Group work - reproducibility answers

Assignment

Lars is involved in a study on grip strength in patients with rheumatoid arthritis, and he is interested in finding out more about the reproducibility of the procedures he uses.

Therefore, he has designed a study which includes measuring the grip strength (Nm2) in 20 patiens with rheumatoid arthritis at two different time points (1 week in between). He obtains the results outlined in Table 1.

Table 1. Test-retest data on grip strength in 20 rheumatoid arthritis patients

Patient no.	Measurement 1	Measurement 2
1	86	92
2	40	47
3	50	55
4	52	57
5	69	74
6	62	64
7	84	84
8	68	72
9	58	62
10	71	74
11	92	97
12	76	74
13	77	81
14	77	83
15	64	67
16	35	34
17	88	94
18	76	74
19	103	110
20	117	125

1. Pearson's correlation coefficient

Lars calculates the Pearson's correlation coefficient to be: 0.991 ($p<0.0001$).

Questions

1.1 Discuss what the correlation coefficient tells us about the reliability between the two measurements?

> ***Answer***
>
> The Person's correlation coefficient:
>
> 1. Tells us if there is a linear association between the two measurements
> 2. Depends on the variance of the data. E.g.: if the SD is large, we get a higher correlation compared to a small SD
> 3. Does not take systematic error or variance differences between the measurements into account
>
> The variance of the data is: Measurement 1 [35 - 117], Measurement 2 [47 - 125]
>
> Therefore, the reliability is high, but this does not mean that the two measurements find exactly the same values for grip strength.

1.2 Explain what the p-value means?

> ***Answer***
>
> The P-value tells us that the correlation coefficient is statistically significant different from 0. This is of no value as we would like to know if how close the coefficient is to 1.

2. Intraclass correlation coefficients

The numbers in Table 2 below are an extraction from an SPSS analysis (General Linear Models, variance components, restricted maximum likelihood) of the data from the Table 1.

Table 2. Variance components

Variance component	Estimate
Var(patients)	424,579
Var(measurements)	6,811
Var(error)	4,414

Questions

2.1 Using the variance estimates, please calculate ICC-consistency (model 2.1), ICC-agreement (model 2.1) SEM-consistency and SEM-agreement.

> ***Answer***

$$ICC_{consistency} = ICC\ [2.1c] = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{po,e}^2} = 424.579/(424.579+4.414) = 0.99$$

$$ICC_{agreement} = ICC\ [2.1a] = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2 + \sigma_{po,e}^2} = 424.579/(424.579+6.811+4.414) = 0.97$$

$$SEM_{consistency} = \sqrt{\sigma_{po,e}^2} = 4.414 = 2.10$$

$$\text{SEM}_{agreement} = \sqrt{\sigma_o^2 + \sigma_{po,e}^2} = \sqrt{6.811 + 4.414} = 3.35$$

2.2 What do you think about the reliability and the measurement error?

Answer

Both ICC's are high:

- ICC-consistency is highest
- ICC-agreement is slightly lower due to a small systematic difference between measurement 1 and 2

SEM-consistency = 2.10 and SEM-agreement = 3.35, therefore the measurement error is 2.10 Nm2 or 3.35 Nm2 (if including the systematic error):

- The measuremet error is very small compared to the variance of the data (range: [35 - 125])
- Therefore, good reliability and small measurement error

3. Limits of agreeement and systematic differences

A paired t-test of measurement 1 (score1) and measurement 2 (score2) is shown in Table 3.

Table 3. Paired t-test

```
. ttest score1 = score2

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  score1 |      20       72.25    4.508982    20.16478    62.81259    81.68741
  score2 |      20          76    4.750623    21.24543    66.05683    85.94317
---------+--------------------------------------------------------------------
    diff |      20       -3.75    .6644151    2.971354   -5.140637   -2.359363
------------------------------------------------------------------------------
     mean(diff) = mean(score1 - score2)                            t =  -5.6441
 Ho: mean(diff) = 0                             degrees of freedom =       19

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
```

```
Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Questions

3.1 What is the systematic difference between measurement 1 (score1) and measurement 2 (score2)? Which measurement has the highest average score?

Answer

- The systematic difference is -3.75 Nm2
- Score 1 (measurement 1) has the highest score of 72.25 Nm2

3.2 What is LOA (limits of agreement)?

Answer

- LOA = -3.75 ± 1.96 x 2.97 = -3.75 ± 5,82 Nm2 = [2.07; -9.57]

3.3 What will happen with ICC-consistency and ICC-agreement if measurement 2 (score2) always measures 20 Nm2 lower compared to measurement 1 (score1)?

Answer

If measurement 2 consequently measures 20 Nm2 lower compared to measurement 1, then:

- ICCconsistency will stay the same
- ICCagreement will be lower as it includes systematic error

3.4 What will happen to the 'limits of agreement'?

Answer

- If measurement 2 consequently measures 20 Nm2 lower compared to measurement 1, the mean score of measurement 2 will be 52.25 Nm2 (72.25 – 20). Therefore, the systematic error will be (mean-score 1 – mean-score 2): 72.25 - 52.25 = 20 Nm2.
- The width of the LOA stays the same, but we get higher LOA limits: LOA = 20 ± 1.96*2.97 = (14.18; 25.82)

4. Reproducibility of shoulder measurements

In a study by De Winter et al. (2004) two physiotherapists have measured shoulder abduction (Figure 1) in 155 patients with pain in one shoulder. They used an electronic inclinometer, showing the abduction angle in degrees (Figure 2).

Figure 1. Shoulder abduction

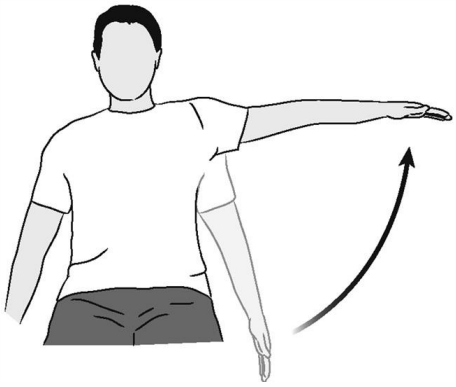

Figure 2. Inclinometer measuring the abduction angle

Table 4 shows the results in degrees as well as other measures of reproducibility for both shoulders.

Table 4. Results

	Phys A	Phys B	Mean diff. A-B	Agreement		
	Mean (SD)	Mean (SD)	(SD_{A-B})	< 5°	< 10°	ICC
Painful shoulder	69.5 (17.6)	68.8 (16.3)	0.8 (10.0)	43%	72%	0.83
Healthy shoulder	79.8 (7.6)	78.9 (8.4)	0.9 (9.6)	43%	72%	0.28

Questions

4.1 Explain the large difference between the two ICC's in light of the exact same results for the 5 and 10% agreement.

Answer

Look at the variance of the results (SD):

- Abduction of the healthy shoulder is almost the same, therefore SD is small: low reliability (ICC)
- Abduction of the painful shoulder varies more, therefore SD is large: high reliability (ICC)

4.2 Discuss which parameter is preferable?

Answer

Whether one prefers degrees or ICC from 0-1 is a personal preference. Personally, I prefer degrees as it is more meaningful clinically and because ICC's are more difficult to interpret.

5. Design of a reproducibility study

A researcher is designing a reproducibility study. On a course in questionnaire technique he has heard that the reliability will improve if the study population is more heterogenous. He therefore designs the study to include very different patients.

Questions

5.1 Discuss this strategy?

Answer

Reliability of a questionnaire must be tested in the same patient population as where is should be used. Therefore, this design is flawed.