University of York Department of Health Sciences

M.Sc. in Evidence Based Practice

Measuring Health and Disease

Suggested Answers For Assessment, June 2007

Question 1.

This question relates to the paper ‘Muscle strength testing with one repetition maximum in the arm/shoulder for people aged 75 — test-retest reliability’.

(a) *In Figure 3 we see the results of sessions 1 and 2 in kg for the combined group, with the line of equality. What is the line of equality? Why is used here, rather than a regression line?* The line of equality is the straight line on which observations would fall for each patient if the results of sessions 1 and 2 were exactly the same. If there is a consistent bias, so that one session tends to produce higher measurements than another, this will be reflected by more of the points being on one side of the line than on the other. A regression line is fitted to the data and so always goes though the middle of the plotted points. It would not show bias. Also, we expect the slope of the regression line to be less than one and the line would give a misleading view of the relationship between repeated measurements.

(b) *In Figure 3, the authors quote a correlation coefficient $r = 0.97$, $P<0.0001$. What null hypothesis is being tested and what is its value here? What feature of the study design is likely to inflate the value of this correlation coefficient? What feature of the results would be missed by the use of a correlation coefficient?* The null hypothesis being tested is that the measurements for session 1 and for session 2 are not related. We expect two measurements of the same thing to be related, so this is not very interesting. All it tells us is that we are measuring something. The correlation coefficient may be inflated because we have two different groups of subjects, one of which had had muscle strength training and the other had not. If those who had experienced training had a different mean muscle strength to the untrained group, this would increase the variability between subjects and so increase the correlation. A correlation coefficient would miss any systematic bias between groups.

(c) *Figure 4 shows 95% limits of agreement analysis comparing measurements in the first and second sessions for subjects in group 1. What are 95% limits of agreement? What is the purpose of this plot and what does it show here? What should we conclude about the limits of agreement estimated?* Limits of agreement are a way of measuring the agreement between observations made under two conditions, such as by different methods of measurement. Here we have the first and second session, so we are looking at the effect of practice. The limits are the mean difference between pairs plus or minus 1.96 (or 2) standard deviations of the difference, which will include 95% of possible differences. The plot is to check that the mean and standard deviation of the differences are unrelated to the magnitude of the measurement. In Figure 4, this is not the case, as the differences appear more variable for higher strength measurements. This would make the limits too wide for subjects with lower strength and too narrow for stronger subjects.

Question 2.

This question relates to the paper ‘Development of a pictorial scale of pain intensity for patients with communication impairments: initial validation in a general population’.

(a) *In Table 3, the authors present the test-retest reliability of the visual analogue scale and numeric rating scale by intraclass correlation coefficients. What is meant by ‘test-retest reliability’? Why are intraclass correlation coefficients used to estimate it for these two scales?* ‘Test-retest reliability’ means looking at how closely measurements made on two different occasions are related. We usually do this for quantitative data using a correlation coefficient. An intraclass correlation is used when we have repeated observations on each of a group of subjects and which we regard as equivalent. It does not make any distinction between the first and second observations, as the ordinary product-moment correlation coefficient would. It is used here because we have quantitative data, either continuous or discrete, and we assume that the two measurements are true replicates, no changes having taken place between them.

(b) *In Table 3, the authors present the test-retest reliability of the scale of pain intensity evaluated by weighted kappa coefficients. What is a weighted kappa coefficient and how would we interpret the kappa values presented here? In the Methods, these are described as quadratic-weighted kappa statistics. What does this mean and why is it useful information?* Kappa is a measure of agreement between two assessments or measurements using a categorical variable. It uses the proportion of cases for which there is agreement and the proportion we would expect to agree if there were the agreement we would expect by chance in the absence of any relationship between the two assessments. We take the amount by which the observed agreement exceeds chance, agreement minus expected agreement, divided by the maximum value this could have, one minus expected agreement. The maximum value is 1.00, for perfect agreement. The ordinary kappa statistic does not take into account any ordering of the categories. All disagreements are treated the same. If we have ordered categories, such as the six visual pain categories, we might want to regard a disagreement between two observations close together as less important than a disagreement where observations are at opposite ends of the scale. Weighted kappa does this, by attaching arbitrary weights to pairs of categories. Quadratic weights attach weight to the disagreement equal to the square of the number of categories apart. It is useful because we should know what the weights are to interpret weighted kappa.

(c) *In Table 3, the authors present 95% confidence intervals for the weighted kappa and the intraclass correlation coefficients. Why are several of these obviously wrong?* Some of these confidence intervals have upper limits greater than 1.0. As neither the intra-class correlation coefficient nor the weighted kappa statistic can be greater than 1.0, it is impossible for these values to be true for the population. Hence the confidence intervals must be wrong.

Question 3.

This question relates to the paper ‘Validity of Hamilton Depression Inventory in Parkinson’s Disease’.

(a) *Table 2 shows sensitivity, specificity, positive predictive value, and negative predictive value for different cut-off points and different target diagnoses. What are ‘sensitivity’, ‘specificity’, ‘positive predictive value’, and ‘negative predictive value’? For each scale, as we go down the table, sensitivity decreases and specificity increases. Why is this?* ‘Sensitivity’, ‘specificity’, ‘positive predictive value’, and ‘negative predictive value’ are statistics used to examine the properties of diagnostic tests. ‘Sensitivity’ is the proportion of those who have the disease in question who are positive on, or detected by, the test. ‘Specificity’ is the proportion of those who do not have the disease who are negative on, or detected as not having the disease by, the test. ‘Positive predictive value’ is the proportion of those who are positive on the test who have the disease. ‘Negative predictive value’ is the proportion of those who are negative on the test who do not have the disease. The table shows the effect of changing the cut-off value on the scales. If we raise the cut-off value, fewer subjects will be classified as positive. Hence fewer cases of the disease will be detected and sensitivity will go down. On the other hand, there will be more people not detected with the disease and more of those without the disease will be correctly classified and specificity will go up.

(b) *Figures 1 and 2 show ROC curves. What is an ROC curve and why might it be useful? Why are the areas under the ROC curves given and how can they be interpreted?* A ROC curve is a plot of sensitivity and specificity for varying cut-off values. We usually plot sensitivity, the proportion of people with the disease detected by the test, against one minus specificity, the proportion of those without the disease detected as positive by the test. A ROC curve shows us how sensitivity and specificity are related and can help us decide which cut-off will give the best compromise between high sensitivity and high specificity. The maximum possible area under the ROC curve is 1.0 and the closer to this the observed area is, the better the test will be. It enables us to compare tests. The area under the ROC curve estimates the probability that a member of one population (disease positives) chosen at random will have a test score which exceeds that for a member of the other population (disease negatives) chosen at random. In this case HAMD-17 has a larger area than either of the other tests and would appear to be a more effective tool for diagnosing depression in this population.

(c) *The authors describe this as a study of the concurrent validity of the HDI-17. What is meant by ‘concurrent validity’? In this study, would this be best described as an aspect of criterion validity or of construct validity, and why?* Concurrent validity is examining the relationship between the instrument to be validated and other variables measured at the same time. Here, the relationship between the HDI-17 and a clinical interview is examined. The clinical interview is described as a gold standard and the authors look at how well the HDI-17 can be used to predict the result of the interview. Hence the interview is regarded as the criterion and we have an example of criterion validity.