Advanced course in questionnaire technique and clinimetrics

Exercises in validity

PRACTICAL INFORMATION

No computer is needed: all results required to solve the exercises are given in the tables in this text and in the associated graphs. A pocket calculator is sufficient for the calculations.

BACKGROUND INFORMATION

SF-36:

The SF-36 is a 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 and mental health summary measures (PCS and MCS). It is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. In these exercises we will only be using the PCS and MCS. The two summary measures (PCS and MCS) have been standardized to have a mean of 50 and a standard deviation of 10 in 1998 general U.S. population, with higher scores representing better health. The SF-12 is a short version of the SF-36, consisting of 12 of the 36 items, from which PCS and MCS can also be computed.

Whiteley-7 index:

The Whiteley 7-Index (WI-7) measures health anxiety (aka. hypochondriasis). It is a short version of the original 14-item Whiteley Index, and the items concern worry about illness, conviction of having a disease despite reassurance, and preoccupation with bodily symptoms. It consists of 7 items and is scored on a rating scale from 1-5 with a total sumscore ranging from 7-35. The scores were transformed to a 0-100 score point scale with higher scores indicating better health.

SCL-8:

The Symptom Checklist (SCL-8) consists of 8 items measuring self-reported symptoms of emotional distress. The items are scored on a rating scale from 1-5 with a total sumscore ranging from 8-40. The scores were transformed to a 0-100 score point scale with higher scores indicating better health.

EXERCISE 1.1 (criterion validity, continuous instrument and gold standard)

The data we will consider in this exercise originates from a randomized controlled trial (RCT). The main purpose of the RCT was to investigate the effect of Acceptance and Commitment Therapy (ACT) on 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 time-points. The SF-12 and the SF-36 were completed as two separate questionnaires, and the data in this exercise are from baseline.

In this exercise though, we will consider the concurrent criterion validity of the Physical Component Summary (PCS), as computed by the SF-12, by comparing it to PCS calculated by SF-36 (the gold standard). The results needed are given in Table 1 and in the graphs qq_diff_1_1.png and Scatter plot.png.

Table 1. PCS from SF-12 and SF-36 at baseline (n = 123; 3 patients with missing values)

	n	Mean	SD	95% CI for mean
SF-12 PCS	123	44.82	7.57	43.47 ; 46.17
SF-36 PCS	123	47.37	7.83	45.97 ; 48.76
Difference (SF-12 − SF-36)	123	−2.54	3.71	−3.21 ; −1.88

Paired t-test of H0: mean difference = 0: t = −7.61, df = 122, p < 0.001.

- Go through step one, two and four in the process of criterion validation (see slide 15 from today's lecture). That is; discuss whether the choice of gold standard and sample is appropriate for the purpose, and whether the scores from the instrument and the gold standard were obtained independently.

It is known, that for this group of patients, a change in PCS larger than ±5 (equal to half the standard deviation in the norm population) is considered clinically significant. Also, changes less than ±5 are considered as "no change". Hence we choose the required level of agreement to be between -5 and 5.

- Calculate the limits of agreement (approximately a 95% prediction interval for the difference between SF-12 PCS and SF-36 PCS in a new patient) using Table 1 (see slide 26).
- Is there a systematic difference between the two scores (Table 1)?
- Make a rough sketch of the Bland-Altman plot based on "Scatter plot.png": draw the axes, indicate roughly where the points lie, and add horizontal lines at 0, at the mean difference and at the limits of agreement (see slide 26).
- Check the assumptions behind the analysis (see slide 27 from today's lecture). In particular: is there any sign of a relationship between the differences and the averages?
- Make a summary of your findings, interpret and discuss the results. Would the required level of agreement be met if there were no systematic difference between the two scores?

EXERCISE 1.2 (criterion validity, dichotomous instrument and gold standard)

In a study similar to the one used in the example of concurrent validity in section 6.4.1 in "Measurement in Medicine", 16,060 patients were screened for cancer using a gold standard and a new test instrument.

The researchers hypothesized that the new instrument is sufficiently valid for its purpose, defined as a sensitivity of at least 85%, a specificity of at least 95%, a positive predictive value (PPV) of at least 80% and a negative predictive value (NPV) of at least 90%. The cross-table of the test result against the gold standard is given in Table 2.

Table 2. New test instrument against the gold standard

	Gold standard +	Gold standard −	Total
Test +	969	231	1,200
Test −	164	14,696	14,860
Total	1,133	14,927	16,060

- Identify the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in Table 2 (see slide 17).

- Calculate the sensitivity and specificity by "hand" (see slide 17).
- Calculate the negative and positive predictive value by "hand" (see slide 19).
- Find the sensitivity, specificity, the negative predictive value, the positive predictive value, and their 95% CI's in Table 3 (at the end of this exercise).
- Why are the 95% CI's for the specificity and negative predictive value so narrow compared to the CI's for the sensitivity and positive predictive value?
- Calculate the prevalence of cancer in the sample. Which PPV and NPV would you expect at this prevalence if the sensitivity and specificity were exactly as hypothesized? Are the researchers' four hypotheses consistent with each other (see slide 19)? Hint: PPV = sens·p / (sens·p + (1−spec)·(1−p)) and NPV = spec·(1−p) / (spec·(1−p) + (1−sens)·p), where p is the prevalence.
- Summarize and interpret your findings. Are the data consistent with the researchers' hypotheses about the sensitivity, specificity, PPV and NPV? Consider both the point estimates and the 95% CI's: which hypotheses are demonstrated, and which are merely not contradicted by the data?

Table 3. Estimates with 95% CI's

	Estimate (%)	95% CI (%)
Sensitivity	85.5	83.3 ; 87.5
Specificity	98.5	98.2 ; 98.6
Positive predictive value (PPV)	80.8	78.4 ; 82.9
Negative predictive value (NPV)	98.9	98.7 ; 99.1

Exact binomial confidence intervals.

EXERCISE 1.3 (Construct validity)

In this exercise we will consider only a part of the data from the RCT described in Exercise 1.1 (baseline data from 64 of the 126 patients), but now our focus is on assessing the convergent, discriminant and discriminative validity of the WI-7. The results needed are given in Tables 4 and 5 and in the graphs qq_1_3.png, scatter_1_3.png, qq_1_3_no_comor.png, qq_1_3_comor.png, qq_1_3_female.png and qq_1_3_male.png.

Hypotheses regarding convergent and discriminant validity (see slide 38):

- H1: A positive correlation of at least 0.50 between WI-7 and SCL-8 (similar construct).
- H2: A positive correlation between 0.30 and 0.50 between WI-7 and MCS (related, but dissimilar construct).
- H3: A correlation below 0.30 between WI-7 and PCS (unrelated construct).
- H4: The correlation between WI-7 and SCL-8 is higher than the correlation between WI-7 and MCS, which in turn is higher than the correlation between WI-7 and PCS.

Hypotheses regarding discriminative validity:

- H5: Patients with at least one psychiatric comorbidity have a mean WI-7 score at baseline that is at least 15 points lower (indicating worse health) than patients without psychiatric comorbidity.
- H6: Men have a mean WI-7 score at baseline that is at least 10 points lower (indicating worse health) than women.

A hypothesis is considered confirmed if the estimate is in accordance with it. Following COSMIN, construct validity is considered sufficient if at least 75% of the hypotheses are confirmed. Focus on effect sizes and CI's, not statistical significance (see slide 39).

Table 4. Correlations between WI-7 and the other scores at baseline

WI-7 vs.	n	Pearson r (95% CI)	Spearman ρ (95% CI)
SCL-8	64	0.68 (0.52 ; 0.79)	0.67 (0.51 ; 0.79)
SF-36 MCS	63	0.44 (0.21 ; 0.62)	0.45 (0.23 ; 0.63)
SF-36 PCS	63	0.38 (0.14 ; 0.57)	0.38 (0.15 ; 0.57)

95% CI's based on Fisher's z-transformation; for Spearman's ρ computed on the ranked variables.

Table 5. WI-7 at baseline by psychiatric comorbidity and by gender

	n	Mean	SD	95% CI
No psychiatric comorbidity	28	57.0	19.8	49.3 ; 64.7
Psychiatric comorbidity	36	33.0	23.6	25.1 ; 41.0
Difference (no − yes)		24.0		12.9 ; 35.1 (p < 0.001)
Women	43	45.2	25.5	37.3 ; 53.0
Men	21	40.1	24.0	29.2 ; 51.1
Difference (women − men)		5.0		−8.3 ; 18.4 (p = 0.45)

Differences estimated by unpaired t-test assuming equal standard deviations in the two groups.

- Evaluate the hypotheses regarding convergent and discriminant validity (H1–H4) using Table 4, and check the assumptions underlying the analyses using the graphs.
- Evaluate the hypotheses regarding discriminative validity (H5–H6) using Table 5, and check the assumptions underlying the analyses using Table 5 and the graphs.
- Discuss the rationale behind each hypothesis. Is the sample size adequate for each of the analyses (see slide 43), and how does it affect your conclusions?
- Summarize and interpret your findings. How many of the hypotheses were confirmed, and what do you conclude about the construct validity of the WI-7?

EXERCISE 1.4

Think of your own data and choose a specific variable/scale. Specify hypotheses regarding direction and strength of correlations between your chosen scale and other variables/scales (see slide 38).