chapter 2-2

Norms and Rating Scales

Introduction

In the previous chapter we discussed descriptive statistics. One group of descriptive statistics that we use extensively in Biomedical Physiology and Kinesiology, so-called 'rating statistics', gives comparisons or ratings of values against meaningful distributions. This chapter is dedicated to that group, which includes z-scores, percentiles, and arbitrary scales such as T Scores, Hull Scores and Stanines – each has valuable properties that can help us to make judgements about data.

When a subject has been measured, the first question they almost always ask is, "Was that good?" It is the rating statistics that give you the answer, assuming there is an appropriate comparative distribution. We often call the comparative distribution a 'Standard' or a 'Norm' (short for normative distribution). The first step is to determine if an appropriate Norm exists for use with the subjects we have. Then, the second step is to decide which rating score to use.

Standard Scores

A standard score is a quantification of a measured score (observation) in comparison to a distribution of the score from some normative (comparison) sample. The data are assumed to be normally distributed. Numerically, a standard score indicates how many standard deviations an observation is above or below the mean of the comparison distribution. Standard scores are also called **z-scores**. To calculate a z-score, the difference between the score (X) and mean of the distribution ($\overline{X}$) is divided by the standard deviation of the sample (s).

$$Z = \frac{(X - \overline{X})}{s}$$

An advantage of the z-score is that a single number quantifies the measured value and its location with respect to the normative distribution. Thus, a z-score of -0.5 tells you that the measured score is half a standard deviation below the mean of the distribution, therefore a little below average. Whereas a z-score of +2.5 tells you that the measured value is quite large and uncommon relative to the norm.

Z-scores are dimensionless because the standard deviation is divided into the difference between the observation and the mean. In the case of height, centimetres are divided by centimetres. Similarly for weight, kilograms are divided by kilograms, and so on. This is a valuable property because z-scores for measures with different units can be compared. Consider the example of measuring an individual's grip strength and VO_2max. How would you determine if his grip strength is better or worse than his VO_2max? It is not possible to address this question without normative data. However, if normative data are available, then you can calculate z-scores and compare them. If his grip strength corresponds to a z-score of +1.2 and his VO_2max corresponds to a z-score of -0.2, then you can conclude that this man's strength is better, relative to the norm, than his aerobic fitness.

A warning here is that the norms used for the two variables must be comparable. Ideally the two distributions would be from the same sample data. If you have distributions from different samples then the comparison becomes invalid if the samples are intrinsically different. Therefore, it is not a recommended tactic to gather arbitrary normative distributions, just to have norms for each test in your battery. It is much better to have a homogeneous set of norms based upon the same sample. What that sample is depends on the use you have for it. Often a national sample is deemed appropriate. In Canada, the Canada Fitness Survey (1981) fulfils that role for many fitness and anthropometric measures. It forms the basis for the current Canadian Physical Activity, Fitness & Lifestye Approach Protocol (CPAFLA) used in personal fitness appraisals in Canada. It however, does not include every fitness test that one might want to evaluate, so you are limited to the measures that it includes.

Composite Scores: Because the z-score has no units it is possible to combine z-scores for different variables into a single composite score. If you have a test battery you want to administer then you can rate each test against its own norm using z-scores.

But, if you want a single composite rating of the performance of a subject in all the tests, this too can be achieved using z-scores. Because the z-scores have no units, you can average the z-scores for all the tests for each subject to come up with an overall rating. Table 2-2.1 shows the results of calculating z-scores for a subject in four fitness tests: sum of 5 skinfolds, grip strength, vertical jump, and shuttle run. Each test has different measurement units; thus z-scores help us equate the variables by comparing to normative distributions and giving unitless scores. There are two sets of z-scores in table 2-2.1 based upon the same data of an individual. If we simply calculate z-scores for the

Variable	z-scores for profile A	z-scores for profile B
Sum of 5 Skinfolds (mm)	1.5	-1.5
Grip Strength (kg)	0.9	0.9
Vertical Jump (cm)	-0.8	-0.8
Shuttle Run (sec)	1.2	-1.2
Overall Rating	0.7	-0.65

Table 2-2.1: Z-Scores for profiles A and B shown in Figure 2-2.1

measured values we will get the z-scores shown as profile A (also shown in Figure 2-2.1). An overall rating of the subject on the four tests can then be simply achieved by averaging the four z-scores. Thus the overall rating for profile A is a z-score of +0.7. However, if our inference is the bigger the z-score the better, this is incorrect for two of the scores. For grip strength and vertical jump it is true that higher scores are better or represent greater fitness. However, the opposite is true for skinfolds and shuttle runs. A fatter person will have higher skinfolds, and a less fit person will have a slower or larger time for the shuttle run. We therefore need to reverse the rating for these two variables so as to show that a lower score is preferable. In our example, this was achieved by simply reversing the sign of the z-score. This holds true if the scores are normally distributed. In calculating z-scores, in order to get a reverse rating you can simply change the z-score equation around to get a reverse z-score as it were. Our example profile B shows the result of correcting the z-scores for skinfolds and shuttle run. The composite overall rating is now easier to interpret. The overall rating in profile B of -0.65 gives a truer indication of the situation. The overall rating is less than average (negative z-score), based upon the person being above average in grip strength, but below average for skinfolds, vertical jump, and shuttle run performance.

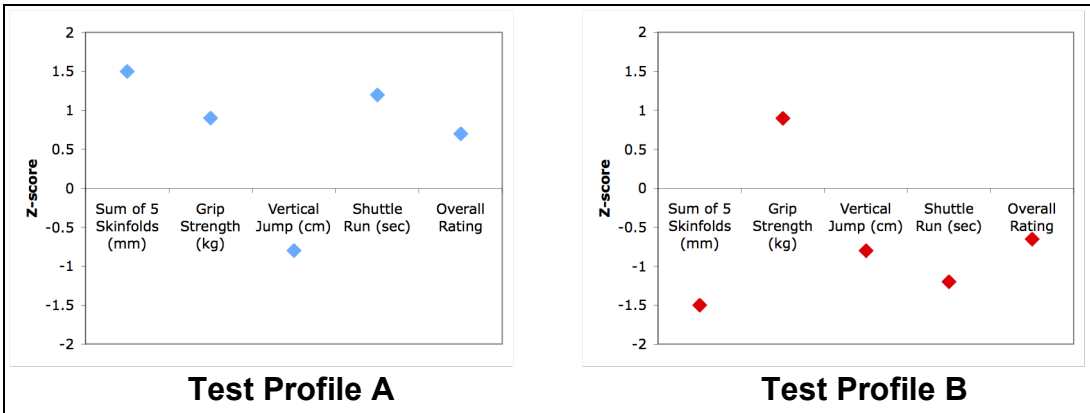

Figure 2-2.1: Z-Score profiles A and B

In combining z-scores in this way, the assumption is that it makes sense to equally weight the component scores. For instance, in our example we infer that all four tests should contribute equally to the overall rating. Without evidence to the contrary, this is a reasonable approach. On the other hand, one could arbitrarily assign differential weighting to the four tests based upon experience or intuition, but such an approach is hard to defend scientifically. However, if a criterion score exists, it is possible to predict that score from the component tests and determine their relative contribution to the prediction. In our example, a criterion score might be a rating of subjects in an athletic event based upon performance scores or subjective ratings of judges. This criterion score is then predicted from the four tests using linear regression. A special form

of regression is used however, called standardized regression. In standardized regression the independent variables are converted to z-scores before the regression is carried out. The relative size of the regression coefficients then indicates the relative importance of the four tests and could then be used as weighting factors in an overall rating from combination of the four z-scores. Standardized regression will be discussed in more detail in chapter 2-5.

External vs Internal Norm: In the discussion so far, the z-scores have been calculated by reference to the means and standard deviations of an external norm or standard. Z-scores can simply be calculated for all scores in the sample by using the mean and standard deviation of the sample itself. The sample therefore becomes an internal norm for the calculation of the z-scores. Although sometimes by choice, this is often done when no appropriate external norm is available, but the characteristics of the z-score are required for analysis. By definition, assuming the variable is normally distributed, the mean of the sample z-scores will be 0 and the standard deviation 1. Whereas the mean and standard deviation of z-scores in comparison to an external norm would depend upon the extent to which the study sample deviates from the external norm.

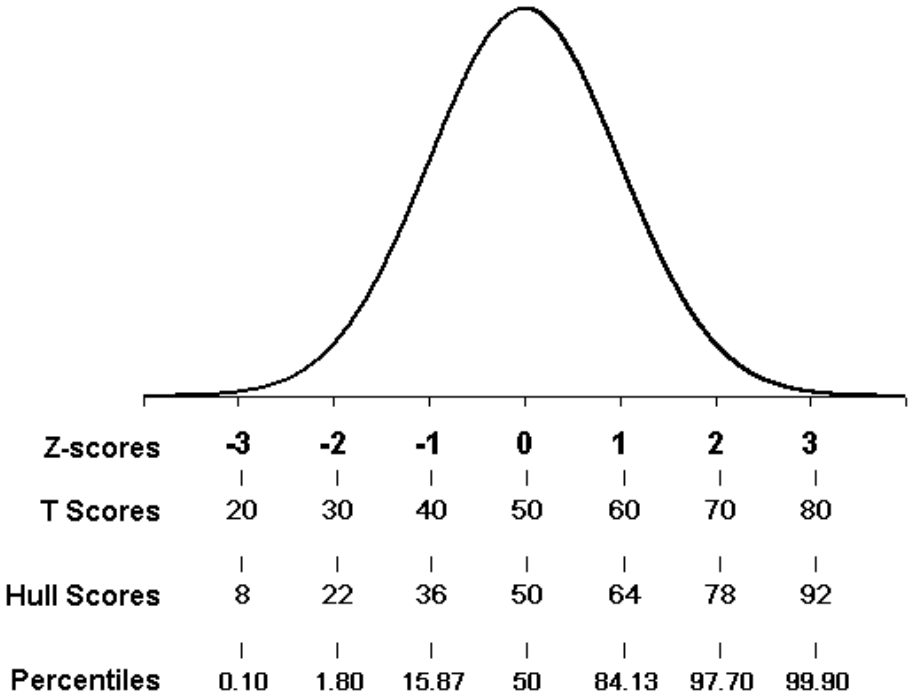

Figure 2-2.2: Arbitrary Rating Scales in comparison to the normal distribution

T Scores

Figure 2-2.2 illustrates several arbitrary scales in comparison to the normal distribution based upon z-scores. Each scale is calculated from the calculated z-score. The first is the T score,

which is the most commonly used of these rating scales. The mean and standard deviation of the T-Score are arbitrarily set at 50 and 10 respectively. The z-score can simply be converted into T-scores by multiplying the z-score by 10 and adding it to 50:

T-Score = 50 + (z-Score x 10)

The T-score is often seen in physical education literature, especially when comparing physical performance scores to normative data. The justification for the use of T-scores over z-Scores is that the values for the T-Score lie between 0 and 100 rather than having awkward numbers such as -0.87 as a z-Score. The distribution of the scores is the same as that for z-scores, and the scores can be used in exactly the same way, it is just that they have a nicer look to the numbers. You just have to remember that 50 is a mean rating, with ratings above or below 50 meaning above or below average ratings respectively. The Hull score discussed below is a similarly derived arbitrary rating scale.

Hull Scores

The arbitrary nature of the defined T-Score mean and standard deviation is highlighted by the existence of the Hull Score. In the Hull Score, the mean is 50 but the standard deviation is 14. The z-score can be converted into Hull Scores by multiplying the z-score by 14 and adding it to 50:

Hull Score = 50 + (z-Score x 14)

This results in a score that conveniently uses more of the range of numbers from 0 to 100. For instance, as mentioned previously the range in a sample can be estimated as 3 x standard deviations above and below the mean. This range would be defined as T-scores of 20 to 80, whereas in Hull Scores the values would be 8 and 92. Since values are rarely found outside the range, a score of 90 is highly unlikely as a T-score but more likely as a Hull Score. The choice of standard deviation of 14 is therefore purely arbitrary based upon an attempt to use as much of the range of numbers between 0 and 100. You could in fact define any arbitrary scale you wanted. All you need to do is set the values for mean and standard deviation to define your new score. If you decided that the mean should be 500 and the standard deviation should be 100 you would have a score that uses most of the 0 to 1000 range. If this sounds familiar, it is how the SAT and MCAT results are reported. The decision on what mean and standard deviation is used simply depends upon the form of the numbers that is required. It should be noted that the distribution of the scores will always stay the same, whether z-scores, Hull scores or T-scores are used.

Percentiles

The most common form of rating a measure is to express it as a percentile. The percentile is defined as that percentage of the comparison sample that are **at or below** the given score. Thus if a value is rated as being at the 25th percentile, then 25% of the comparison sample is **at or below** that value, with 75% being above. So far it has been stated that the scores assume the normal distribution. Figure 2-2.2 shows the relationship of the various scores to the normal distribution. As discussed in chapter 2-1, -1 to +1 standard deviations encompasses 68.26% of the population. In fact, when normally distributed, any z-score can be associated with the specific percentage of the population up to and including that score. Thus, a z-score of +2 has 97.7% of the population at or below it and it could be termed the 97.7th percentile.

Using a table found in most stats books you can find the area under the normal distribution (therefore percentage in the population) for any given standard score. Using this, you can determine how many standard deviations above or below the mean any specific percentile is located. Say for example you have looked up the table and found that the 97.5th percentile is located 1.96 standard deviations (a z-score of 1.96) above the mean. If you know the mean is 56 and the standard deviation is 2, then you can determine that the value for the 97.5th percentile is 56 + (2 * 1.96) = 59.2. This method does not work well if the data are not normally distributed. If the data are positively skewed, as is the case for variables like weight and skinfolds, you would tend to underestimate the percentile values below the mean and overestimate the percentile values above the mean as the standard deviation does not truly represent the variability on both non-symmetrical sides of the mean.

Percentile	Z-score for Percentile	Predicted Percentile value based upon Mean = 170 Sd = 10
5	-1.645	153.55
25	-0.675	163.25
50	0	170
75	-0.675	176.75
95	-1.645	186.45

Table 2-2.2: Predicted percentiles based upon mean = 170 and sd = 10, assuming a normal distribution.

If only means and standard deviations are reported it is possible to calculate the required percentiles from these reported mean and standard deviation values for the variable concerned. This relies on the assumption that the data are normally distributed. Assuming normality means that any given percentile has a specific z-score associated with it. Table 2-2.2 shows the results of estimating percentiles for a sample with mean = 170 and sd = 10. The 5th percentile is represented by a z-score of -1.645, the calculated 5th percentile is therefore 170 +(-1.645 x 10)

= 153.55. This can be carried out for any desired percentiles. The z-score for the required percentile can be found by reference to a table of areas under the standard normal curve.

Stanines

The Stanine scale is a categorical scale rather than the continuous values produced by the z-scores, T-scores and Hull Scores. There are nine categories, with the width of the categories being defined by the size of the sample standard deviation. Categories 2 to 8 are defined as being 0.5 standard deviations wide with the mean of the distribution passing through the centre of the category 5. The boundary z-scores for category 5 are therefore -0.25 and +0.25 (as shown in Figure 2-2.3. The other categories increment by 0.5 standard deviations moving out from the centre. Categories 1 & 9 are open ended, encompassing the bottom and top 4% of the distribution respectively. They are open-ended because there is theoretically no maximum or minimum values in the normal probability distribution.

Stanine Category	1	2	3	4	5	6	7	8	9
Percentage of Normal Distribution within Category	4%	7%	12%	17%	20%	17%	12%	7%	4%
Boundary Z-Scores	-1.75	-1.25	-0.75	-0.25	+0.25	+0.75	+1.25	+1.75	
Boundary Percentiles	4	11	23	40	60	77	89	96	

Figure 2-2.3: Stanine Category Cutoff Points

Health Example: T-Scores in Osteoporosis

To diagnose osteoporosis, clinicians measure a patient's bone mineral density (BMD) and then express the patient's BMD in terms of standard deviations above or below the mean BMD for a "young normal" person of the same sex and ethnicity.

$$Osteoporosis\ T-score = \frac{\left(BMD_{\text{patient}} - \overline{BMD}_{\text{young normal}}\right)}{SD_{\text{young normal}}}$$

Although the field of osteoporosis calls this standardized score a T-score, it is really just a z-score where the reference mean and standard deviation come from an **external reference population** (i.e., young normal adults of a given sex and ethnicity).

T-scores are used to classify a patient's BMD into one of three categories:

- T-scores of ≥ -1.0 indicate normal bone density
- T-scores between -1.0 and -2.5 indicate low bone mass ("osteopenia")
- T-scores ≤-2.5 indicate osteoporosis

Decisions to treat patients with osteoporosis medication are based, in part, on T-scores.

For more information about the diagnosis of osteoporosis see Chapter 3 of the National Osteoporosis Foundation's Clinician's Guide to Prevention and Treatment of Osteoporosis (2010). http://www.nof.org/sites/default/files/pdfs/NOF_ClinicianGuide2009_v7.pdf