Group work - Validity answers

Background

The data we will consider originates from a randomized controlled study (RCT). The main purpose of the RCT was to investigate the effect of Acceptance and Commitment Therapy (ACT) in 126 patients with severe health anxiety. During the study patients were asked to complete several questionnaires (among others SF-36 and SF-12) at different timepoints.

Outcome measures

SF-36

The SF-36 is a generic multi-purpose, short-form health survey with only 36 questions. It yields an 8-scale profile of scores (physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional and mental health) as well as physical (PCS) and mental (MCS) health summary measures. The 8-scale profile scores and the summary measures ranges from 0-100 with higher scores indicating better health. The scores are considered continuous.

SF-12

The SF-12 Health Survey is a 12-item subset of the SF-36 that measures the same 8 scales of health including PCS and MCS. It is a brief, reliable measure of overall health status. It is useful in large population health surveys and has been used extensively as a screening tool. The score ranges are the same as for the SF-36 and scores are considered continuous.

Criterion validity - continuous instruments

You choose to design a validation study and look at criterion validity of the Physical Component Summary (PCS), as computed by the SF-12, by comparing it to PCS calculated by the SF-36 (gold standard).

Question 1 - The gold standard

1.1 Discuss whether the choice of gold standard and sample is appropriate for the purpose?

> ***Answer***
>
> The choice of SF-36 as a gold standard is probably the closest we will get to a 'valid' gold standard when the comparison is between two questionnaires. However, one can argue if the SF-36 is actually a valid measure of QoL in our population of patients with anxiety. The SF-36 is generic and may not be an efficient measure of OoL in all populations. If not, the SF-36 is more of a 'silver' standard.
>
> The population is from a randomized clinical trial which include very selected patients with anxiety. Many trials have stringent inclusion and exclusion criteria making the population narrow and homogenous. Consequently,

our results may not be generalizable to other populations of patients with anxiety. This is the price of including a validity study in an RCT.

Question 2 – limits of agreement

A simple paired t-test between the PCS SF36 and the PCS SF12 reveals the following output:

```
. ttest pcs_sf36 = pcs_sf12

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
pcs_sf36 |     123    47.36617    .7058742    7.828523    45.96882    48.76352
pcs_sf12 |     123    44.82154    .6827179    7.571708    43.47003    46.17305
---------+--------------------------------------------------------------------
    diff |     123    2.544632    .3344371    3.709087    1.882581    3.206684
------------------------------------------------------------------------------
     mean(diff) = mean(pcs_sf36 - pcs_sf12)                       t =   7.6087
 Ho: mean(diff) = 0                               degrees of freedom =      122

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

2.1 What is the mean difference between pcs_sf36 and pcs_sf12?

Answer

The mean difference is 2.54 on a scale from 0-100. The PCS of the SF-36 measures higher compared to PCS of the SF-12, and the systematic error is therefore 2.54 points.

2.2 What is the limits of agreement (i.e. prediction interval)?

Hint

The LOA = $\delta \pm 1.96 \times SD$

Answer

The LOA = $\delta \pm 1.96 \times SD = 2.54 \pm 1.96 \times 3.71 = [-4{,}73; 9{,}81]$

2.3 Explain what the limits of agreement tells you?

Answer

The LOA indicates the size of the measurement error between the two instruments, both the systematic (δ) and the random error (1.96 x SD). It is equivalent to the 95% prediction interval and is an estimate of an interval in which 95% of future observations will fall.

Question 3 – Bland and Altman plot

The score range of PCS is 0-100 and the summary statistics of the two variables are:

```
. sum pcs_sf36 pcs_sf12

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    pcs_sf36 |        125    47.37597    7.768135   29.12679   66.29746
    pcs_sf12 |        123    44.82154    7.571708   25.04456   57.46968
```

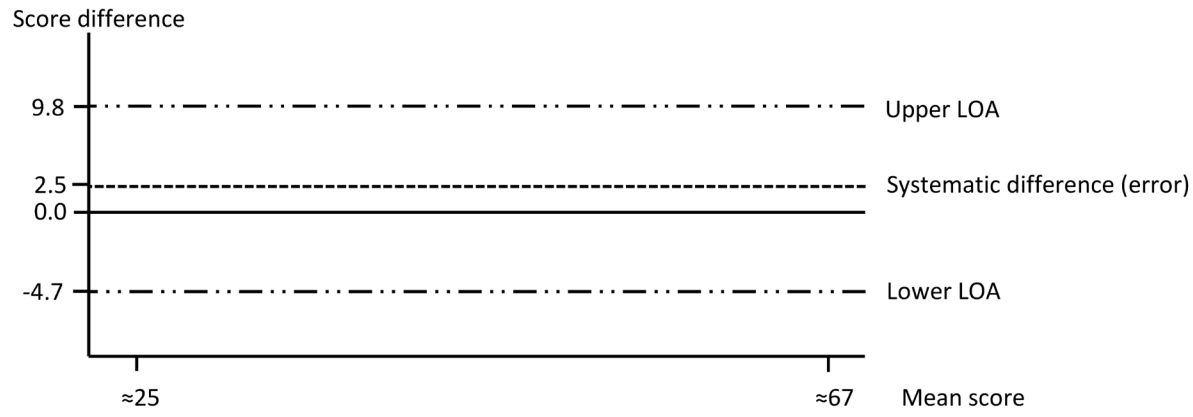
From the paired t-test we have a mean difference between pcs_sf36 and pcs_sf12 = 2.54. The limits of agreement were found in question 2.2.

3.1 Draw a Bland & Altman limits of agreement plot and put values on the x and y axis.

Hint

You need to put labels and numbers on the x and y-axis. Furthermore you need to draw the lines for zero, LOA_{upper}, LOA_{lower}, and the systematic error.

Answer



Criterion validity – dichotomous instruments

In another study, 16060 patients were screened for cancer using a gold standard (criterion) and a new test instrument (test). You hypothesize that the new instrument is as good as the gold standard.

Question 4 – sensitivity and specificity

To test your hypothesis, you calculate sensitivity and specificity:

```
. diagt test criterion

           |       criterion
      test |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       969        231 |     1,200
    Normal |       164     14,696 |    14,860
-----------+----------------------+----------
     Total |     1,133     14,927 |    16,060
True abnormal diagnosis defined as test = 1
```

			[95% Confidence Interval]	
Prevalence	Pr(A)	7.5%	7.1%	7.9%
Sensitivity	Pr(+|A)	80.8%	78.4%	82.9%
Specificity	Pr(-|N)	98.9%	98.7%	99.1%
ROC area	(Sens. + Spec.)/2	0.90	0.89	0.91
Likelihood ratio (+)	Pr(+|A)/Pr(+|N)	73.17	62.68	85.41
Likelihood ratio (-)	Pr(-|A)/Pr(-|N)	0.19	0.17	0.22
Odds ratio	LR(+)/LR(-)	375.90	304.60	463.88
Positive predictive value	Pr(A|+)	85.5%	83.3%	87.5%
Negative predictive value	Pr(N|-)	98.5%	98.2%	98.6%

4.1 How large do you think sensitivity and specificity should be before the two methods agree?

Hint

Explain to the group what sensitivity and specificity means. What is the misclassification?

Answer

Sensitivity is the proportion of people *with the disease* (gold standard +) who are correctly identified by a positive test result (true positive rate).

Specificity, on the other hand, is the proportion of people *free of the disease* (gold standard -) who are correctly identified by a negative test result (true negative rate).

We want both to be as high as possible and preferably balanced, as this identifies most TP and TN. Our results show a sensitivity of 80.8% and specificity of 98.9% meaning that we have a false negative rate of 19.2% (1-sensitivity) and a false positive rate of 1.1% (1-specificity). To determine if the sensitivity and specificity are acceptable, we have to gage if the misclassification is clinically acceptable, and this obviously depends on the specific circumstances.

4.2 Explain the positive and negative predictive values and summarise your findings.

Answer

A positive predictive value (PPV) refers to the proportion of individuals with *positive test results* who have the disease (gold standard +).

A negative predictive value (NPV) refers to the proportion of individuals with *negative test results* who do not have the disease (gold standard -).

At a prevalence of 7.5%, our results show a PPV of 85.5% and the NPV of 98.5%. This would probably be very different if the prevalence was higher or lower.

Perhaps more intuitive is the + and - likelihood ratios (not included in the lecture):

A positive likelihood ratio (LR+) is a measure of how much more likely a positive test result is among people who have the condition of interest than it is among people who do not have the condition of interest.

A negative likelihood ratio (LR-) is a measure of how much more likely a negative test result is among people who have the condition of interest than it is among people who do not have the condition of interest.

Our findings show a LR+ of 73.2 and a LR- of 0.2. This means that a positive test result is 73% more likely among those who have cancer compared to those who do not. Conversely, a negative test result is about 20% more likely among people with who have cancer compared to those who do not.