Usual study design

Commonly chosen in PPUKM

- Clinical trial
- Cross-sectional
- Case-control
- Cohort (PURE & Malaysian Cohort)

Specific for Patho/Diagnostic

- Diagnostic testing
- Sensitivity, Specificity & ROC
- Kappa/ Agreement

Universiti Kebangsaan di Persada Antarabangsa

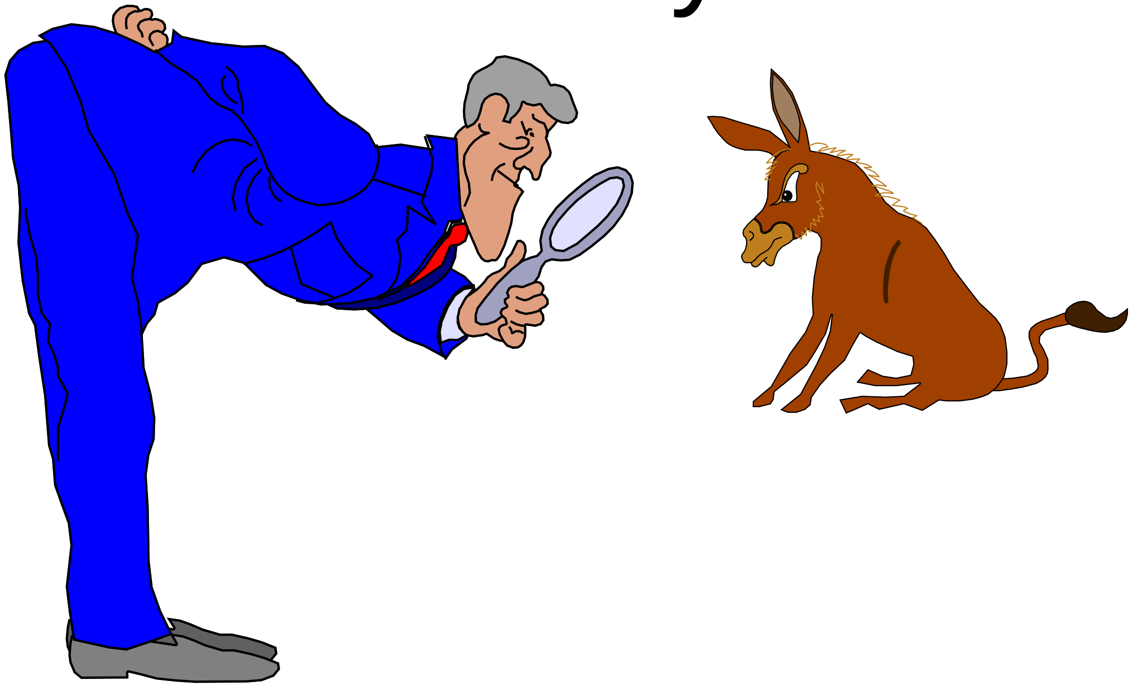
Universiti Kebangsaan Malaysia
National University of Malaysia

Clinical Agreement & Reliability Analysis

Dr Azmi Mohd Tamil

www.ukm.my

Assessing the validity and reliability of test



Issues in assessing clinical tests

- Validity of the tests – how good is the test to identify the sick and the healthy individuals
- Reliability – how stable is the results of the test .
- Efficiency and cost-effectiveness

How to measure?

- Accuracy – Sensitivity & Specificity
- Reliability
  - Qualitative
    - Kappa Analysis
    - AC-1 Statistic
  - Quantitative
    - Bland-Altman Plot
    - Cronbach α Coefficient

Reliability

RELIABILITY OF THE TEST

- Intra-observer (self) vs. inter-observer variation
- Repeatability or reproducibility – how stable is the results if repeated many time in similar condition / situation.

How to measure reliability?

- Qualitative
  - Kappa Analysis
  - AC-1 Statistic
- Quantitative
  - Bland-Altman Plot
  - Cronbach α Coefficient
  - Intra Class Correlation (ICC)

Reliability – Qualitative Data

Kappa Analysis

Table 2-6. Agreement between two examinations of the same 100 fundus photographs by one clinician

		Second examination	
		Little or no retinopathy	Moderate or severe retinopathy
First examination	Little or no retinopathy	69	11
	Moderate or severe retinopathy	1	19

Data modified from N. Aoki, H. Horibe, Y. Ohno, et al. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn. Circ. J.* 41:11, 1977.

Table 2-5. Agreement between two clinicians examining the same set of 100 fundus photographs

		Second clinician	
		Little or no retinopathy	Moderate or severe retinopathy
First clinician	Little or no retinopathy	46	10
	Moderate or severe retinopathy	12	32

Data modified from N. Aoki, H. Horibe, Y. Ohno, et al. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn. Circ. J.* 41:11, 1977.

Table 2-5. Agreement between two clinicians examining the same set of 100 fundus photographs

		Second clinician	
		Little or no retinopathy	Moderate or severe retinopathy
First clinician	Little or no retinopathy	46	10
	Moderate or severe retinopathy	12	32

Data modified from N. Aoki, H. Horibe, Y. Ohno, et al. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn. Circ. J.* 41:11, 1977.

OBSERVED AGREEMENT

= (46+32)/100

=78%

Table 2-5. Agreement between two clinicians examining the same set of 100 fundus photographs

		Second clinician	
		Little or no retinopathy	Moderate or severe retinopathy
First clinician	Little or no retinopathy	46 **Expected =32.5**	10 **Expected 23.5**
	Moderate or severe retinopathy	12 **Expected =25.5**	32 **Expected =18.5**

Data modified from N. Aoki, H. Horibe, Y. Ohno, et al. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn. Circ. J.* 41:11, 1977.

Table 2-5. Agreement between two clinicians examining the same set of 100 fundus photographs

		Second clinician	
		Little or no retinopathy	Moderate or severe retinopathy
First clinician	Little or no retinopathy	**32.5**	**23.5**
	Moderate or severe retinopathy	**25.5**	**18.5**

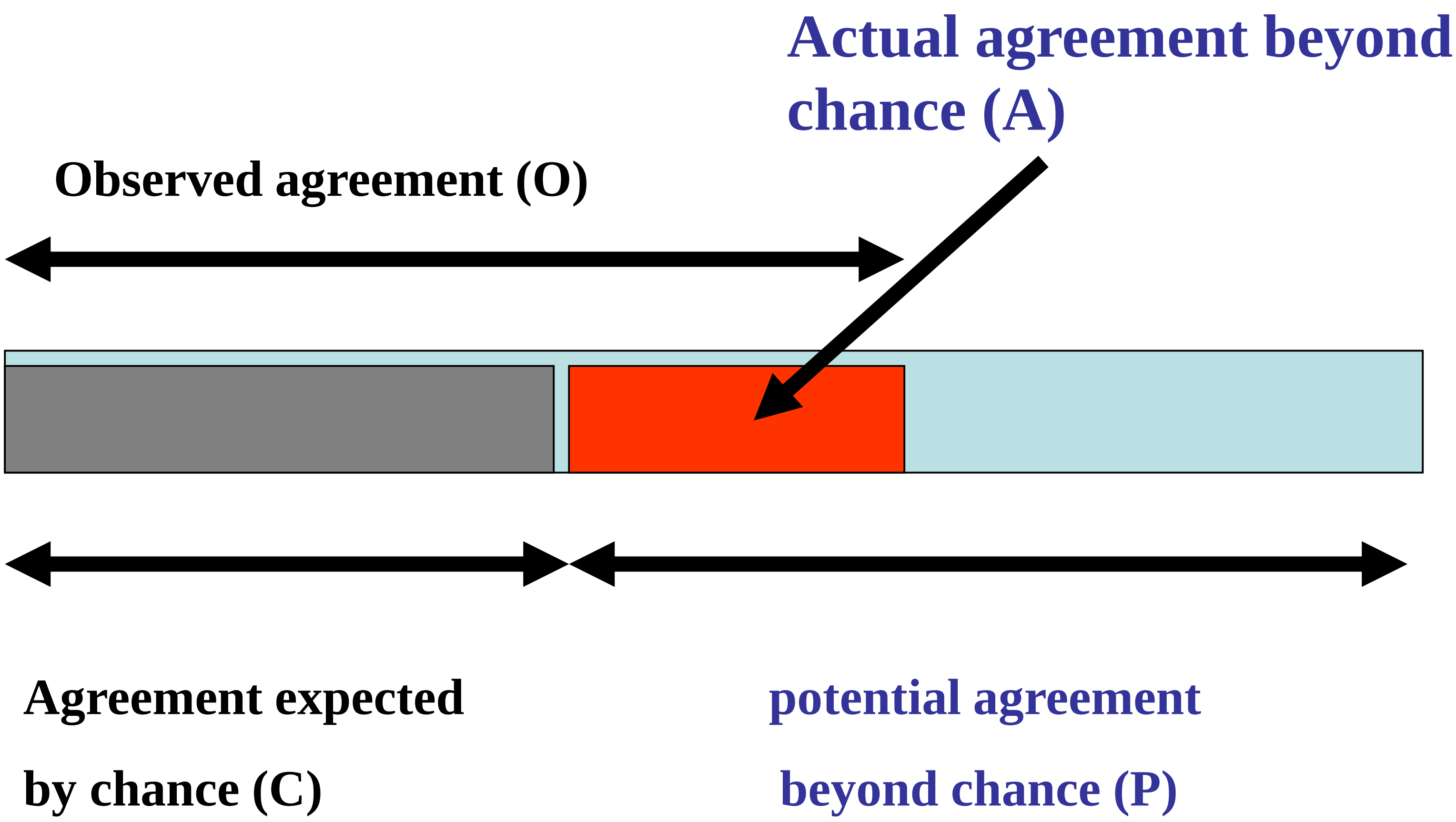
Data modified from N. Aoki, H. Horibe, Y. Ohno, et al. Epidemiological evaluation of funduscopic findings in cerebrovascular diseases: III. Observer variability and reproducibility for funduscopic findings. *Jpn. Circ. J.* 41:11, 1977.

EXPECTED AGREEMENT

= (32.5+18.5)/100

=51%

Clinical agreement



kappa

Kappa = $\dfrac{\text{actual agreement beyond chance}}{\text{potential agreement beyond chance}}$

= O – C / 100 – C

= (78-51) / 100-51

= 27 / 49 = 0.551

- 27/49 is the proportion of potential agreement beyond chance that was actually achieved for the absence/little or mod/severe retinopathy.

Interpretation of kappa values

- <0 No agreement
- 0.0-0.19 Poor agreement
- 0.20-0.39 Fair agreement
- 0.40-0.59 Moderate agreement
- 0.60-0.79 Substantial agreement
- 0.80-1.00 Almost perfect agreement

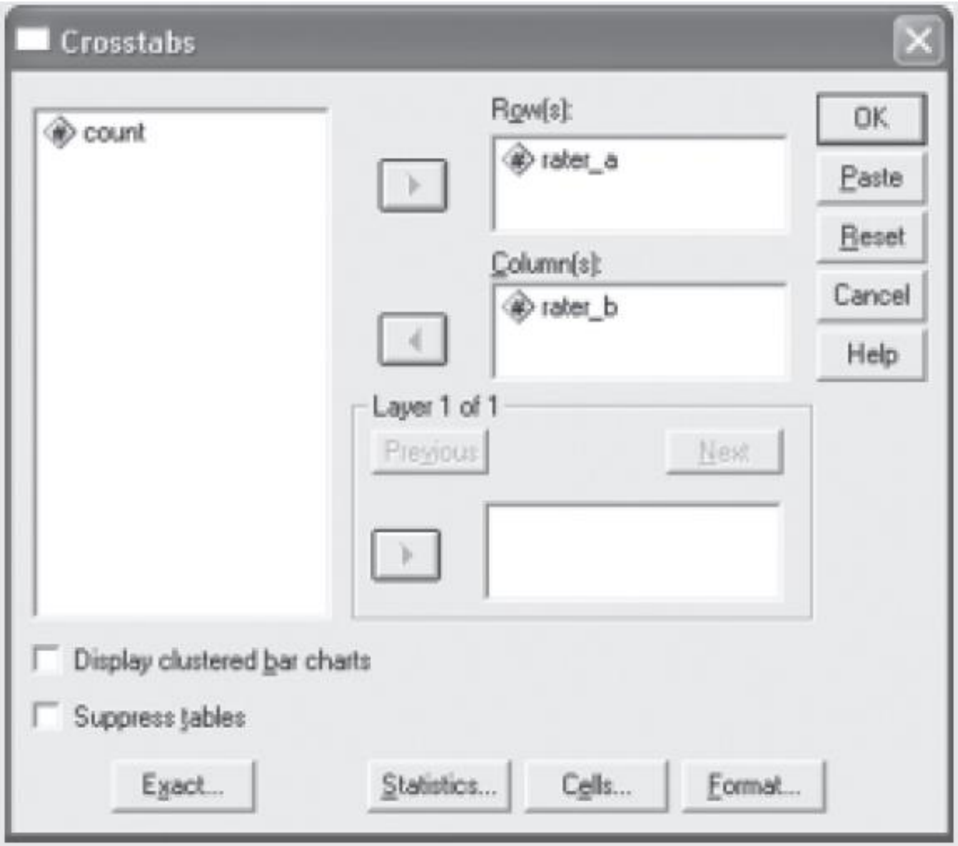
Can also use r as example

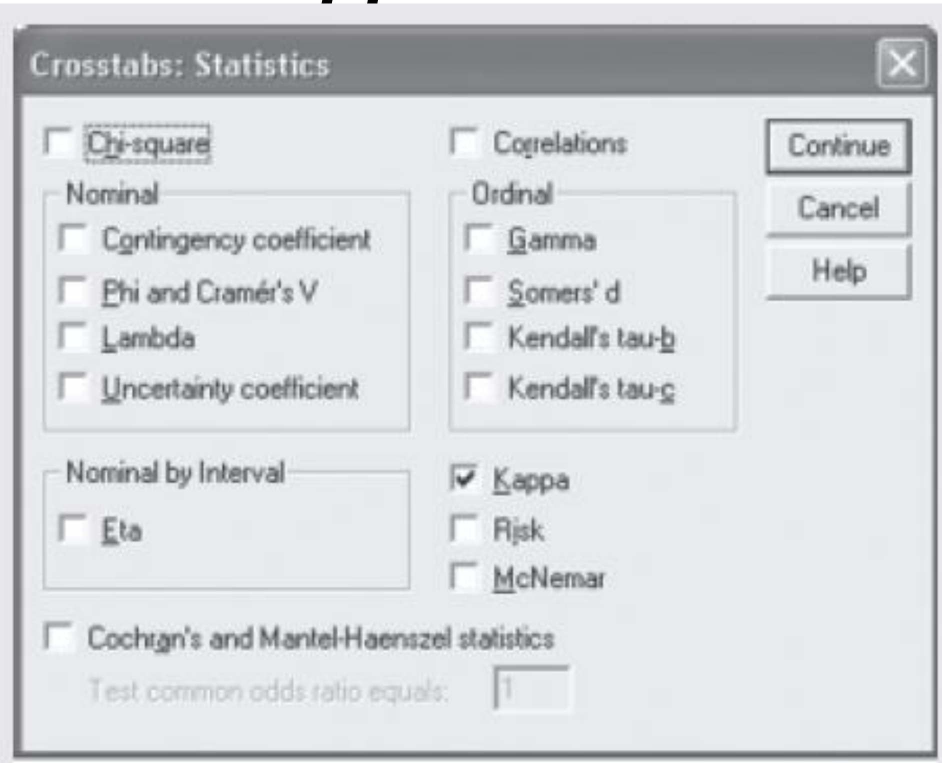
Table II. Strength of linear relationship.

Correlation Coefficient value	Strength of linear relationship
At least 0.8	Very strong
0.6 up to 0.8	Moderately strong
0.3 to 0.5	Fair
Less than 0.3	Poor

Can use SPSS to calculate kappa

- ***Analyse, Descriptive Statistics, Crosstab - choose Statistics and tick on Kappa.***





SPSS Output

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.551	.084	5.518	.000
N of Valid Cases		100			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Reliability – Qualitative Data

AC-1 Statistic

Limitation of Kappa Analysis

- When one of the category in agreement is low, the Kappa value is artificially low despite the high agreement value.
- For example; the table shows an observed agreement of 90% and yet the kappa value is only 0.444.
- In such situations, an alternative statistic known as AC1-statistic is suggested since it is more consistent with the percentage of agreement between raters in all situations.

RATER_A * RATER_B Crosstabulation

Count

		RATER_B		
		no disease	disease	Total
RATER_A	no disease	85	5	90
	disease	5	5	10
Total		90	10	100

AC-1 Statistic

Rater B	Rater A		
	1	2	Total
1	A	B	B1
2	C	D	B2
Total	A1	A2	N

$$AC1 = \frac{p - \varnothing}{1 - \varnothing} \text{ where } p = \frac{A + D}{N} \text{ and } \varnothing = 2q(1-q),\ q = \frac{A1 + B1}{2N}$$

Example

RATER_A * RATER_B Crosstabulation

Count

		RATER_B		
		no disease	disease	Total
RATER_A	no disease	85	5	90
	disease	5	5	10
Total		90	10	100

- p=(85+5)/100=0.9
- q=(90+90)200=0.9
- Φ=1.8(1-0.9)=0.18

- $AC1=\frac{(0.9-0.18)}{(1-0.18)}$
  =0.72/0.82
  =0.878
- This is more consistent with the two raters having the same rating for 90% of the subjects

Reliability – Quantitative Data

Bland Altman Analysis

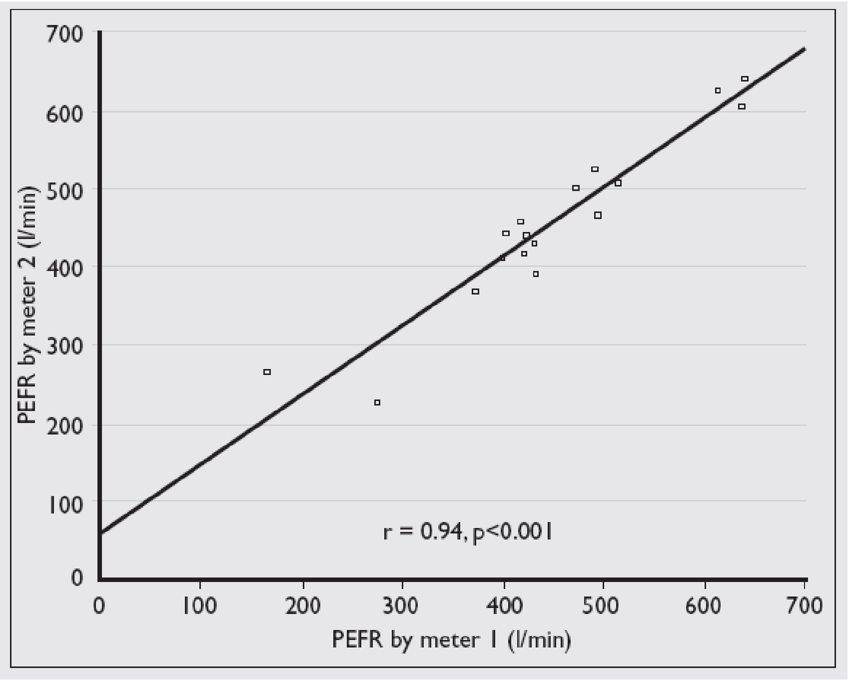
Correlation for agreement analysis?

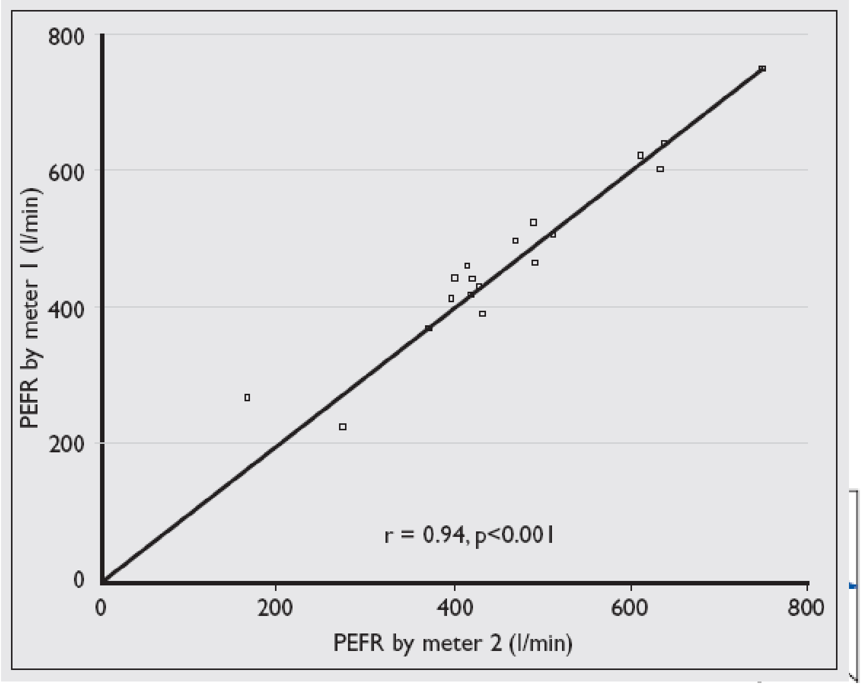
- paired t-test is not a suitable test to show agreement between two quantitative measurements (for example, two instruments measuring temperature) since a high correlation does not imply agreement.
- To show agreement on correlation, the "line of agreement" should be a 45 degrees ($x = y$) line.

Correlation for agreement analysis?

Universiti Kebangsaan Malaysia

Both scatter plots show high correlation but which one shows agreement?





Bland Altman plot

Universiti Kebangsaan di PERSADA ANTARABANGSA

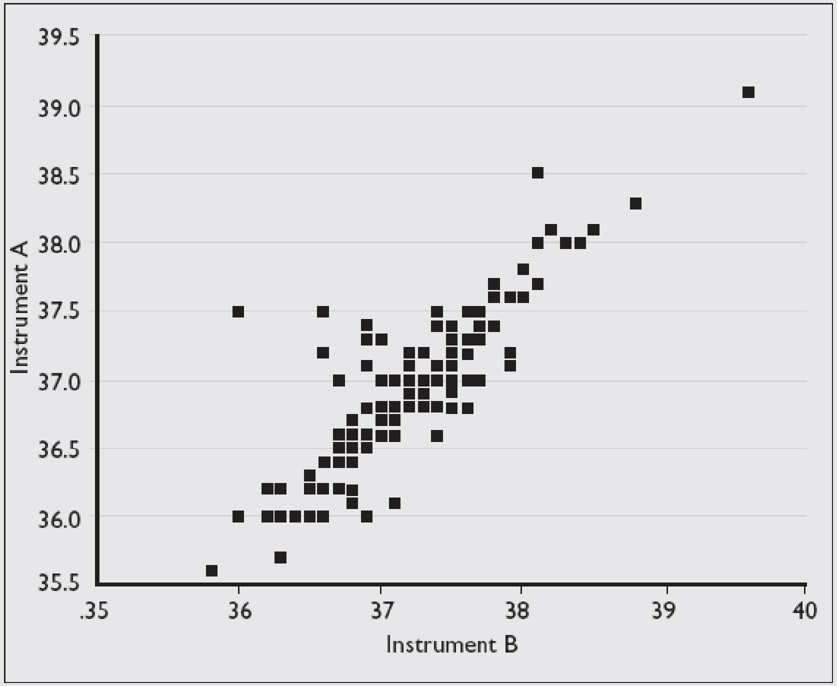
UNIVERSITI KEBANGSAAN MALAYSIA
National University of Malaysia

- Bland-Altman plot describes agreement between two quantitative measurements.
- There's no p-value available to describe this agreement but rather a "quality control" concept.
- The difference of the paired two measurements is plotted against the mean of the two measurements and it is recommended that 95% of the data points should lie within the $\pm$2sd of the mean difference.

Example

- Let us analyse agreement of two temperature-measuring instruments.
- The diagram shows the scatter plot between instrument A vs instrument B, the correlation is 0.871, p<0.001.

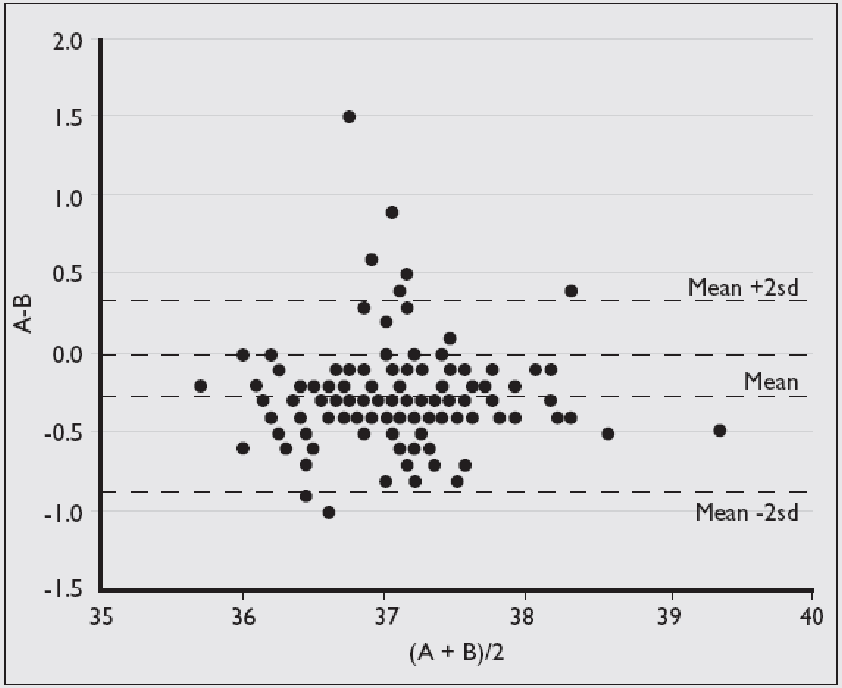
Fig. 5 Scatter plot between Instruments A and B.



Plotting the B-A Plot

- To plot, compute the differences between the instruments (A-B) and the mean of both instruments ((A+B)/2) for all the paired values.
- The mean (sd) of the differences is -0.2665 (0.3022), thus the +2sd of the mean are -0.8709 and 0.3379.
- We want the cluster of points to be around the difference = 0 line.
- Any large deviated differences have to be checked since they might be due to
  - Wrong data-entry,
  - Operator error or
  - Flaw of the instrument(s)

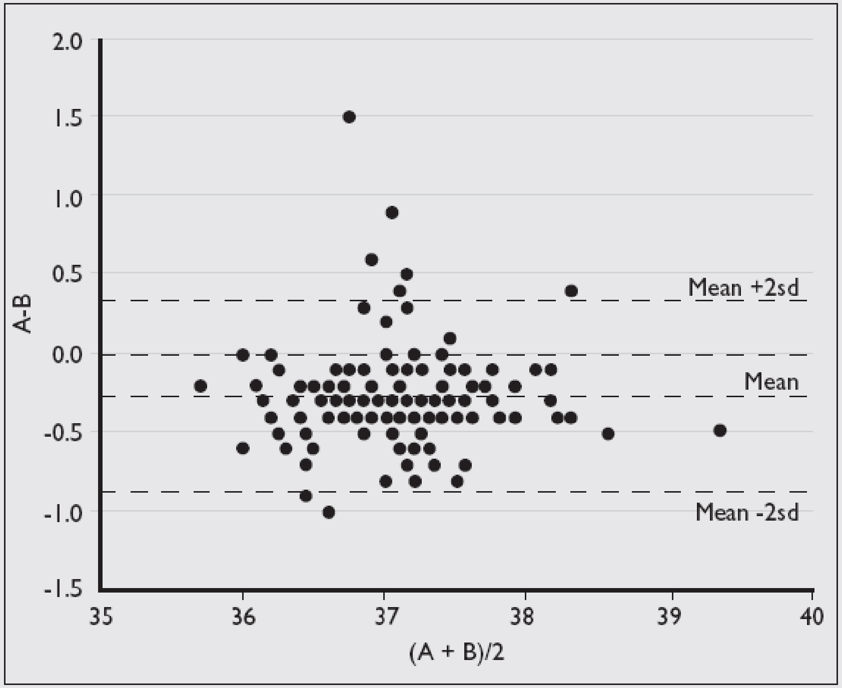
Fig. 6 Bland Altman plot.



Interpreting the B-A Plot

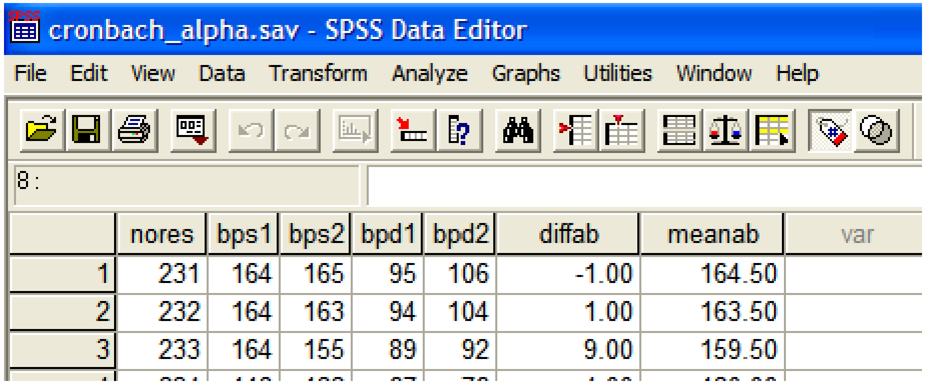
- 8/150 (5.3%) of the points are beyond the +2sd lines
- Instrument A seems to be measuring 'low' most of time
- A trend of high scores amidst the 37 degree value.
- The extreme difference is 1.5 degree Celsius.
- Therefore are the two instruments agreeable?
- Well, if one does not mind a one-degree difference, then they are agreeable; otherwise you have to make the judgement call.

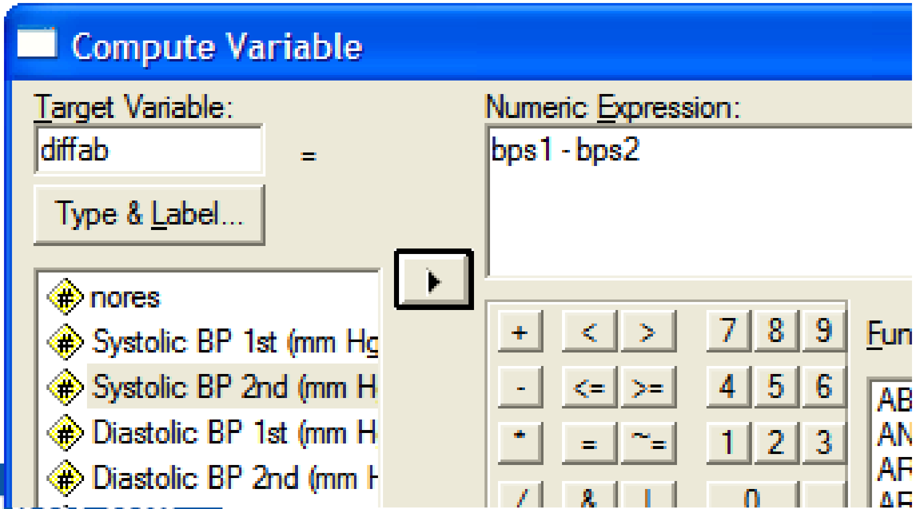
Fig. 6 Bland Altman plot.



B-A Plot Using SPSS

- These data came from BP measurements using two different BP sets, mercury sphygmomanometer (bps1 &bpd1) and an electronic BP set (bps2 & bpd2).
- Using the systolic pressure as an example, use “Transform -> Compute” to create the difference and mean of A & B.





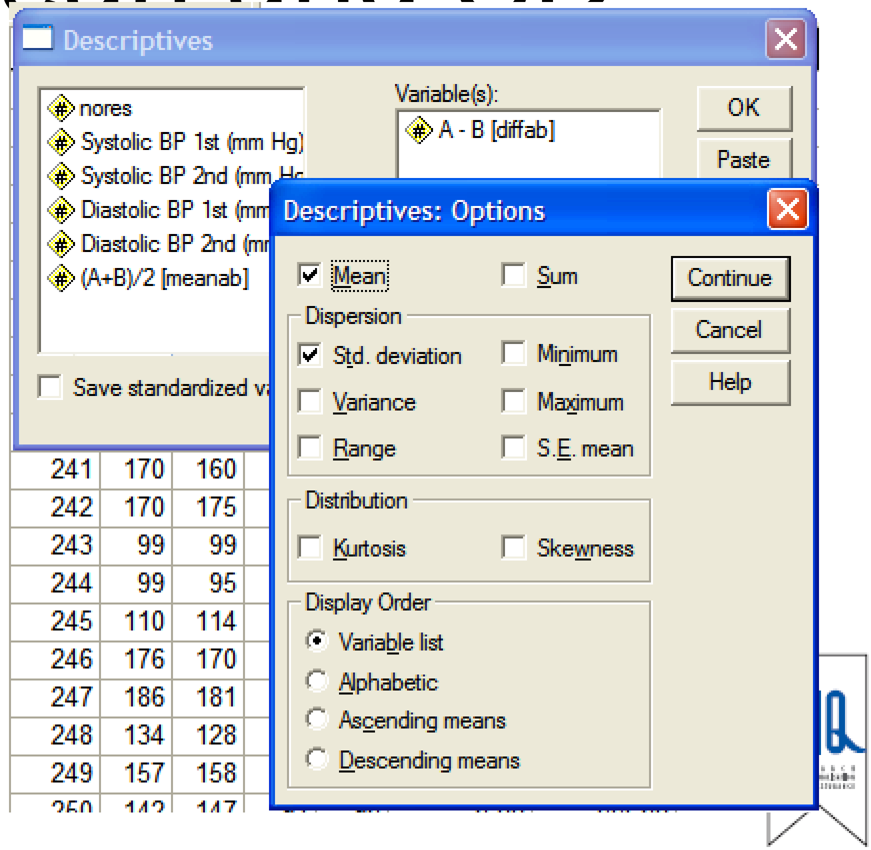
Calculate Mean and SD

- Click on menu “Analyse->Descriptive->Select A-B; for “Statistics”, select Mean & Std deviation”.

Descriptive Statistics

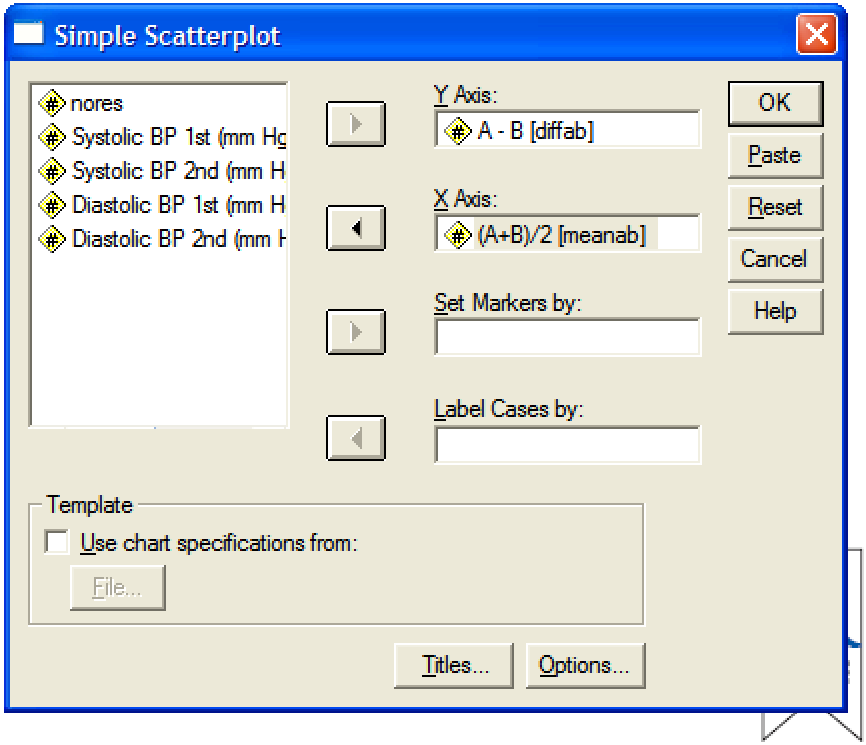
	N	Mean	Std. Deviation
A - B	100	2.8000	7.19427
Valid N (listwise)	100		

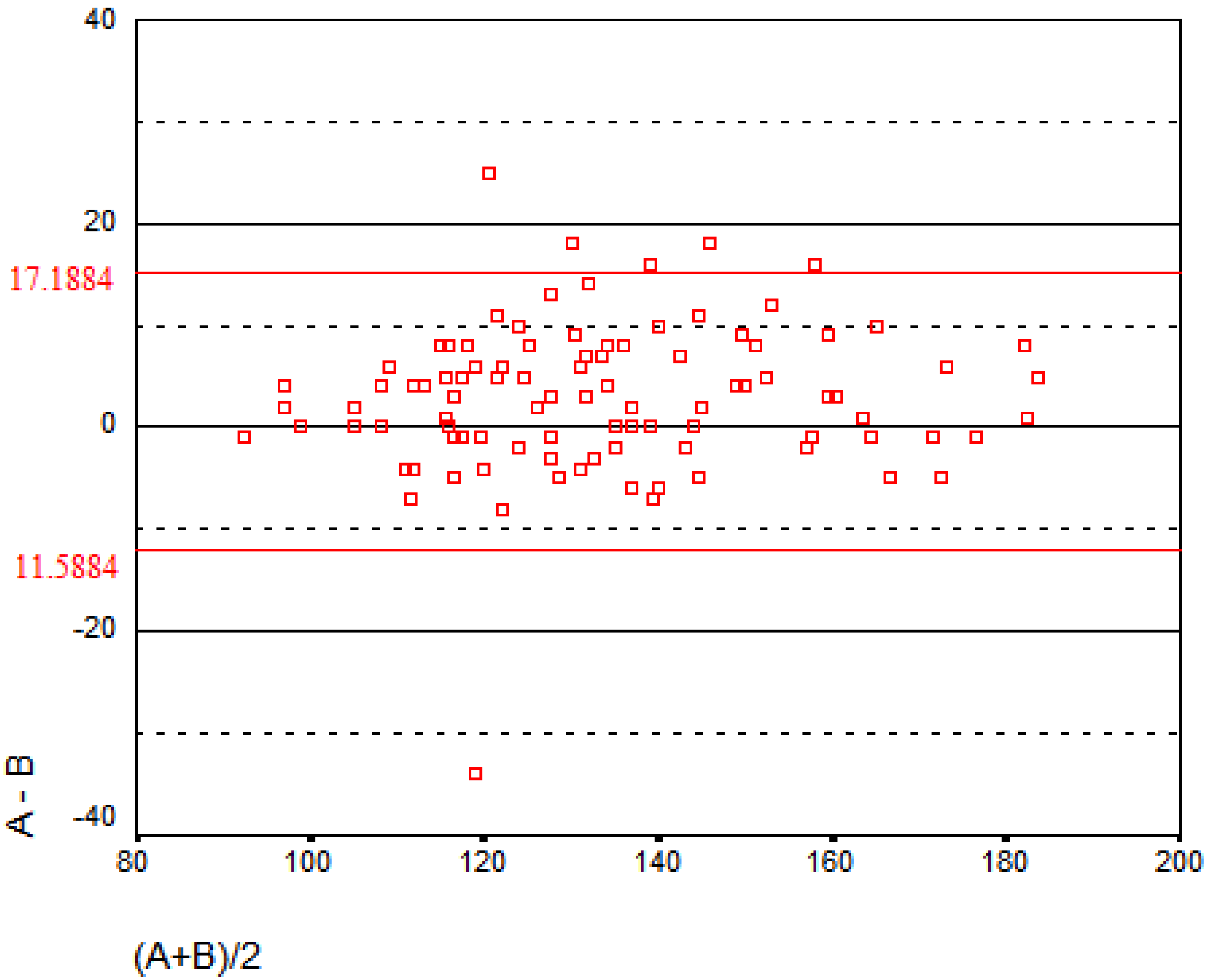
- Mean difference is 2.8000 with sd of 7.1942. Therefore +2sd of mean is -11.5884 and 17.1884.



Plotting the B-A Plot Using SPSS

- Click on menu; “Graph->Scatter->Simple->Define”.
- Fill up the requester accordingly as in illustration. X axis is “(A+B)/2”, Y axis is “A-B”.





- 94% of the points are within the +2sd lines. Yet some of the readings differ up to 25mm Hg.

Interpreting the Tolerance

- Another way to assess the agreement is to set acceptable tolerances for the differences between the two instruments; |A-B|.
- For example, how many percent in the 0.1 degree difference, 0.2, etc.
- At least 87% of the measurements agree on the 0.5 degree tolerance.
- Is it acceptable to have 13% being more than 0.5 degrees difference?

Table VII. "Tolerance" table: absolute differences between instruments A & B.

Tolerance (degrees)	n (%)
0.2	29 (19.3)
0.4	117 (78.0)
0.5	131 (87.3)
0.6	133 (88.7)
0.8	144 (96.0)
1.0	149 (99.3)

Reliability – Quantitative Data

Cronbach α Coefficient

Cronbach α Coefficient

- Usually used to measure internal consistency.
- In scales, the value should be above 0.7.
- In SPSS, select Analyze, Scale, Reliability Analysis.
- Select the two variables that is being compared and move them into the box marked **Items**.
- In the **Model** section, make sure **Alpha** is selected.
- Click on the **Statistics** button. In the **Descriptives for** section, click on **Item** & **Scale.**
- Click on **Continue** and then **OK.**

Example

- These data came from BP measurements using two different BP sets, mercury sphygmomanometer (bps1 &bpd1) and an electronic BP set (bps2 & bpd2).

SPSS Output

R E L I A B I L I T Y   A N A L Y S I S - S C A L E (A L P H A)

		Mean	Std Dev	Cases
1.	BPS1	135.0300	21.2087	100.0
2.	BPS2	132.2300	20.5052	100.0

Statistics for	Mean	Variance	Std Dev	N of Variables
SCALE	267.2600	1688.7802	41.0948	2

Reliability Coefficients

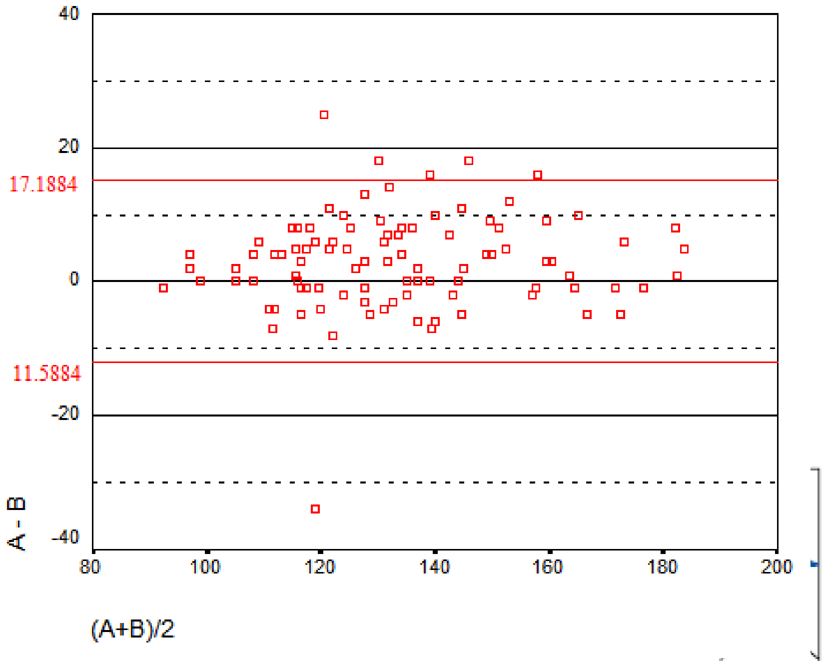
N of Cases =   100.0   N of Items = 2

Alpha =   .9694

- The output indicated a Cronbach apha coefficient of 0.9694.
- Comparing the output with a B-A plot; mean of difference is 2.8000 with sd of 7.1942.

Cronbach α Coefficient vs B-A Plot

- 94% of the points are within the +2sd lines. Yet some of the readings differ up to 25mm Hg.
- And the Cronbach apha coefficient is as high as 0.9694, indicating that the measurements were reliable.
- So is the electronic BP set as reliable as a mercury sphygmomanometer?



Conclusion

- Use of which reliability test depends on the type of data being analysed.
- Some reliability tests requires subjective judgement during interpretation.

TERIMA KASIH