Theory of Measurement

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Australia

Overview

- Some concepts of measurement
- Methods for assessment of reliability
- Flawed methods
- Consequences of measurement error
- Control of measurement error
- Summary

Three concepts about science

Classificatory – Phân loại

place objects within a certain class

Comparative – So sánh

relationships between objects (warmer/cooler)

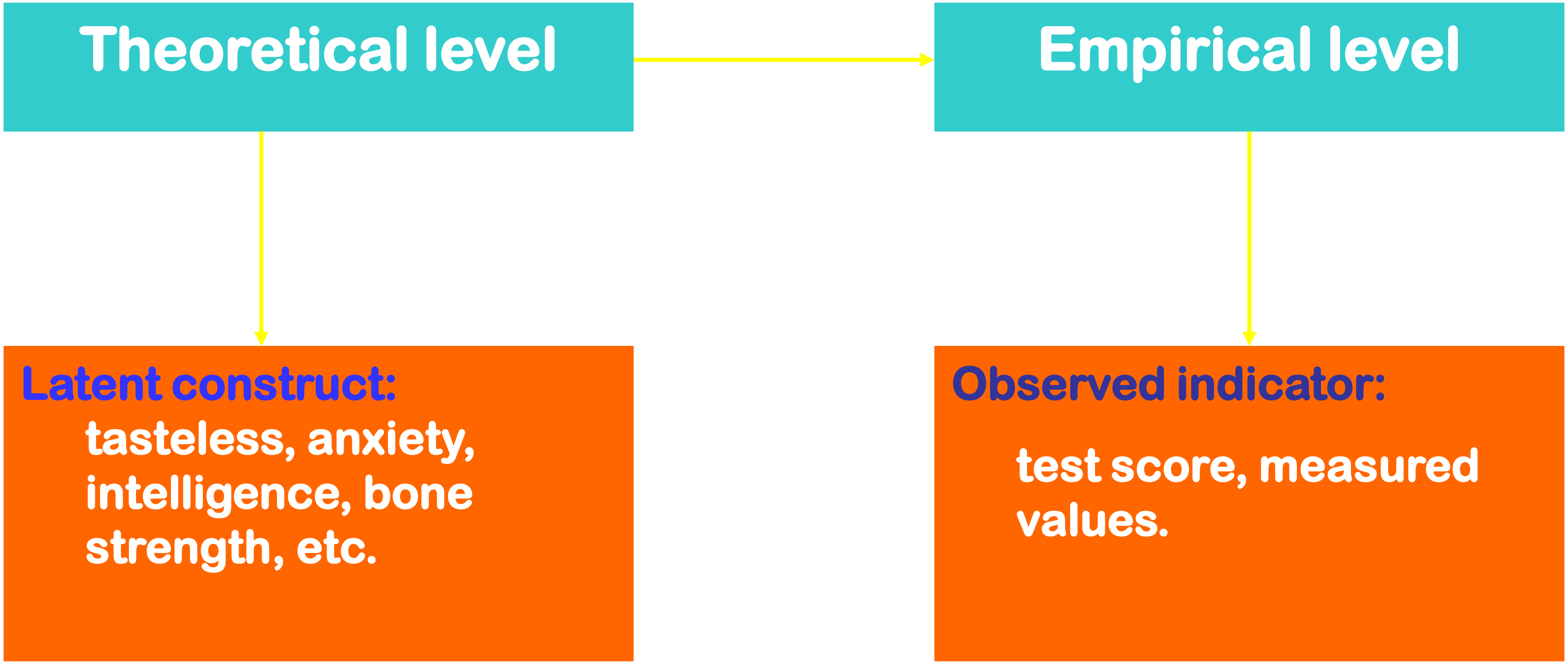
Prediction – Tiên đoán

evolution from the comparative concept

The criteria of “science”

Science	**Pseudoscience**
Logic, experimental evidence	Belief, loyalty
Results are **repeatable**	Results are *not* repeatable
Falsiability	Not falsifiable
Peer-reviewed journals	Not in peer reviewed journals
Evolution / learn from mistakes	Constant, unchanged belief

Theoretical vs empirical level
(Lí thuyết và thực nghiệm)



Theoretical level

Empirical level

Latent construct:

tasteless, anxiety, intelligence, bone strength, etc.

Observed indicator:

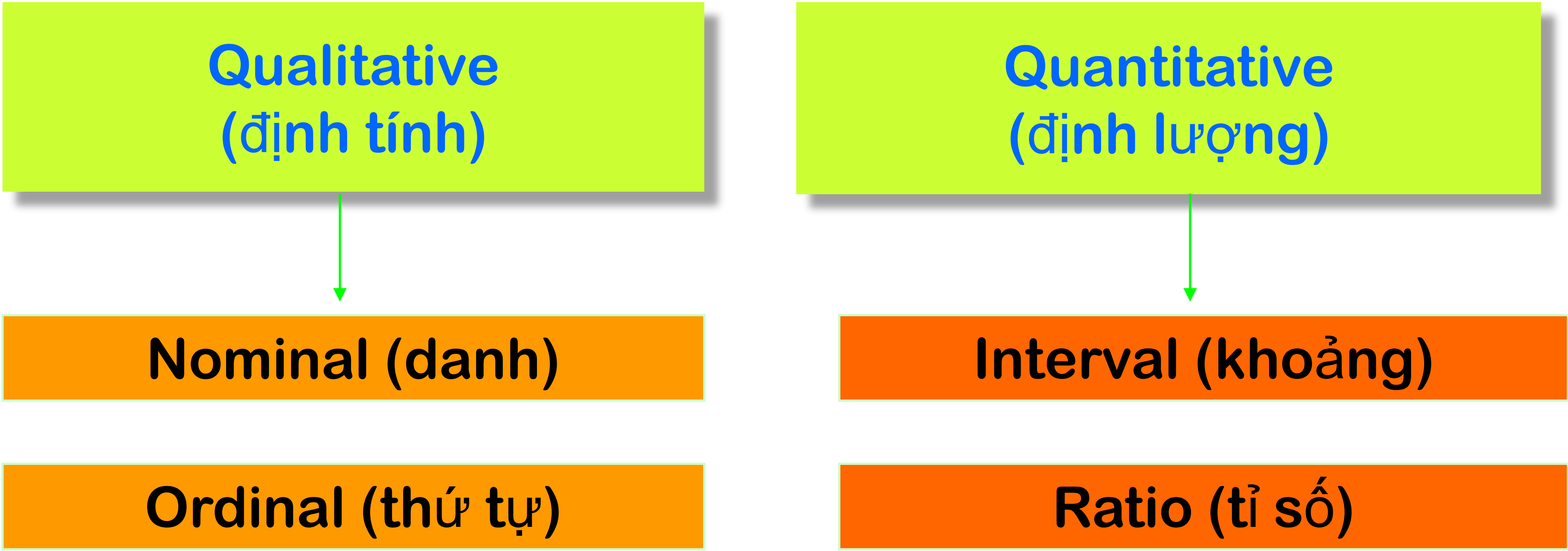
test score, measured values.

Measurements

- **The assigning of numbers to the values of a variable**
  (SS Stevens, Science 1946;103:677-80)

- **Rules specify procedures to assign numbers to values**

Types of measurement



Qualitative (định tính)

Nominal (danh)

Ordinal (thứ tự)

Quantitative (định lượng)

Interval (khoảng)

Ratio (tỉ số)

Qualitative measurements

Nominal level

- Classification
- A set of objects can be classified into exhaustive, mutually exclusive and unique symbol
- *Ex: religion, sex, location, etc*

Ordinal level

- Classification + Ordering
- A set of numbers can be assigned rank values and nothing more.
- *Ex: socio-economic status, education, levels of satisfaction, bitterness, etc*

Quantitative measurements

Interval level

- Classification + Ordering + Standard distance
- A set of objects can be described by units that indicate how far one case is from another case
- *Ex: temperature*

Ratio level

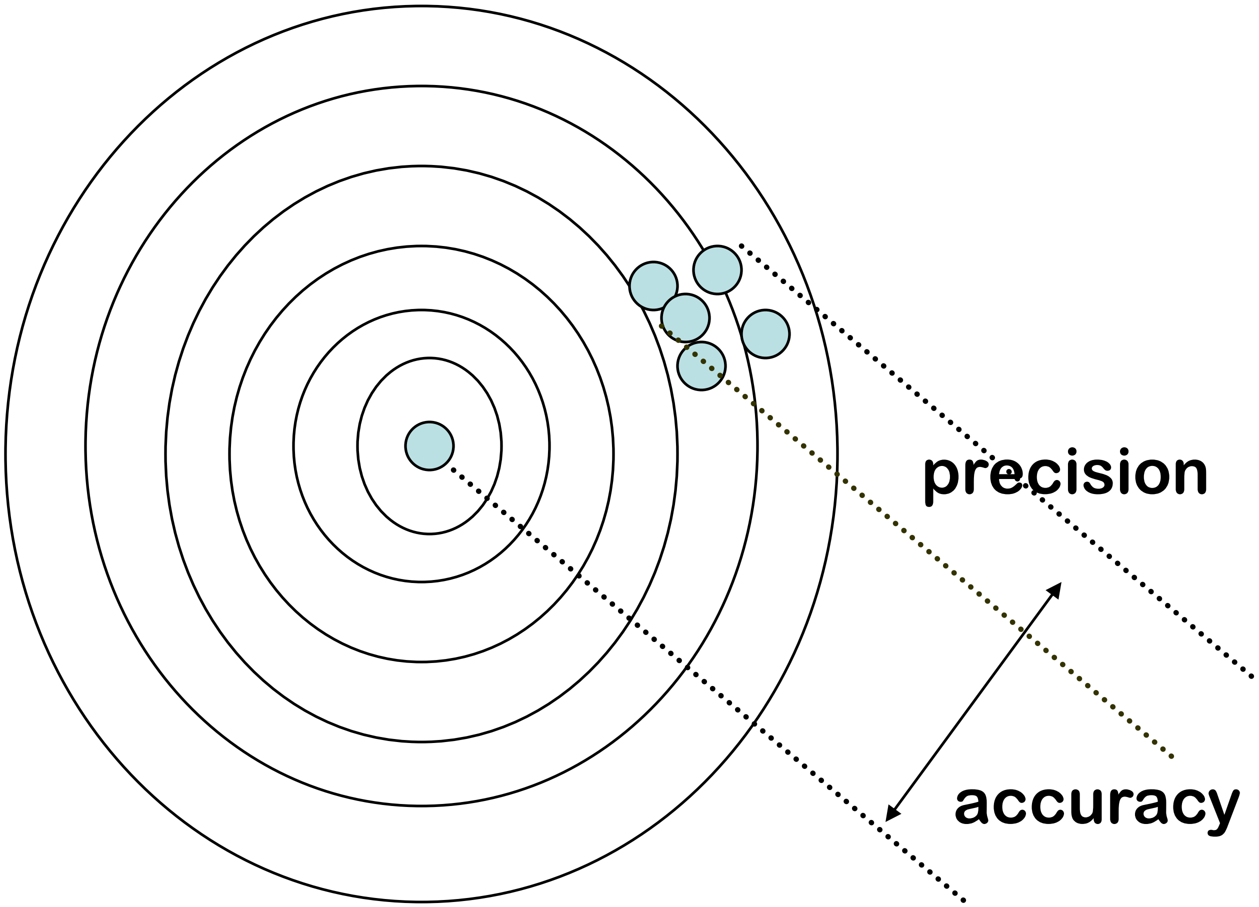
- Classification + Ordering + Standard distance + Natural zero
- Quantitative variable with natural zero
- *Ex: income, age, weight, bone mineral density*

Criteria of measurements

- **Validity** measures what it purports to
- **Accuracy** - the degree of “truthfulness” of an attribute that is being measured.
- **Reliability** (consistency and repeatability)
- **Sensitivity** to important variation

Accuracy vs reliability (precision)

Tính chính xác và độ tin cậy



Measurement error decreases the accuracy of measurement

Evaluation of reliability

Reliability (repeatability, reproducibility)

- **Stability**. Degree of stability exhibited when a measurement is repeated under identical conditions
- **Equivalence.** Same results by different operators

Evaluation of reliability

Validity (*validus = strong*)

- **Constructive validity**. The extent to which the measurement corresponds to theoretical concepts (constructs). *Ex: Bone density changes with advancing age.*
- **Content/Face validity**. The extent to which the measurement incorporates the domain of the phenomenon under study. *Ex: functional health status should encompass activities of daily living, occupation, family, etc.*
- **Criterion validity**. The extent to which the measurement correlates with an external criterion of the phenomenon under study. *Ex: academic aptitude test is validated against subsequent academic performance.*

Assessment of Reliability

Questions of interest

Example: A patient has bone mineral density (BMD) of 0.75 g/cm^2, is considered osteoporotic, and treated with Alendronate. After two months, BMD is 0.80 g/cm^2.

- How reliable is the measurement?
- What is the "true" baseline BMD?
- How large should a change be, to be sufficient certain that a true change did occur?
- How can reliability be improved?

Statistical indices of reliability

Quantitative

- Standard error of measurement (độ sai chuẩn)
- Coefficient of variation (hệ số biến thiên)
- Coefficient of reliability (hệ số tin cậy)
- Coefficient of concordance (hệ số đồng hợp)
- Limit of agreement (giới hạn đồng nhất)

Qualitative

- Kappa statistic
- Cronbach’s alpha coefficient
- Coefficient of concordance
- Intraclass correlation coefficient (hệ số phương sai trong một đối tượng)

Estimation of reliability: quantitative measurements

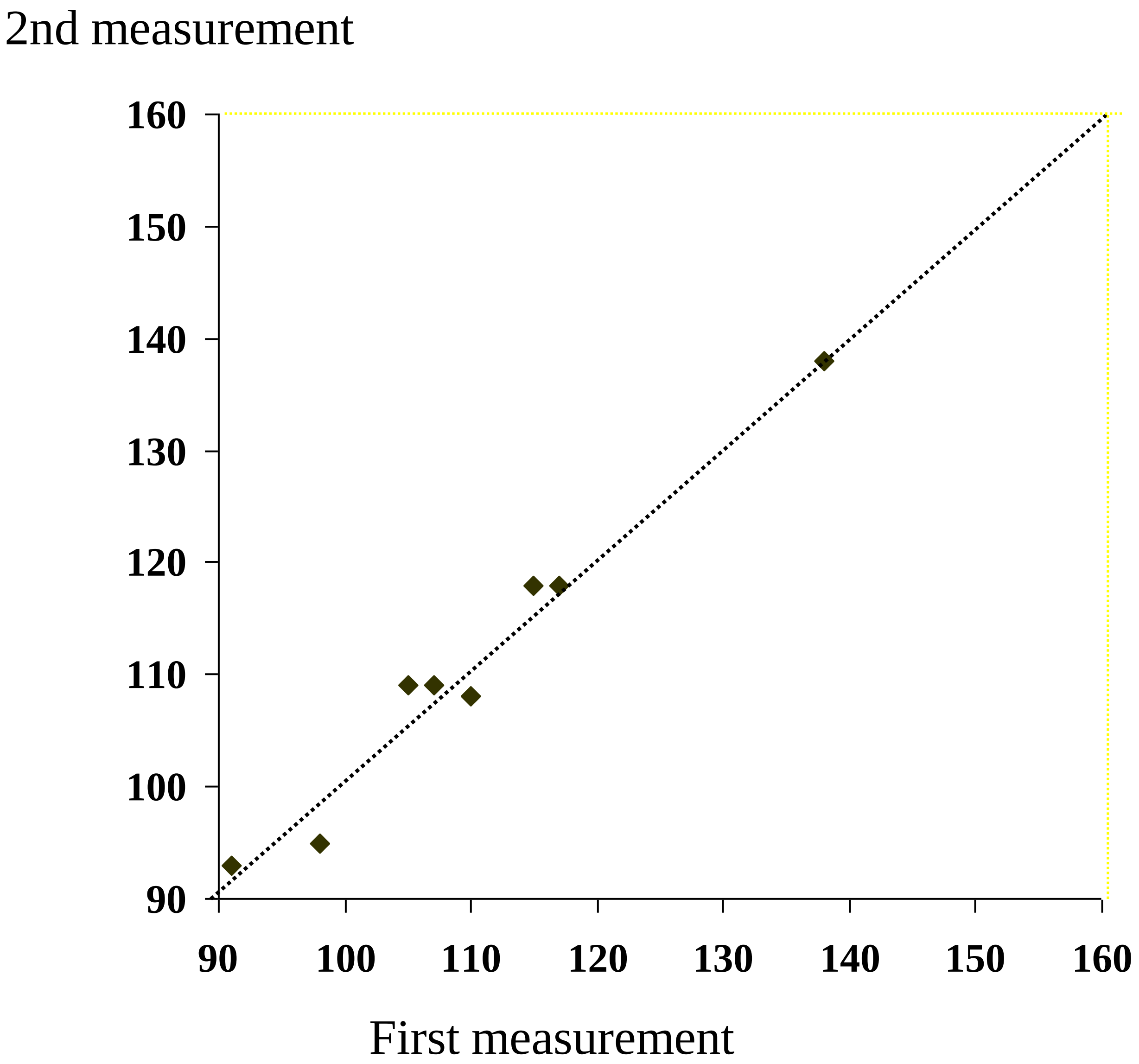
General case

	Measurement			
Patient	1	2	. . . .	k
1	x_{11}	x_{12}	· · ·	x_{1k}
2	x_{21}	x_{22}	· · ·	x_{2k}
3	x_{31}	x_{32}	· · ·	x_{3k}
.				
.				
.				
N	x_{n1}	x_{n2}	· · ·	X_{nk}

Bone mineral density

Patient	First	Second
1	117	118
2	115	118
3	110	108
4	91	93
5	138	138
6	85	90
7	107	109
8	110	108
9	98	95
10	105	109

Plot of 1st and 2nd measurements



Estimation of reliability: quantitative measurements

Bone mineral density

Patient	First	Second	Mean	Variance
1	117	118	117.5	0.5
2	115	118	116.5	4.5
3	110	108	109.0	2.0
4	91	93	92.0	2.0
5	138	138	138.0	0.0
6	85	90	87.5	12.5
7	107	109	108.0	2.0
8	110	108	109.0	2.0
9	98	95	96.5	4.5
10	105	109	107.0	8.0
Mean	**107.6**	**108.6**	**108.1**	**3.8**

Standard error of measurement (SEM)

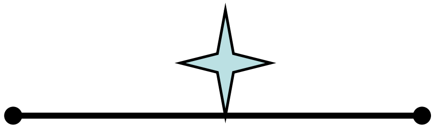
$$SEM = \sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i^2}$$

$$SEM = \sqrt{3.8} = 1.95$$

n = number of subjects

s_i^2 = intra-subject variances

Interpretation: The difference between a subject's measurement and the "true" value would be expected to be less than 1.96x1.95 = 3.8 for 95% of observations.



Coefficient of variation (CV)

Let X be the overall mean, and S be the within-subject standard deviation.

In our case: X = 108.1, S = sqrt(3.8) = 1.95

Coefficient of variation

CV = S / X

= 1.95 / 108.1

= 1.8%

Coefficient of variation (CV): interpretation

CV = 1.8%

- All variability between repeated measurements within a subject is 1.8%?
- Assuming Normality:
  - 68% of the differences between measurements lie within 1.8% of the mean;
  - 95% of the differences between measurements lie within 1.8x2 = 3.6% of the mean

Limits of agreement (LoA)

Assumption: Individual differences are Normally distributed.

Concept: The variability of reproducibility (intrasubject difference) for individual subjects may be expressed as 95% CI of the difference.

$$LoA = \bar{x}_d \pm 1.96 S_d$$

Limit of Agreement: estimation

Bone mineral density

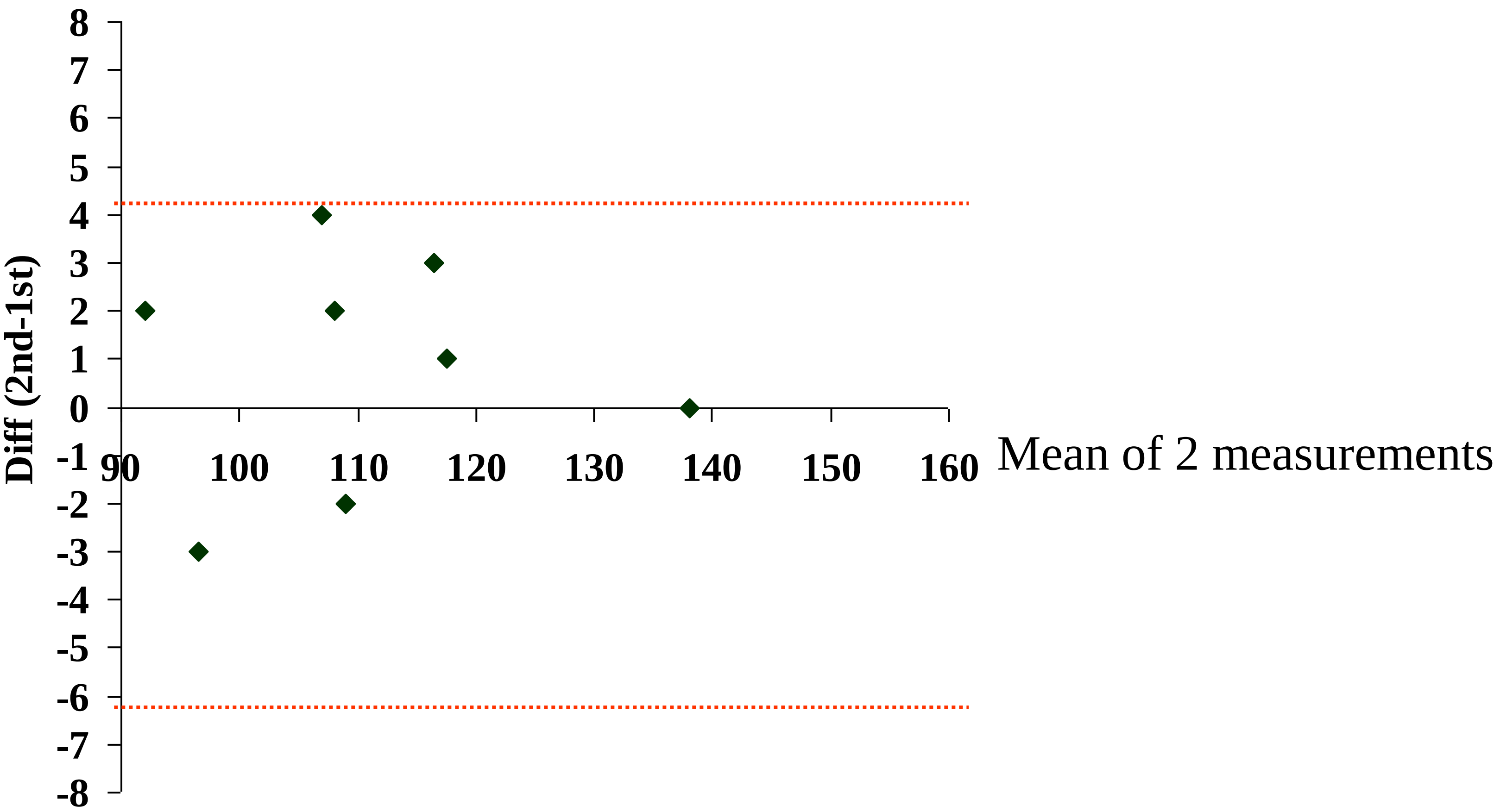
Patient	First	Second	Difference
1	117	118	-1
2	115	118	-3
3	110	108	+2
4	91	93	-2
5	138	138	0
6	85	90	-5
7	107	109	-2
8	110	108	+2
9	98	95	+3
10	105	109	-4
Mean	107.6	108.6	-1
SD	14.8	14.2	2.7

$$LoA = -1 \pm 1.96(2.7)$$

$$= -6.3 \text{ to } +4.3$$

The repeated BMD measurements may be 6.3 below or 4.3 above an average value for a subject.

Bland-Altman plot



Coefficient of reliability: concept

Observed score = “True” score + Random Error

$$X = T + E$$

$$Var(X) = Var(T) + Var(E)$$

Coefficient of reliability

$$\boldsymbol{R = var(T) \,/\, var(X)}$$

It measures the correlation between the “true” and observed values.

Estimation of reliability coefficient

Analysis of variance

Source	variance
Between patients	206.3
Within patients	3.8

$Var(T) = 206.3$

$Var(E) = W = 3.8$

$R = 206.3 / (206.3 + 3.8)$

$= 0.98$

Coefficient of concordance: concept

- Take into account the difference in means between first and second measurements

$$C = \frac{2Cov\left(x_1, x_2\right)}{s_1^2 + s_2^2 + \left(\bar{x}_1 - \bar{x}_2\right)^2}$$

$Cov(x_1,x_2)$: Covariance between 1st and 2nd measurements

s_1, s_2 : Standard deviation of 1st and 2nd measurements.

$Xbar_1$, $Xbar_2$: sample means

Coefficient of concordance: concept

10 judges were asked to score the bitterness of a wine twice.

Judge	1st time	2nd time
1	76	78
2	72	74
3	60	60
4	80	76
5	87	83
6	75	80
7	78	76
8	81	79
9	74	74
10	69	72

Sample statistics

	1st	2nd
Means:	75.2	75.2
SD:	7.3	6.2

Covariance = 41.9

$$\frac{2(41.9)}{7.3^2 + 6.2^2 + (75.2 - 75.2)^2}$$

$= 0.90$

Kappa: a measure of reliability for qualitative measurements

- Two judges score an attribute
- The scores are categorical: A, B and C.
- The outcomes may be summarized as follows

Judge 2’ s scores	Judge 1’ s scores			Total
	A	B	C	
A	n_{11}	n_{12}	n_{13}	$N_{1.}$
B	n_{21}	n_{22}	n_{23}	$N_{2.}$
C	n_{31}	n_{32}	n_{33}	$N_{3.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

Kappa

- Proportion of agreement: $$P_A = \frac{n_{11} + n_{22} + n_{33}}{N}$$
- Proportion of change agreement: $$P_C = \frac{(n_{.1} \times n_{1.}) + (n_{.2} \times n_{2.}) + (n_{.3} \times n_{3.})}{N^2}$$
- Kappa statistic $$\kappa = \frac{P_A - P_C}{1 - P_C}$$
- Variance of κ $$\operatorname{var}(\kappa) = \frac{P_C + P_C^2 - \sum_{i=1}^{3} \left(\frac{n_{i.}^2 n_{.i} + n_{i.} n_{.i}^2}{N^3} \right)}{N(1 - P_C)^2}$$

Kappa: Example of analysis

- Two judges scored the sweetness of 466 ice cream samples
- The scores are: *very sweet (A), sweet (B), not sweet (C)*
- Results:

Judge 2’s scores	Judge 1’s scores			Total
	A	B	C	
A	302	27	5	334
B	40	55	9	104
C	1	9	18	28
Total	343	91	32	466

Kappa : Example of analysis

- Proportion of agreement: $P_A = 0.805$

- Proportion of change agreement: $P_C = 0.575$

- Kappa statistic: $\kappa = 0.54$

- Variance of κ: 0.00161

- Standard error of κ: sqrt(0.00161) = 0.04

- 95% confidence interval of κ : $0.54 \pm 2(0.04) = 0.46$ to 0.62

Summary

- **Reliability** (reproducibility, repeatability) is different from **accuracy** (validity) concept.

- Analysis of reliability for continuous measurements: coefficient of reliability, coefficient of variation, limit of agreement.

- Analysis of reliability for categorical measurements: Kappa statistic.