University of York Department of Health Sciences

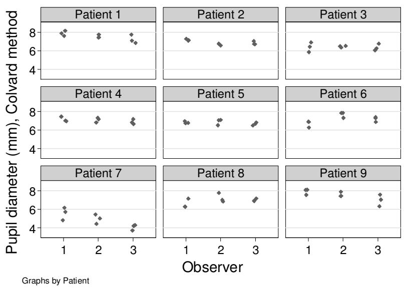
Measurement in Health and Disease

Observer variation

Martin Bland

http://martinbland.co.uk/

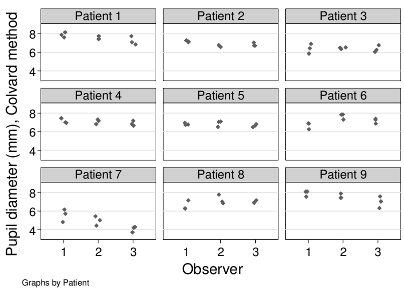
What do we mean by observer variation?



First 9 patients from a study of 28, each patient was measured 3 times by each of 3 observers.

More variation between observations by different observers than when the the same observer measures a patient.

How do we measure observer variation?

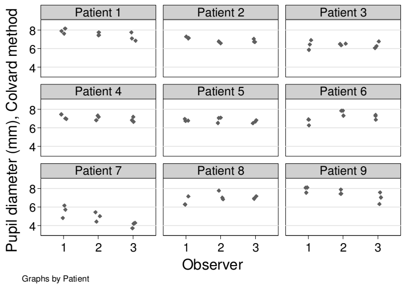


Within-subject standard deviation and coefficient of variation, and correlation coefficients (ICCs).

Intra-observer:
SD = 0.38 mm,
ICC = 0.80.

Inter-observer:
SD = 0.48 mm,
ICC = 0.72.

(all 28 subjects).



Standard deviation is greater with different observers and ICC is smaller.

Both reflect the greater error when different observers are used.

Why investigate observer variation?

Several reasons.

Sometimes focus of interest is the properties of the measurement method itself:

Sometimes the focus may be on the observers rather than the measurement method.

Why investigate observer variation?

Several reasons.

Sometimes focus of interest is the properties of the measurement method itself:

- ❖ to see whether a new measurement technique can be reproduced by a second investigator,
- ❖ to see whether some aspects of a measurement are more subject to observer variation than others,
- ❖ to estimate the extra variation in measurement which would occur in practice, using different observers drawn from the group who might use the method for clinical purposes.

Main problem is getting enough observers to have a reasonable sample to represent observers in general.

Why investigate observer variation?

Several reasons.

Sometimes the focus may be on the observers rather than the measurement method:

- we may need to train observers in a large investigation,
- we may wish to evaluate the benefits of training.

Purpose determines design.

Usual design: get several observers each to measure several subjects, preferably more than once.

Representative sample of observers, make repeated observations on each of a sample of subjects, the order being randomized.

1. can rarely obtain a representative sample of observers.
2. many measurements which involve subjective assessment cannot be repeated by the same observer without the result of the first measurement influencing the second.
3. many methods of measurement are either uncomfortable or invasive, and a long series of measurements cannot be done on the same subject.

Most observer comparison studies are a compromise between the ideal study design and practical and ethical limitations.

Other possibilities:

- could have every subject measured by two different observers, using new observers every time,
- carry out several small replicates of the ideal design and then combine them,
- physical model or video recording.

Several small replicates of the ideal design combined:

Ultrasound abdominal circumference measurements (cm) by 16 observers (L. Chitty, personal communication)

Observer	Subject 1			Subject 2			Subject 3		
1	13.6	13.3	12.9	14.7	14.8	14.7	17.1	17.1	18.3
2	13.8	14.2	13.2	14.9	14.1	14.5	17.2	17.5	17.6
3	13.2	13.1	13.1	14.5	14.2	13.8	16.3	15.2	16.1
4	13.7	13.7	13.4	14.4	14.3	13.6	16.8	16.8	17.5
	Subject 4			Subject 5			Subject 6		
5	14.8	14.6	14.8	18.3	18.5	18.5	12.6	12.6	12.4
6	14.9	14.4	14.2	17.4	17.9	17.0	12.3	12.1	12.1
7	14.3	14.4	14.3	17.7	17.0	18.3	12.5	12.2	12.6
8	13.8	14.1	14.1	17.4	17.9	16.4	13.0	12.6	12.7
	Subject 7			Subject 8			Subject 9		
9	12.4	11.7	11.6	16.0	16.0	16.2	11.3	11.6	10.7
10	11.5	12.5	12.8	16.1	15.8	15.4	9.7	10.2	9.8
11	14.6	12.7	11.5	16.7	16.5	16.2	10.7	10.3	9.8
12	13.5	13.4	12.5	17.0	16.6	17.2	10.9	11.2	11.3
	Subject 10			Subject 11			Subject 12		
13	14.3	14.4	14.8	15.6	15.9	16.1	20.2	20.9	21.1
14	14.3	15.5	14.6	15.7	15.0	16.5	20.1	20.7	20.9
15	14.6	14.8	15.4	16.3	16.1	15.6	19.2	20.0	20.0
16	14.1	14.6	13.7	14.4	15.1	15.2	20.5	20.5	21.1

Several small replicates of the ideal design combined.

Ultrasound abdominal circumference measurements (cm) by 16 observers (L. Chitty, personal communication)

Inter-observer: SD = 0.62, ICC = 0.95.
Intra-observer: SD = 0.46, ICC = 0.97.

Analysis of observer variation studies

Pupil diameter (mm) measured by 3 observers on 28 subjects, left eye

Patient	Obs 1			Obs 2			Obs 3		
1	8	7.5	8	7.5	7.5	7.5	8	7	7
2	7	7	7	7	6.5	7	7	6.5	7
3	6.5	6	7	6.5	6.5	6.5	6	6	6.5
4	7	7.5	7	7	7	7	7	7	6.5
5	7	7	7	6.5	7	7	7	6.5	7
6	7	6.5	7	8	8	7.5	7.5	7.5	7
7	5.5	5	6	5	5.5	4.5	4.5	4	4.5
8	6.5	6.5	7	7	7	7.5	7	7	7
9	8	8	7.5	7.5	8	7.5	6.5	7	7.5
10	6.5	6.5	6.5	6	6.5	6.5	7	7	7
11	8	8	8	8	8	8	8	8.5	8.5
12	7	7	7.5	7.5	7	7	7	7	7
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
27	6	6	6	6	6	6	6	6.5	6.5
28	7	7	7	7	7	7	6	6	5.5

Analysis of observer variation studies

To estimate the increase in variation when different observers are used, we use analysis of variance.

Compared to the simple measurement error problem, the analysis of variance is more complicated, because we have more sources of variation.

- repeated observations by the same observer on the same subject: s_w^2,
- between subjects, that is between the true values of the quantity being measured: s_b^2,
- between observers (an observer consistently measures higher or lower than others): s_o^2,
- heterogeneity (where the observer measures higher than others for some subjects and lower for others): s_h^2.

Analysis of observer variation studies

Hence we have four different variances, and if we have measurements on different subjects made by different observers, the variance will be the sum of all of them:

$s^2 = s_b^2 + s_o^2 + s_h^2 + s_w^2$

s_b^2 is the variance between subjects, i.e. between the true values for subjects,

s_o^2 is the variance between observers, i.e. between the average measurements made by different observers,

s_h^2 is the variance between different observers on different subjects, over and above the variance between the average values of the observers and of the subjects,

s_w^2 is the variance of observations by one observer on one subject.

Analysis of observer variation studies

Hence we have four different variances, and if we have measurements on different subjects made by different observers, the variance will be the sum of all of them:

$s^2 = s_b^2 + s_o^2 + s_h^2 + s_w^2$

These four variances are called the **components of variance**.

We shall assume that all the errors, between observers, the heterogeneity, and the measurement error, are independent of one another and of the magnitude of the measurement.

If we do not assume this, we cannot estimate the errors.

We shall also assume that they follow a Normal distribution.

Analysis of observer variation studies

We can estimate the components of variance by analysis of variance, which is straightforward provided we have the same number of repeated measurements by each observer on each subject.

For the pupil diameter data, the anova table is:

Source	Partial SS	df	MS	F	Prob > F
Subject	153.74107	27	5.69411376	39.31	0.0000
Observer	3.43056	2	1.71527778	11.84	0.0000
Sub × Obs	19.62500	54	0.36342593	2.51	0.0000
Residual	24.33333	168	0.14484127		
Total	201.12996	251	0.80131458		

Analysis of observer variation studies

Source	Partial SS	df	MS	F	Prob > F
Subject	153.74107	27	5.69411376	39.31	0.0000
Observer	3.43056	2	1.71527778	11.84	0.0000
Sub × Obs	19.62500	54	0.36342593	2.51	0.0000
Residual	24.33333	168	0.14484127		
Total	201.12996	251	0.80131458		

Expected values of mean squares in a two-way analysis of variance table for *o* observers each measuring *n* subjects *m* times:

Source of variation	Degrees of freedom	Mean square
Total	$mno-1$	
Subjects	$n-1$	$mos_b^2 + ms_h^2 + s_w^2$
Observers	$o-1$	$mns_o^2 + ms_h^2 + s_w^2$
Subjects × observers	$(n-1)(o-1)$	$ms_h^2 + s_w^2$
Residual error	$(m-1)no$	s_w^2

Analysis of observer variation studies

The components of variance are found as follows:

$s_w^2 = 0.14484127$, $s_w = 0.38$

$s_h^2 = (0.36342593 - 0.14484127)/3 = 0.07286155$, $s_h = 0.27$

$s_o^2 = (1.71527778 - 0.36342593)/(3\times28) = 0.01609347$, $s_o = 0.17$

$s_b^2 = (5.69411376 - 0.36342593)/(3\times3) = 0.59229865$, $s_b = 0.77$

Analysis of observer variation studies

$s_w^2 = 0.14484127$, $s_h^2 = 0.07286155$,

$s_o^2 = 0.01609347$, $s_b^2 = 0.59229865$.

Intra-observer:

Within-subject standard deviation is $\sqrt{0.14484127} = 0.38$.

Intraclass correlation coefficient is

$$ICC = s_b^2/(s_b^2 + s_w^2)$$
$$= 0.59229865/(0.59229865 + 0.14484127) = 0.80.$$

Analysis of observer variation studies

$s_w^2 = 0.14484127$, $s_h^2 = 0.07286155$,

$s_o^2 = 0.01609347$, $s_b^2 = 0.59229865$.

Inter-observer:

Within-subject standard deviation is

$$\sqrt{(0.01609347 + 0.07286155 + 0.14484127)} = 0.48.$$

Intraclass correlation coefficient is

$$ICC = s_b^2/(s_b^2 + s_o^2 + s_h^2 + s_w^2) =$$
$$0.59229865/(0.59229865 + 0.01609347 + 0.07286155 + 0.14484127) = 0.72.$$

Analysis of observer variation studies

$s_w^2 = 0.14484127$, $s_h^2 = 0.07286155$,

$s_o^2 = 0.01609347$, $s_b^2 = 0.59229865$.

Intra-observer:

Within-subject standard deviation = 0.38.

Intraclass correlation coefficient = 0.80.

Inter-observer:

Within-subject standard deviation = 0.48.

Intraclass correlation coefficient = 0.72.

Analysis of observer variation studies

```
    Source | Partial SS   df       MS           F    Prob > F
-----------+--------------------------------------------------
   Subject | 153.74107    27  5.69411376   39.31   0.0000
  Observer |   3.43056     2  1.71527778   11.84   0.0000
 Sub x Obs |  19.62500    54  0.36342593    2.51   0.0000
  Residual |  24.33333   168  0.14484127
-----------+--------------------------------------------------
     Total | 201.12996   251  0.80131458
```

Note that both main observer effect and the eye times observer interaction (heterogeneity) are highly significant (P<0.0001).

Checking assumptions

Same assumptions as for measurement error:

For within-subject standard deviation:

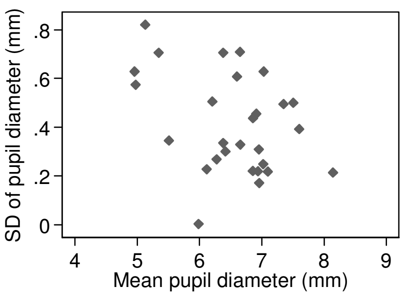
- subject standard deviation is independent of the mean,
- distribution within the subject is approximately Normal,

For correlation:

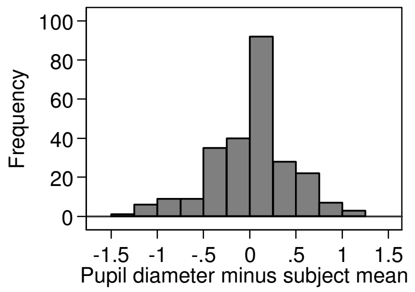
- representative sample
- Normal distribution for the measurement itself

We can check these assumptions graphically.

Plot standard deviation against mean:

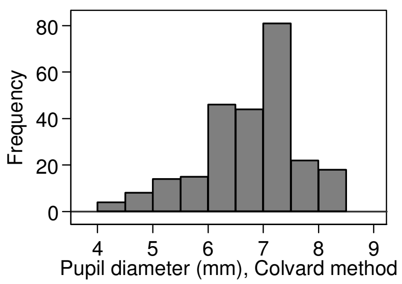


Observation minus subject mean:



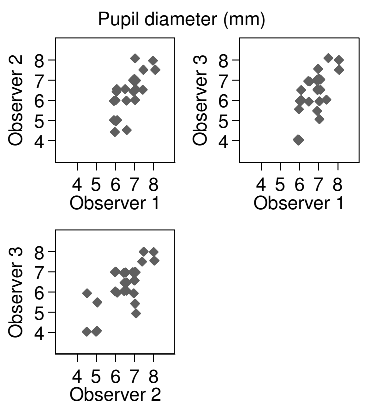
Symmetrical, peak a bit high, but not too bad.

Histogram of all observations



Negatively skew.

Scatter plots for observers, first measurement:



The necessary assumptions are not met for these data and the estimates obtained can only be approximate.