Advanced couse in questionaire research
Solutions to exercises, day 1

EXERCISE 1.1 (criterion validity, continuous instrument and gold standard)
The data we will consider in this exercise, "*criterion validity sf12.dta*", 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 timepoints.

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). As a help you can find SF-36 norm data from a general U.S population, as well as from a Danish sample of patients with health anxiety, below. Below you find all the Stata commands that are necessary to do the analysis.

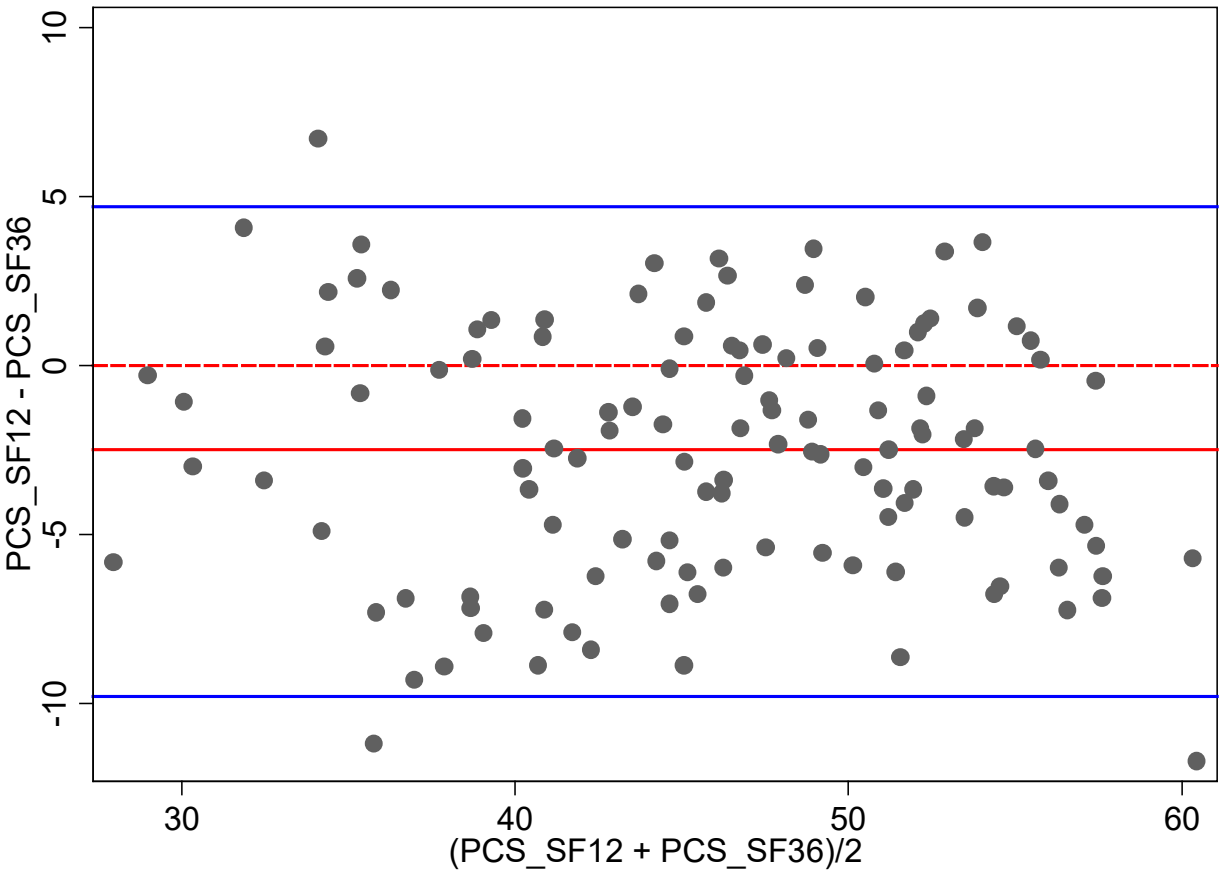
1. Go through step two and three in the process of criterion validation (see slide 13 from todays lecture). That is; discuss whether the choice of gold standard and sample is appropriate for the purpose.
   The choice of gold standard is acceptable since we are trying to validate a short version (SF-12) of an already existing measurement instrument (SF-36). Remember that validity should always be assessed in the population of interest. The purpose of the study was to asses the criterion validity of the SF-12 in people with severe health anxiety. The SF-12 and the SF-36 was used in the same sample, so assuming that this sample is an adequate representation of all people with severe health anxiety the sample is appropriate. But, notice that the data originates from a RCT, and not from a validation study. It is, in general, not a good idea to use the same data to show treatment efficacy and validate the instrument score.

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

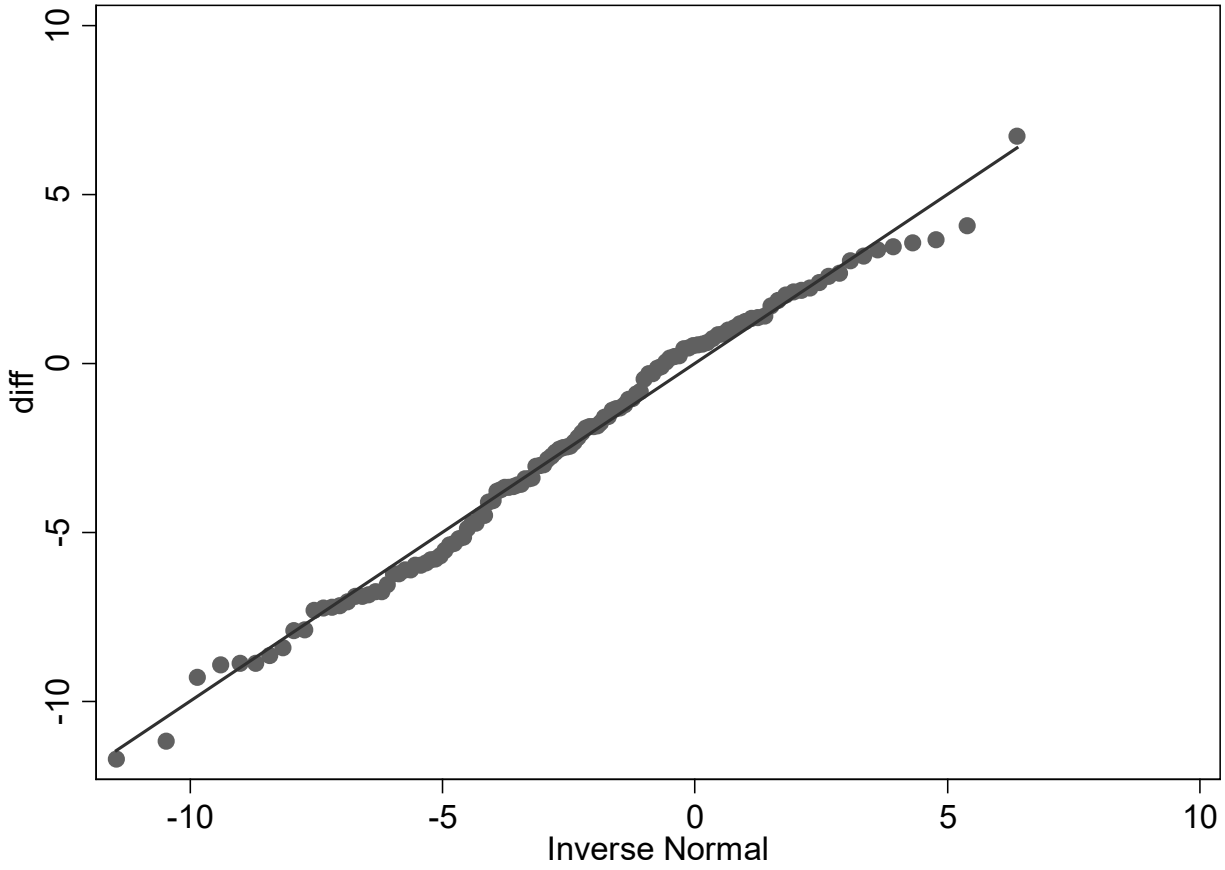
2. Find the limits of agreement (aka. the 90% prediction interval).
   The limits of agreement are (-9.9 ; 4.7), and can be found using the **`-centile-`** command in Stata.

3. Test whether there is a systematic difference between the two scores.
   That is, find the mean difference between the two scores. As usually this is done by a paired t-test. The mean PCS score is 44.8, 95% CI (43.5 ; 46.2), on SF-12 and 47.4, 95% CI (46.0 ; 48.8), on SF-36, corresponding to a mean difference of -2.5, 95% CI (-3.2 ; -1.9), $p < 0.001$. This means that on average the PCS score from SF12 is 2.5 points lower than the PCS score from SF36.

4. Make the Bland-Altman plot.



5. Check the assumptions (see slides from today).
    i. Independence is as always checked by going through the design. The important part is to make sure that people are not sampled in clusters (eg. families, schools etc.). This was not the case in this study, and hence we have no reason to doubt that the differences are independent.
    ii. Assumption 2 is checked by looking at the Bland-Altman plot. There should be no obvious patterns in the distribution of the points. Here everything looks fine.
    iii. Normality is assessed by graphical inspection of a QQ-plot (see below). Since the points lie approximately on the straight line we have no reason to reject the assumption of a normality.



6. Make a summary of your findings and discuss the results.
   Criterion validity of the SF-12 was assessed by calculating the limits of agreement. Since all the assumptions was fulfilled we have no reason to doubt the results from our analysis. We saw that the PCS score was, on average, 2.5 (1.9 ; 3.2) points higher on the SF-36 as compared to the SF-12, and that this difference was statistically significant different from zero, $p < 0.001$ (if the limits of agreement corresponds to our predetermined required level of agreement, this is not necessarily a problem). Remember that observations within the limits of agreement are likely due to measurement error and that observations outside the limits of agreement correspond to real “differences”. Hence if the limits of agreement are far apart (as determined by our a priori specified levels of agreement) there is a high amount of measurement error and therefore we don’t have criterion validity. Our limits of agreement was (-9.8 ; 4.7). Since this interval is wider (even when subtract the mean difference of 2.5) that our prespecified level of agreement (-5.0 ; 5.0) we don’t have criterion validity.

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 data can be found in the file "*criterion validity cancer.dta*".

The researchers hypothesized that the new instrument is as good as the gold standard, with a sensitivity of 85%, specificity of 95%, positive predicted value (PPV) of 80% and a negative predicted value (NPV) of 90%. Below you find all the Stata commands that are necessary to do the analysis.

1. Make a crosstable comparing the two methods. Identify the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

test	criterion 0	1	Total
0	14,696	164	14,860
1	231	969	1,200
Total	14,927	1,133	16,060

   We see that TP = 969, TN = 14696, FP = 231 and FN = 164.

2. Calculate the sensitivity and specificity by "hand".
   Using the **-display-** command in Stata we find that

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{969}{969 + 164} = 0.855$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{14696}{14696 + 231} = 0.985$$

3. Calculate the negative and positive predicted value by "hand".
   Again, using the **-display-** command in Stata we find that

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{969}{969 + 231} = 0.808$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{14696}{14696 + 164} = 0.989$$

Download the **-diagt-** command in Stata by typing **-ssc install diagt-** in Stata's command window.

4. Use the **-diagt-** command to find the sensitivity, specificity, the negative predicted value, the positive predicted value, and their 95% CI's.
   The output from the **-diagt-** command can be seen below

criterion	test Pos.	Neg.	Total
Abnormal	969	164	1,133
Normal	231	14,696	14,927
Total	1,200	14,860	16,060

True abnormal diagnosis defined as criterion = 1

			[95% Confidence Interval]	
Prevalence	Pr(A)	7.1%	6.7%	7.46%
Sensitivity	Pr(+|A)	85.5%	83.3%	87.5%
Specificity	Pr(-|N)	98.5%	98.2%	98.6%
ROC area	(Sens. + Spec.)/2	.92	.91	.93
Likelihood ratio (+)	Pr(+|A)/Pr(+|N)	55.3	48.5	62.9
Likelihood ratio (-)	Pr(-|A)/Pr(-|N)	.147	.128	.169
Odds ratio	LR(+)/LR(-)	376	305	464
Positive predictive value	Pr(A|+)	80.8%	78.4%	82.9%
Negative predictive value	Pr(N|-)	98.9%	98.7%	99.1%

5. Why are the 95% CI's for the specificity and negative predicted value so narrow compared to the CI's for the sensitivity and positive predicted value?
   This is due to the fact that both the sensitivity and specificity directly depends on the 14.860 true negatives in our data.

6. Summarize and interpret your findings.
   The sensitivity was 85.5 (83.3 ; 87.5)% which corresponds to our hypothesis since 85 is contained in the 95% CI.

   The specificity was 98.5 (98.2 ; 98.6)%) and since the lower boundary of the 95% CI is higher than 0.9, we cannot reject the hypothesis that the specificity is at least 0.9.

   The positive predicted value is 80.8 (78.4 ; 82.9)% which also corresponds to out hypothesis since 0.8 is contained in the 95% CI.

   The negative predicted value is 98.9 (98.7 ; 99.1)% and since the lower boundary of the 95% CI is higher than 0.9, we cannot reject the hypothesis that the specificity is at least 0.9.

   Since all our hypotheses checked out the researchers found the instrument sufficiently valid for its purpose.

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, but now our focus is on assessing the convergent, discriminant and discriminative validity of the WI-7 (see page for a description of WI-7). The data can be found in the file "*construct validity.dta*".

Hypotheses regarding convergent and discriminant validity:

- A moderate to strong positive correlation of 0.5 between WI-7 and MCS. A strong correlation of 0.8 between WI-7 and SCL-8.

- A low correlation between of 0.2 between WI-7 and PCS

Hypotheses regarding discriminative validity:

- Men had a 10 point lower (indicating better health) mean score than women on the WI-7 at baseline.
- Patients with at least one psychiatric comorbidity had a 15 point lower score (indicating worse health) than patients without psychiatric comorbidity on the WI-7 at baseline.

The `-corrci-` command can be used to compute 95% CI's for the Pearson correlation between two or more numerical variables (and with a little extra manual labour it can also be used to compute 95% CI's for the Spearman correlation). You can download the command bytyping `-search corrci-` in Stata's command window and follow the instructions.

1. Test the hypotheses regarding convergent and discriminant validity, and check the assumptions underlying the tests.
   We assume that the observations are pairwise independent. From the scatter- and QQ-plots we see that there is an approximate linear relationship between WI-7, MCS, PCS and SCL-8, and that they all follow (approximately) follow a normal distribution. Hence we can calculate the Pearson correlation to check our hypotheses. If the two last assumptions did not check out we could instead have computed Spearmans rank correlation (see the do-file for how to calculate Spearmans rank correlation and the corresponding 95% CI using the `-corrci-` command).

   The correlation between WI-7 and MCS is 0.44 (0.21 ; 0.62) and the correlation between WI-7 and SCL-8 is 0.68 (0.51 ; 0.79). Thus we cannot reject the hypotheses regarding WI-7 and MCS, but we reject the hypothesis regarding WI-7 and SCL-8.

   The correlation between WI-7 and PCS is 0.38 (0.14 ; 0.57) and hence we cannot reject the hypothesis.

2. Test the hypotheses regarding discriminative validity, and check the assumptions underlying the tests.
   We assume that the observations are independent. Since the data follow a normal distribution, (see the QQ-plots in the do-file) and the standard deviations are of similar size (see p-value from the **`-sdtest-`** commnds in Stata in both gender and comorbidity groups. We can compare the groups using an unpaired t-test.

   We see that the mean WI-7 score is 57.0 (49.3 ; 64.6) in the group without comorbidity and 33.0 (25.1 ; 41.0) in the group with comorbidity. Hence the group without comorbidity has, on average, a WI-7 score that is 24.0 (12.9 ; 35.1) points higher than the group with comorbidity. Since 15 is contained in the 95% CI we cannot reject our hypothesis.

   We see that men, on average score 5.1 (-8.3 ; 18.4) lower than women. Since 10 is contained in the 95% CI we cannot reject the hypothesis.

3. Summarize and interpret your findings
   All, but one, of the hypotheses could not be rejected, and hence the researchers found that the WI-7 is sufficiently valid for its purpose.