the same estimating equation faster by the method of moments, so what is the advantage of going the ML route?

iii. In fact, the manufacturer's 91.8% sensitivity was the proportion 78/85, and the 99.5% specificity was 369/371. The smaller pilot dataset from Stanford gave $se = 25/37 = 67.6\%$ and $sp = 30/30 = 100\%$. How would you incorporate the inherent sampling error in $se$ and $sp$ into the CIs for $\theta$.

(h) The concepts of sensitivity and specificity are typically taught in the context of positive and negative predictive values of diagnostic tests. In this SARS-CoV-2 immunoassay context, what would predictive values refer to? And how does this focus differ from the focus of the Santa Clara study?

The first version of the report was the subject of a very extensive statistical blog (as well as many many politically-based) ones.

These resulted in a second version, that prompted another round of statistical attention, and even a full paper. This topic is an old one, and just-now-retired McGill biostatistics professor Lawrence Joseph played a pioneering role – well before prevalence estimation became so important. Interestingly, the references in the full paper don't go back many years.

For a newer [as of July 28, but this topic is moving fast!] report covering more space and time, see this JAMA article *Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020.*

You will see that a key driver remains the *sp*ecificity of the test, and its critical presence in both the numerator and denominator of the prevalence estimator.

As you will find by Googling say 'dashboard seroprevalence covid' there are now several dashboards, of variable quality, some showing international comparisons. Here is one. It is not always clear if all of the reported prevalences make the correction for the operating characteristics of the tests used, or how estimates are combined, or how calendar time is taken into account.

23. **A 165-year rewind: John Snow's data on cholera deaths in customers of two Water Companies, South London, Fall 1854, and his additional 1856 report**:

Of the internet 'shrines' to John Snow and his work elucidating the mode of transmission of cholera, the foremost, and the earliest is the UCLA site. Others are the one at Michigan State University, maintained by authors of the definitive biography of Snow, and the John Snow Society at the London School of Hygiene and Tropical Medicine.

On JH's site that he prepared for his lecture in the EBOH 2014 lecture series for the public, you will find further material and links.

While the full John Snow story is quite extensive, most courses limit their coverage to the Broad Street Pump episode. Interestingly, in his 1855 book, Snow gave this fewer pages (and less weight?) to this than he gave to the South London data, which arose from what he thought of as 'The Grand Experiment.' The account of this begins in what the UCLA site calls part 3 of the (now online) recreation of the book. Snow's own observations began in 1832 in the coal mines in northeastern England 1849, and it was during the 1949 epidemic that he published his short pamphlet. So you might want to start in 1849, and his section *Influence of the water supply on the epidemic of 1849, in London*, and how he exploited the fact of the 'New water supply of the Lambeth Company,' and the ' **Intimate mixture of the water supply of the Lambeth with that of the Southwark and Vauxhall Company**'.

For this exercise, focus on the 'Result of the inquiry as regards the first four weeks of the epidemic in 1854' – an inquiry he carried out by personally visiting the houses of the first 334 who died. The purpose was to learn which water company (or other source) they received their water from. The results are summarized in his Table VII. He went on to say:

> According to a return which was made to Parliament, the Southwark and Vauxhall Company supplied 40,046 houses from January 1st to December 31st, 1853, and the Lambeth Company supplied 26,107 houses during the same period; consequently, as **286 fatal attacks of cholera took place, in the first four weeks of the epidemic, in houses supplied by the former Company, and only 14 in houses supplied by the latter**, the proportion of fatal attacks to each 10,000 houses was as follows. Southwark and Vauxhall 71. Lambeth 5. **The cholera was therefore fourteen times as fatal at this period, amongst persons having the impure water**

**of the Southwark and Vauxhall Company, as amongst those having the purer water from Thames Ditton.**

The additional tables extend the study to the end of the year, but he considers the data from the first 4 weeks as the most trustworthy, and also as the ones that gave the sharpest contrast (the least diluted by information quality, and other factors: indeed the morality rate ratios declined as larger time windows are considered).

He returns to this data quality issue in his J Public Health, Oct. 1856 article. The main purpose of his article was to use newer and more fine-grained (district level) number-of-persons denominators than the overall number-of-houses denominators available to him when he completed the book at the end of 1854.

The main result was in Table V, which covered the entire 1854 epidemic,

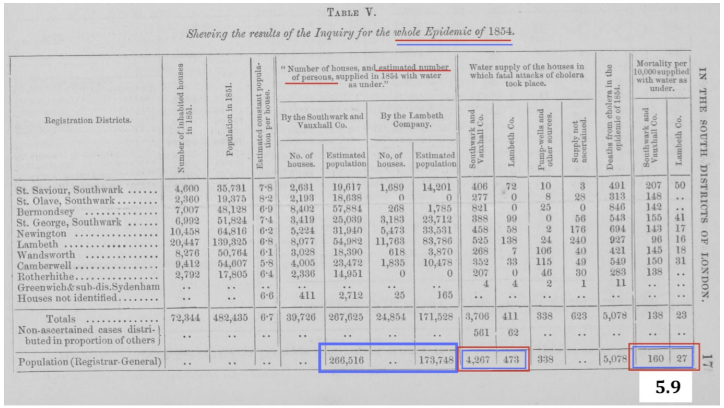


TABLE V.

*Shewing the results of the Inquiry for the whole Epidemic of 1854.*

| Registration Districts. | Number of inhabited houses in 1851. | Population in 1851. | Estimated constant population per house. | "Number of houses, and estimated number of persons, supplied in 1854 with water as under." By the Southwark and Vauxhall Co. No. of houses. | Estimated population | By the Lambeth Company. No. of houses. | Estimated population | Water supply of the houses in which fatal attacks of cholera took place. Southwark and Vauxhall Co. | Lambeth Co. | Pump-wells and other sources. | Supply not ascertained. | Deaths from cholera in the epidemic of 1854. | Mortality per 10,000 supplied with water as under. Southwark and Vauxhall Co. | Lambeth Co. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| St. Saviour, Southwark | 4,600 | 35,731 | 7·8 | 2,631 | 19,617 | 1,689 | 14,201 | 406 | 72 | 10 | 3 | 491 | 207 | 50 |
| St. Olave, Southwark | 2,360 | 19,375 | 8·2 | 2,193 | 18,698 | 0 | 0 | 277 | 0 | 8 | 28 | 313 | 148 | .. |
| Bermondsey | 7,007 | 48,128 | 6·9 | 8,402 | 57,884 | 268 | 1,785 | 821 | 0 | 25 | 0 | 846 | 142 | .. |
| St. George, Southwark | 6,992 | 51,824 | 7·4 | 3,419 | 25,039 | 3,183 | 23,712 | 388 | 99 | 0 | 56 | 543 | 155 | 41 |
| Newington | 10,458 | 64,816 | 6·2 | 5,224 | 31,940 | 5,473 | 33,531 | 458 | 58 | 2 | 176 | 694 | 143 | 17 |
| Lambeth | 20,447 | 139,325 | 6·8 | 8,077 | 54,982 | 11,763 | 83,786 | 525 | 138 | 24 | 240 | 927 | 96 | 16 |
| Wandsworth | 8,276 | 50,764 | 6·1 | 3,028 | 18,390 | 618 | 3,870 | 268 | 7 | 106 | 40 | 421 | 145 | 18 |
| Camberwell | 9,412 | 54,607 | 5·8 | 4,005 | 23,472 | 1,835 | 10,478 | 352 | 33 | 115 | 49 | 549 | 150 | 31 |
| Rotherhithe | 2,792 | 17,805 | 6·4 | 2,336 | 14,951 | 0 | 0 | 207 | 0 | 46 | 30 | 283 | 138 | .. |
| Greenwich & sub-dis. Sydenham | .. | .. | .. | .. | .. | .. | .. | 4 | 4 | 2 | 1 | 11 | .. | .. |
| Houses not identified | .. | .. | 6·6 | 411 | 2,712 | 25 | 165 | .. | .. | .. | .. | .. | .. | .. |
| Totals | 72,344 | 482,435 | 6·7 | 39,736 | 267,625 | 24,854 | 171,528 | 3,706 | 411 | 338 | 623 | 5,078 | 138 | 23 |
| Non-ascertained cases distributed in proportion of others | .. | .. | .. | .. | .. | .. | .. | 561 | 62 | .. | .. | .. | .. | .. |
| Population (Registrar-General) | .. | .. | .. | .. | 266,516 | .. | 173,748 | 4,267 | 473 | 338 | .. | 5,078 | 160 | 27 |

IN THE SOUTH DISTRICTS OF LONDON.

17

5.9

In his discussion, he shows a keen **insight** – It is just as relevant today, in the era of 'Big Data' – concerning the **consequences of errors in data**, in this case the ***addresses*** *of the persons in the Registrar General's list of deaths from cholera.*[24]

> [Big Data] **'can bear no comparison in point of accuracy to a personal inquiry, made on the spot, at the time of the epidemic.'**

It concerns the quality of the **numerators**, ie. how the cholera deaths might have been (mis)classified into the (wrong) water companies, especially as the addresses were collected by people who might have realized how critical they would be in the clolera story.

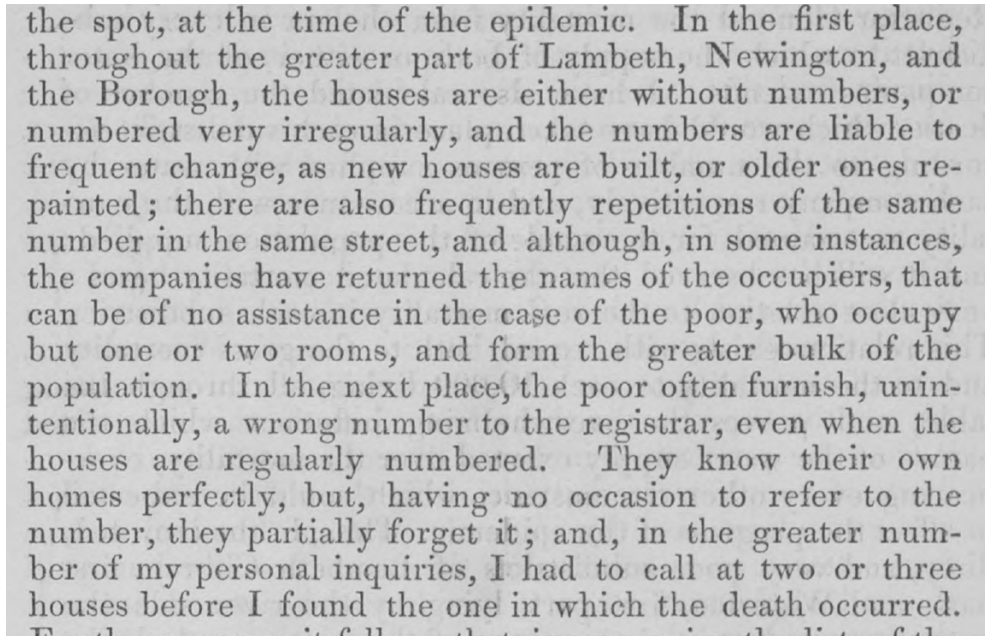
[24] Samantha.

As we recount in this recent item "The 25-page appendix to his 1855 book provided a detailed record of the 334 deaths from cholera that occurred in South London between 8 July and 5 August 1854. According to Snow, *this information was included 'as a guarantee that the water supply was inquired into, and to afford any person who wishes it an opportunity of verifying the results* (p. 80)'. 6 The information that John Snow recorded included the address at which each cholera-related death occurred, the date of death, the occupation and the age of the deceased, the duration of symptoms before death and the water source ( Figure 1 )."

In that item, we recount another little-known fact: when the members of the house could not supply evidence as to which was the water supplier, he used a 'high-tech' method to determine it (1st column, p. 1795).

But, even though it is difficult to imagine that this would be a big isue today, his 1856 article is **quite concerned with the numbers of wrong addresses.**

.



the spot, at the time of the epidemic. In the first place, throughout the greater part of Lambeth, Newington, and the Borough, the houses are either without numbers, or numbered very irregularly, and the numbers are liable to frequent change, as new houses are built, or older ones repainted; there are also frequently repetitions of the same number in the same street, and although, in some instances, the companies have returned the names of the occupiers, that can be of no assistance in the case of the poor, who occupy but one or two rooms, and form the greater bulk of the population. In the next place, the poor often furnish, unintentionally, a wrong number to the registrar, even when the houses are regularly numbered. They know their own homes perfectly, but, having no occasion to refer to the number, they partially forget it; and, in the greater number of my personal inquiries, I had to call at two or three houses before I found the one in which the death occurred.

and with the **consequences for the rate difference and the rate ratio**:

For these reasons it follows that, in comparing the lists of the water supply with the lists of deaths, many errors must have occurred; and as the deaths were six times as numerous in the houses supplied by the Southwark and Vauxhall Company as in those supplied by the Lambeth Company, the evident result would be that out of every six mistakes five would transfer a death from the former company to the latter, and only one would transfer a death from the latter company to the former. Another source of error, but operating

## ***Exercise*** **for** `bios601`

(a) Assume the numbers of deaths (4,267 and 473) in the columns of Table V are *correct*. Let P be the probability that the address of a diseased person is recorded correctly. Let it vary from perfect (1), to 'almost' (say 0.99), to 0.9 to 0.4. Assume also that the 'mixing' is so intimate that there is no 'clustering' of water supplier by street. Snow tells us that it was close to random:

> In the sub-districts enumerated in the above table as being supplied by both Companies, the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. ***In many cases a single house has a supply different from that on either side.*** Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies.
>
> The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity.
>
> To turn this grand experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur. I regret that, in the short days at the latter part of last year, I could not spare the time to make the inquiry; and, indeed, I was not fully aware, at that time, of the very intimate mixture of the supply of the two Water Companies, and the consequently important nature of the desired inquiry.

Show what the ratio of the expected mortality rates would be if the quality of the address reporting was as low as Snow suggested. Since it is difficult to convert '*in the greater number of my personal inquiries, I had to call at two or three houses before I found the one in which the death occurred*' into an exact probability, compute and plot/tabulate the ratio against the `P =c(1, 0.99, (9:4)/10)` values suggested above. Some might call this a *sensitivity analysis*.

(b) Assume the split of the numbers of deaths is *incorrect*, i.e., that they are *already distorted* by the errors in the addresses. Thus, your job is to reverse-engineer what the *correct* split of the 4740 is. Repeat part (a) but with the 'truth' starting at more extreme splits of the 4740 deaths than the 4,267:473 observed. If (say) P=0.9, what would the true split (and thus the mortality ratio) have to be to produce the 'observed' ratio?

(c) What principles can you draw from this assignment, and how well does the lesson fit with the 'regression dilution' examples earlier?

(d) Are these errors in the addresses '*differential*' or '*non-differential*'? Be careful with online definitions: the issues are subtle.

---

Snow mentions one other source of error – one he himself did not make in his 'shoe-leather' epidemiology. He made sure to enquire about the water supply in the house the person lived in *before* getting sick.

pany to the former. Another source of error, but operating to a less extent, is, that a number of persons who were attacked with cholera in houses supplied by the Southwark Company died in the workhouses of St. Saviour's, Lambeth, and Newington, which were supplied by the Lambeth Compan. It need excite no surprise, therefore, that the supple-

pan. It need excite no surprise, therefore, that the supplemental inquiry, embodied in the recent Report, instead of showing a mortality of 160 and 27 for the population supplied by the two water companies, or a difference of 6 to 1, showed a mortality of 125 and 37 per 10,000, or a difference of only $3\frac{1}{2}$ to 1. It must be obvious, however, independently of the above facts, that a difference of three and a-half to one would not explain the great difference in the mortality of the various districts and subdistricts. The epidemic of 1853 is included with that of 1854 in Mr. Simon's Report; but as there were but few deaths in 1853, and those chiefly amongst the population supplied by the Southwark Company, this circumstance would not much affect his results.

24. **A more modern (but ? similar type of) example**:

    Suppose you were asked to calculate the difference in the mean ages of death of the males and males in this dataset of 1,000 persons who died of Covid-19.

    The 1000 ages were easy to extract, and had this distribution. But the extraction of the names proved more difficult, and then (electronically – or even manually) converting them to male and female posed added some uncertainty. This link led JH to the `gender` package for `R`. Based just on the first name, it can produce a probability of being male for each person listed (of course, one could modify/override these based on any additional information contained in each 'blurb').

    How would you suggest these probabilities, $p_1$ to $p_{1000}$, (some of which are near/at zero and 1, and some are intermediate) to calculate a point and interval estimate of the difference in mean age at death?

25. **Mixing of the 'real' and the 'noise'**

    Refer to Fig 12.5 in the 'online book', section 12.7.2 (Measurement error), link here for a colour-based depiction of the mixing of 'true' and 'error' distributions.

    There, E has a '2-point distribution, namely -0.5 and +0.5, with equal probabilities.'

    Extend the diagram, and the VAR and ICC calculations, for the following E distributions:

    (a) 3-point distribution, namely -0.5, 0, and +0.5, with probabilities 1/4, 1/2, 1/4.

    (b) 5-point distribution, namely -1, -0.5, 0, +0.5 and +1, with probabilities 1/10, 2/10, 4/10, 2/10 and 1/10.

26. **'Berkson' error model**

    In your own words, explain why Berkson error in $X$ does not flatten the regression slope. Also, show it by algebra.

27. **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.

Endpieces

- **Type IV error**

*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 |
| | — to + | 5,92 | - | 7,3 | - | 5,58 |
| 5 | + to — | 5,9 | - | 7,5 | - | 5,65 |
| | — to + | 5,98 | - | 7,5 | - | 5,57 |
| 6 | m. to — | 3,03 | 2,9 | - | - | 5,53 |
| | — to + | 5,9 | 5,71 | | - | 5,62 |
| 7 | m. to — | 3,15 | 3,03 | 7,4 | 6,57 | 5,29 |
| | — to + | 6,1 | 5,9 | by mean. | | 5,44 |
| 8 | m. to — | 3,13 | 3,00 | | - | 5,34 |
| | — to + | 5,72 | 5,54 | - | - | 5,79 |
| 9 | + to — | 6,32 | - | 6,58 | - | 5,1 |
| 10 | + to — | 6,15 | - | 6,59 | - | 5,27 |
| 11 | + to — | 6,07 | - | 7,1 | - | 5,39 |
| 12 | — to + | 6,09 | - | 7,3 | - | 5,42 |
| 13 | — to + | 6,12 | - | 7,6 | - | 5,47 |
| | + to — | 5,97 | - | 7,7 | - | 5,63 |
| 14 | — to + | 6,27 | - | 7,6 | - | 5,34 |
| | + to — | 6,13 | - | 7,6 | - | 5,46 |
| 15 | — to + | 6,34 | - | 7,7 | - | 5,3 |
| 16 | — to + | 6,1 | - | 7,16 | - | 5,75 |
| 17 | — to + | 5,78 | - | 7,2 | - | 5,68 |
| | + to — | 5,64 | - | 7,3 | - | 5,85 |

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.

`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).

- **Scientific Method, Statistical Method and the Speed of Light**

– – – – – Link

- **Determinations of the parallax of the sun, the mean density of the earth, and the speed of light**

– – – – –Link

- **A Historical View of Statistical Concepts in Psychology and Educational Research**

– – – – – Stigler — Edgeworth (cited) — Peirce (cited)