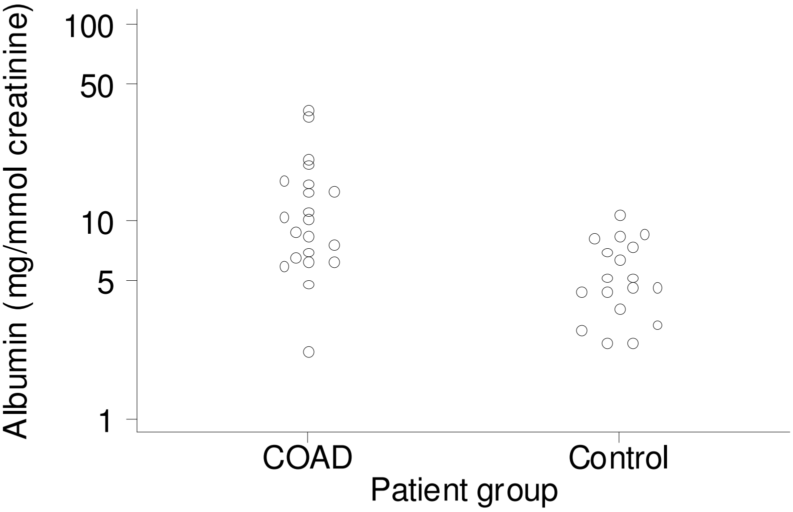
Clinical Biostatistics

Suggested Answers to Exercise: Transformations

Question 1

(a) *Why was the logarithmic transformation of the data used?* From the figure we can see that the albumin data have skew distributions, with a few observations much greater than the rest. The variability is also much greater in the COAD patients where the mean is higher. Using a logarithmic scale (see figure below) stretches the bottom of the scale and compresses the top, making the distribution more like the Normal. The log transformation also makes the variances more uniform. The transformed data matches the assumptions of the t test more closely.



(b) *How can we interpret the antilog of the difference and its confidence interval?* The antilog of the difference between the means on the log scale is the ratio of the means on the natural scale. The antilog of a mean is the geometric mean, so the antilog of the difference is the ratio of two geometric means. So the point estimate for the ratio of the geometric mean albumin for COAD patients is twice that for controls, with interval estimate 1.4 to 3.0.

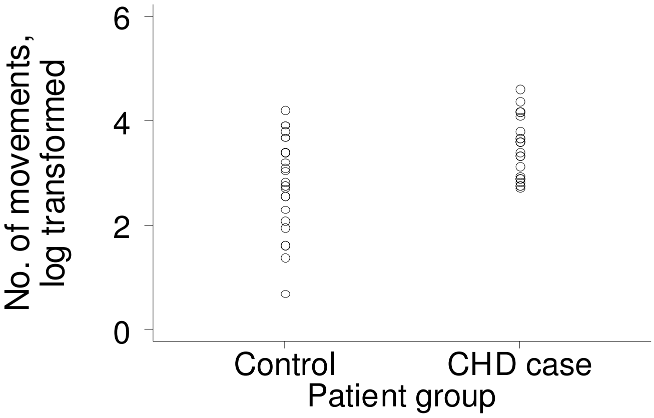
Question 2

(a) *What is meant by a two sample t test and what assumptions about the data does it require?* The two sample t test is a statistical technique used when we wish to compare the means of two independent groups. The method requires that the data come from Normal distributions in each population and that the variance is the same in both populations.

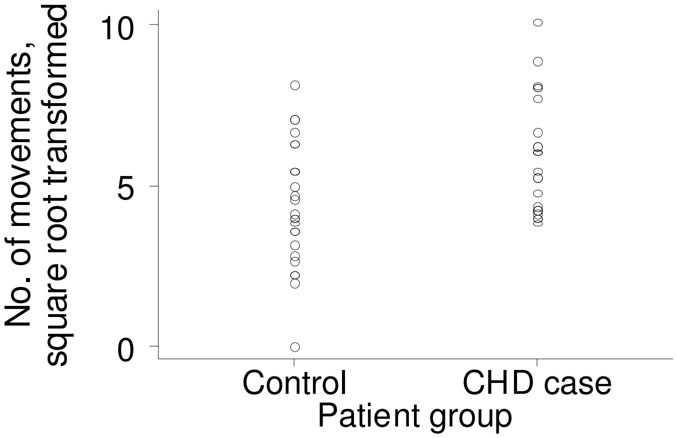
(b) *Are these assumptions likely to be met here?* Inspection of the data re-drawn from the original paper shows that the data are positively skew and the variance is slightly higher in the group with the higher mean (CHD cases). This suggests that the data should be transformed prior to analysis.

(c) *How could we use a transformation for these data? What problems would be caused by the subject who had zero arm movements recorded?* The figure below shows the data after logarithmic and square root transformation. The observation which is zero causes a problem for log transformation, as zero has no logarithm.

We have set this to the log of half the next largest observation (4), which is a bit arbitrary but better than making it a missing value. Another possibility would be to add a small number, such as one, before transforming. The log transformation corrects the skewness. However, the variance in the CHD cases is now less than in the controls so the transformation has over-corrected.



A square root transformation looks better.



The variable is a count of the number movements in a fixed time and therefore the square root transformation is the first transformation we would try.

The p-value from the two sample t test using the square root transformation is just smaller than that for the log transformation (untransformed: P=0.0134, log: P=0.0071, square root: P=0.0066). Making the data fit the assumptions of the method better usually reduces the P value.

Question 3

(a) *Comment on the presentation of the P value.* It would be better to give the actual P value, which is 0.08.

(b) *Why is the paired t test not suitable for these data?* The differences clearly have a very skew distribution. Also, there are 8 zero differences, forming a clump at one end of the distribution.

(c) *Could a transformation of the antibody levels make the paired t test appropriate?* No, the subjects with zero differences will have zero whatever transformation is applied, so the distribution of differences cannot be made Normal.

(d) *What other approaches could be used?* We could use a sign test, or a method based on rank order, the Wilcoxon paired sample test.

(e) *Using the table overleaf, carry out an appropriate significance test.* There are 8 zero differences and 12 differences with a sign. There is one difference with a negative sign. We have 1 negative and 11 positives. Using the table, this would be significant if the number of negatives were 2 or less. Hence we have a significant difference, P<0.05. If we do the test exactly, we find P = 0.006.

(f) *What would this do to the conclusions of the study?* The difference was significant and using an appropriate test showed that there was good evidence for an increase in antibody levels. The original paired t test analysis was incorrect because its assumptions were not met by the data and the non-significant P value was not valid. They failed to detect a difference because they used an inappropriate analysis.