3

The problem of scale

> Exploratory data analysis is detective work–numerical detective work–or counting detective work–or graphical detective work. A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will will not look in the right places. Equally, the analyst of data needs both tools and understanding (p 1: [**?**])

As discussed in Chapter 1.1 the challenge of psychometrics is assign numbers to observations in a way that best summarizes the underlying constructs. The ways to collect observations are multiple and can be based upon comparisons of order or of proximity (Chapter 2). But given a set of observations, how best to describe them? This is a problem not just for observational but also for experimental psychologists for both approaches are attempting to make inferences about latent variables in terms of statistics based upon observed variables (Figure 3.1).

For the experimentalist, the problem becomes interpreting the effect of an experimental manipulation upon some outcome variable (path B in Figure 3.1 in terms of the effect of manipulation on the latent outcome variable (path b) and the relationship between the latent and observed outcome variables (path s). For the observationalist, the observed correlation between the observed Person Variable and Outcome variable (path A) is interpreted as a function of the relationship between the latent person trait variable and the observed trait variable (path r), the latent outcome variable and the observed outcome variable (path s), and most importantly for inference, the relationship between the two latent variables (path a).

Paths r and s are influenced by the *reliability* (Chapter 5), the *validity* (Chapter 7) and the *shape* of the functions r and s mapping the latents to the observed variables. The problem of measurement is a question about the shape of these relationships. But before it is possible to discuss shape it is necessary to consider the kinds of relationships that are possible. This requires a consideration of how to assign numbers to the data.

Consider the set of observations organized into a "data.frame", s.df, in Table 3.1. Copy this table into the clipboard, and read the clipboard into the data.frame, s.df.[1] A data.frame is an essential element in R and has many (but not all) the properties of a matrix. Unlike a matrix, the column entries can be of different data types (strings, logical, integer, or numeric). Data.frames have dimensions (the number of rows and columns), and a structure. To see the structure of a data.frame, use the *str* function.

The *read.clipboard()* function is part of the "psych" package and makes the default assumption that the first row of the data table has labels for the columns. See ?read.clipboard for more details on the function.

```
> s.df <- read.clipboard()

> dim(s.df)
[1] 7 7
> str(s.df)
```

[1] Because θ is read as X., we add the command `colnames(s.df)[4] <- "theta"` to match the table.

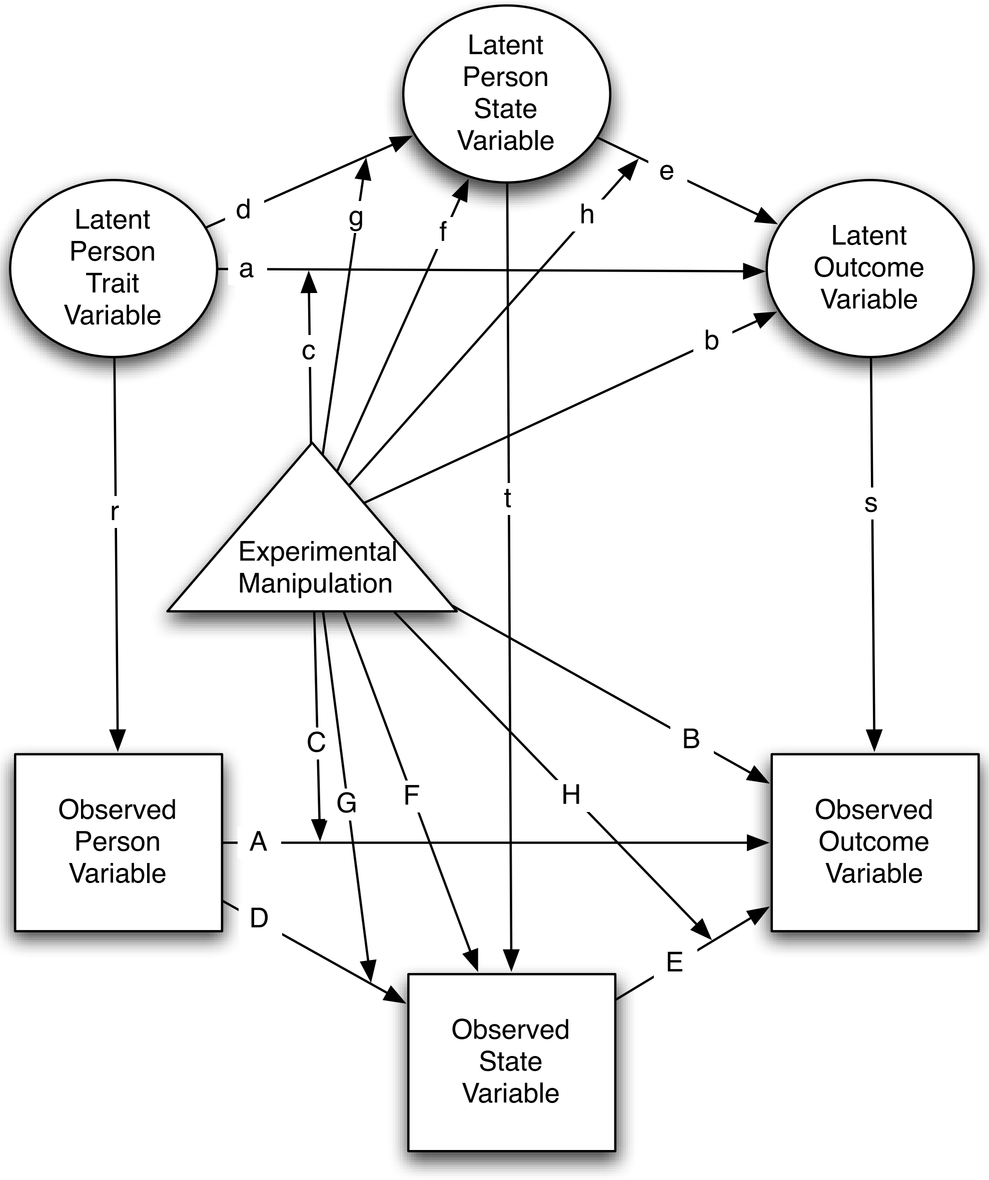
Latent Person State Variable
Latent Person Trait Variable
Latent Outcome Variable
Experimental Manipulation
Observed Person Variable
Observed Outcome Variable
Observed State Variable
d
g
f
h
e
a
c
b
r
t
s
C
G
F
H
B
A
D
E

Fig. 3.1. Both experimental and observational research attempts to make inferences about unobserved latent variables (traits, states, and outcomes) in terms of the pattern of correlations between observed and manipulated variables. The uppercase letters (A-F) represent observed correlations, the lower case letters (a-f) represent the unobserved but inferred relationships. The shape of the mappings from latent to observed (r, s, t) affect the kinds of inferences that can be made(Adapted from [Revelle, 2007])

Table 3.1. Six observations on seven participants

Participant	Name	Gender	θ	X	Y	Z
1	Bob	Male	1	12	2	1
2	Debby	Female	3	14	6	4
3	Alice	Female	7	18	14	64
4	Gina	Female	6	17	12	32
5	Eric	Male	4	15	8	8
6	Fred	Male	5	16	10	16
7	Chuck	Male	2	13	4.	2

```
'data.frame':        7 obs. of  7 variables:
 $ Participant: int  1 2 3 4 5 6 7
 $ Name       : Factor w/ 7 levels "Alice","Bob",..: 2 4 1 7 5 6 3
 $ Gender     : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 2 2
 $ theta      : int  1 3 7 6 4 5 2
 $ X          : int  12 14 18 17 15 16 13
 $ Y          : num  2 6 14 12 8 10 4
 $ Z          : int  1 4 64 32 8 16 2
```

3.0.1 Factor levels as Nominal values

Assigning numbers to the "names" column is completely arbitrary, for the names are mere conveniences to distinguish but not to order the individuals. Numbers could be assigned in terms of the participant order, or alphabetically, or in some random manner. Such nominal data uses the number system merely as a way to assign separate identifying labels to each case. Similarly, the "gender" variable may be assigned numeric values, but these are useful just to distinguish the two categories. In R, variables with nominal values are considered to be *Factors* with multiple *levels*. Level values are assigned to the nominal variables alphabetically (i.e., "Alice", although the 3rd participant, is given a level value of "1" for the "names" variable; similarly, "Females" are assigned a value of "1" on the "Gender" factor).

The "names" and "gender" columns of the data represents "nominal" data (also known as categorical or in R representing levels of a factor), Columns theta, x, and z are integer data, and because of the decimal point appearing in column Y, variable Y is assigned as a "numeric" variable.

3.0.2 Integers and Reals: Ordinal or Metric values?

If the assignment of numbers to nominal data is arbitrary, what is the meaning of the numbers for the other columns? What are the types of operations that can be done on these numbers that allow inferences to be drawn from them? To answer these questions, it is useful to first consider how to summarize a set of numbers in terms of dispersion and central tendency.

3.1 Graphical and numeric summaries of the data

The question is how to best summarize the data without showing all the cases. John Tukey invented many ways to explore one's data, both graphically and numerically [**?**]. One descriptive technique was the "five number summary" which considered the minimum, the maximum, the median, and then the 25th and 75th percentiles. (These later two are, of course, just the median number between the minimum and the median, and between the maximum and the median). The “summary” function gives these five numbers plus the arithmetic mean. For categorical (of Type=Factor) variables, summary provides counts. Notice how it orders the levels of the factor variables alphabetically.

```
> s.df <- read.clipboard()
> summary(s.df)
```

```
 Participant     Name     Gender      theta           X              Y           Z        
 Min.   :1.0   Alice:1   Female:3   Min.   :1.0   Min.   :12.0   Min.   : 2   Min.   : 1.00  
 1st Qu.:2.5   Bob  :1   Male  :4   1st Qu.:2.5   1st Qu.:13.5   1st Qu.: 5   1st Qu.: 3.00  
 Median :4.0   Chuck:1              Median :4.0   Median :15.0   Median : 8   Median : 8.00  
 Mean   :4.0   Debby:1              Mean   :4.0   Mean   :15.0   Mean   : 8   Mean   :18.14  
 3rd Qu.:5.5   Eric :1              3rd Qu.:5.5   3rd Qu.:16.5   3rd Qu.:11   3rd Qu.:24.00  
 Max.   :7.0   Fred :1              Max.   :7.0   Max.   :18.0   Max.   :14   Max.   :64.00  
```

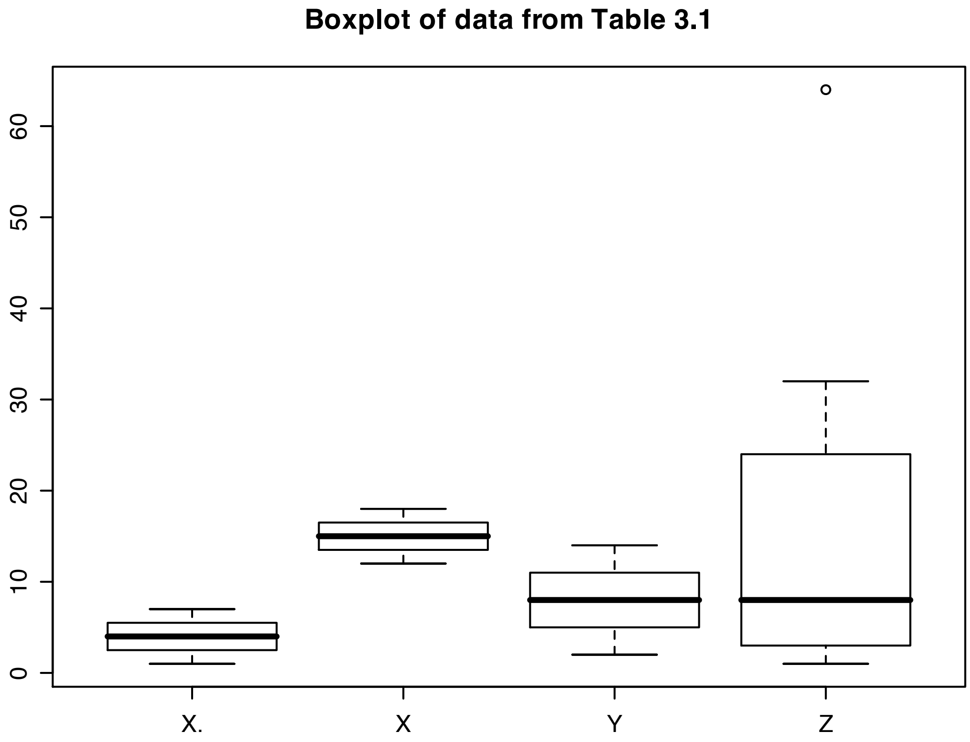

Fig. 3.2. The Tukey box and whiskers plot shows the minima, maxima, 25th and 75th percentiles, as well as the “whiskers” (either the lowest or highest observation or 1.5 times the lower or upper interquartile range, limited to the minimum distance.) Note the outlier on the Z variable).

A graphic representation of the Tukey 5 points is his “BoxPlot” (Figure 3.2) which includes two more numbers, the upper and lower “whiskers”, which are defined as 1.5 the InterQuartileRange (IQR) beyond the upper and lower quartiles. (If the minimum value is less than that distance from the lower quartile, the whisker ends on the data point, similarly for the upper whisker). Several things become immediately apparent in this graph: X is much higher than Y (which has more variability), and z has both greater IQR as well as one very extreme score. Generalizations of the boxplot are "notched" boxplots which give confidence intervals of the median, and "violin" plots which give more graphical representations of the distributions within the distributions.

3.1.1 Sorting data as a summary technique

For reasonable size data sets, it is sometimes useful to sort the data according to a meaningful variable to see if anything leaps out from the data. In this, case, sorting by "name" does not produce anything meaningful, but sorting by the fourth variable, θ, shows that variables 4-7 are all in the same rank order, a finding that was less than obvious from the original data in Table 3.1.

Table 3.2. Sometimes, sorting the data shows relationships that are not obvious from the unsorted data. Two different sorts are shown, the first, sorting by name is less useful than the second, sorting by variable 4. Compare this organization to that of Table 3.1.

```
> s.df <- s.df[order(s.df[,4]),]     #order the data frame by the fourth variable
> n.df <- s.df[order(s.df[,2]),]     #create a new data frame, ordered by the 2nd variable of s.df
> sn.df <- cbind(n.df,s.df)          #combine the two
> sn.df                                           #show them
```

```
Participant  Name Gender theta  X  Y  Z Participant  Name Gender theta  X  Y  Z
3           3 Alice Female     7 18 14 64          1   Bob   Male     1 12  2  1
1           1   Bob   Male     1 12  2  1          7 Chuck   Male     2 13  4  2
7           7 Chuck   Male     2 13  4  2          2 Debby Female     3 14  6  4
2           2 Debby Female     3 14  6  4          5  Eric   Male     4 15  8  8
5           5  Eric   Male     4 15  8  8          6  Fred   Male     5 16 10 16
6           6  Fred   Male     5 16 10 16          4  Gina Female     6 17 12 32
4           4  Gina Female     6 17 12 32          3 Alice Female     7 18 14 64
```

3.2 Numerical estimates of central tendency

Given a set of numbers, what is the single best number to represent the entire set? Unfortunately, although easy to state the question, it is impossible to answer, for the best way depends upon what is wanted. An unfortunately common answer, the mode, is perhaps the worst way of estimating the central tendency

3.2.1 Mode: the most frequent

The mode or modal value represents the most frequently observed data point. But the mode is particularly sensitive to the way the data are grouped or to the addition of a single new data point. Consider 100 numbers randomly sampled, with replacement from a distribution ranging from 0-100. Viewed as real numbers to 10 decimal places, there are no repeats and thus all are equally likely. If we convert them to integers by rounding, table the results, and then sort that table, we find that the most frequent rounded observation was 48 which occurred 5 times. But a "stem and leaf" diagram [?] grouping the data by the first decimal digits, shows that there were just as many numbers in the 70s (14) as in the 40s. Breaking the data into 20 chunks instead of 10, leads to the most numbers being observed between 75 and 80. So, what is the mode?

```
> set.seed(1)     #to allow for the same solution each time
> x <- runif(100,0,100)
>x <- round(x)
> sort(table(x))
> stem(x)
> stem(x,scale=2)
```

```
x
 1  2  7  8 10 11 12 13 14 18 19 25 26 29 32 35 37 40 43 44 49 50 52 53 55 57 63 64 67 73 76 80 81 82 83
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
84 87 89 90 93 94 96 99  6 20 21 24 27 34 39 46 60 66 69 71 72 77 79 86 88 91 33 38 41 65 78 48
 1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  4  5
> stem(x)

  The decimal point is 1 digit(s) to the right of the |

  0 | 126678
  1 | 0123489
  2 | 00114456779
  3 | 2333445788899
  4 | 01113466888889
  5 | 02357
  6 | 003455566799
  7 | 11223677888899
  8 | 01234667889
  9 | 0113469

> stem(x,scale=2)

  The decimal point is 1 digit(s) to the right of the |

  0 | 12
  0 | 6678
  1 | 01234
  1 | 89
  2 | 001144
  2 | 56779
  3 | 233344
  3 | 5788899
  4 | 011134
  4 | 66888889
  5 | 023
  5 | 57
  6 | 0034
  6 | 55566799
  7 | 11223
  7 | 677888899
  8 | 01234
  8 | 667889
  9 | 01134
  9 | 69
```

3.2.2 Median: the middle observation

A very robust statistic of the central tendency is the median or middle number of the ranked numbers. For an odd numbered set, the median is that number with as many numbers above it as below it. For an even number of observations, the mean is the half way between the two middle values.

Tukey's 5 number summaries take advantage of the median, and in addition, define the lower and upper quartiles as the median distance from the median. The median subject will not change if the data are transformed with any monotonic transformation, nor will the median value change if the data are "trimmed" of extreme scores.

The median is perhaps the best single description of a set of numbers, for it that characterization that is exactly above 1/2 and exactly below 1/2 of the distribution. Graphically, it is displayed as a heavy bar on a box plot (Figure 3.2).

3.2.3 3 forms of the mean

Intriguingly, there are at least three different forms of what is known as the "mean" that are seen in psychometrics and statistics. One, the *arithmetic average* is what is most commonly thought of as the mean.

$$\bar{X} = X_{.} = (\sum_{i=1}^{N} X_i)/N \tag{3.1}$$

Applied to the data set in Table 3.1, the arithmetic means for the last four variables are (rounded to two decimals):

```
>round(mean(s.df[,4:7]),2)
```

```
theta     X     Y     Z
 4.00 15.00  8.00 18.14
```

Another, the *geometric mean* is nth root of the n products of X_i:

$$X_{geometric} = \sqrt[N]{\prod_{i=1}^{N} X_i} \tag{3.2}$$

Sometimes, the short function we are looking for is not available in R, but can be created rather easily. Creating a new function (geometric.mean) and applying it to the data is such a case:

```
> geometric.mean <- function(x, ...) { return( exp(mean(log(x))) )}
round(geometric.mean(s.df[4:7]),2)
```

```
theta     X     Y     Z
 3.38 14.87  6.76  8.00
```

The third, the *harmonic mean*, is the reciprocal of the arithmetic average of the reciprocals:

$$X_{harmonic} = \frac{N}{\sum_{i=1}^{N} 1/X_i} \tag{3.3}$$

```
> harmonic.mean <- function(x, ...) {return( 1/(mean(1/x)) )}
> round(harmonic.mean(s.df[4:7]),2)
```

```
  harmonic.mean(s.df[,4:7])
 theta     X     Y     Z
 2.70 14.73  5.40  3.53
```

The latter two means can be thought of as the anti-transformed arithmetic means of transformed numbers. That is, just as the harmonic is the reciprocal of the average reciprocal, so is the geometric mean the exponential of the arithmetic average of the logs of X_i:

$$X_{geometric} = exp\{(\sum_{i=1}^{N} log(X_i))/N)\}. \tag{3.4}$$

The harmonic mean is used in the unweighted means analysis of variance when trying to find an average sample size. Suppose 80 subjects are allocated to four conditions but for some reason are allocated unequally to produce samples of size 10, 20, 20, and 30. The harmonic cell size $= \frac{4}{1/10+1/20+1/20+1/30} = 4/.2333 = 17.14$ rather than the 20/cell if they were distributed equally.

The geometric mean is used when averaging slopes and is particularly meaningful when looking at anything that shows geometric or exponential growth. It is equivalent to finding the arithmetic mean of the log transformed data expressed in the original (un-logged) units. Unfortunately, if any of the values are 0, the geometric mean is 0, and the harmonic mean is undefined.

3.2.4 Comparing variables or groups by their central tendency

Returning to the data in Table 3.1, the four estimates of central tendency give strikingly different estimates of which variable is “on the average greater” (Table 3.3). X has the greatest median, geometric and harmonic mean, while Z has the greatest arithmetic mean. Z is a particularly troublesome variable, with the greatest arithmetic mean and the next to smallest harmonic mean.

Table 3.3. Four estimates of central tendency applied to the data of Table 3.1. The four variables differ in their rank orders of size depending upon the way of estimating the central tendency.

```
           theta     X    Y     Z
Median      4.00 15.00 8.00  8.00
Arithmetic  4.00 15.00 8.00 18.14
Geometric   3.38 14.87 6.76  8.00
Harmonic    2.70 14.73 5.40  3.53
```

3.3 The effect of non-linearity on estimates of central tendency

Inferences from observations are typically based on central tendencies of observations. But the inferences can be affected by not just the underlying differences causing these observations, but the way these observations are taken. Consider the example of psychophysiological measures of arousal. Physiological arousal is thought to reflect levels of excitement, alertness and energy. It may be indexed through measures of the head, the heart, and the hand. Among the many ways to measure arousal are two psychophysiological indicators of the degree of palmer sweating. Skin conductance (SC) taken at the palm or the fingers is a direct measure of the activity of the sweat glands of the hands. It is measured by passing a small current through two electrodes, one attached to one finger, another attached to another finger. The higher the skin conductance, the more aroused a subject is said to be. It is measured in units of conductance, or mhos. Skin resistance (SR) is also measured by two electrodes, and reflects the resistance of the skin to passing an electric current. It is measured in units of resistance, the ohm. The less the resistance, the greater the arousal. These two measures, conductance and resistance are reciprocal functions of each other.

Consider two experimenters, A and B. They both are interested in the effect of an exciting movie upon the arousal of their subjects. Experimenter A uses Skin Conductance, experimenter B measures Skin Resistance. They first take their measures, and then, after the movie, take their measures again. The data are shown in Table 3.4. Remember that higher arousal should be associated with greater skin conductance and lower skin resistance. The means for the post test indicate a greater conductance and resistance, implying both an increase (as indexed by skin conductance) and a decrease (as measured by skin resistance)!

How can this be? Graphing the results shows the effect of a non-linear transformation of the data on the mean (Figure 3.3). The group with the smaller variability (the control group) has a mean below the straight line connecting the points with the greater variability (the movie group). The mean conductance and mean resistance for the movie condition is on this straight line.

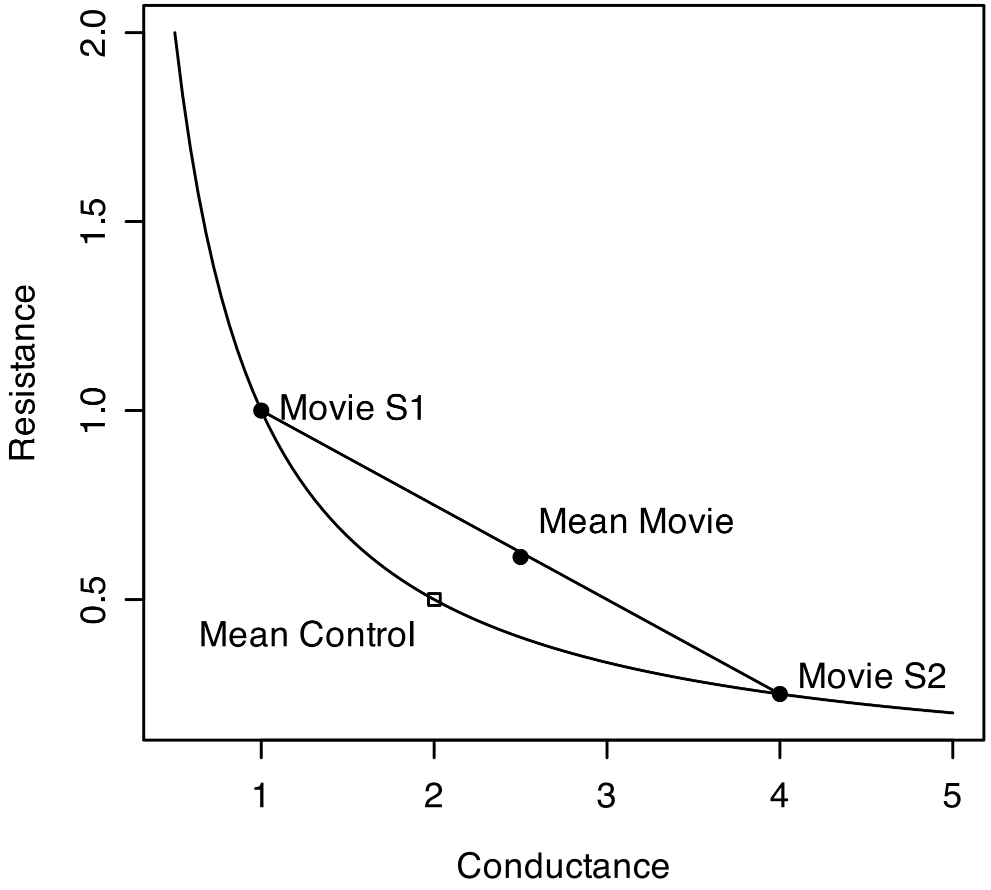

Fig. 3.3. The effect of non-linearity and variability on estimates of central tendency. The movie condition increases the variability of the arousal measures. The "real effect" of the movie is to increase variability which is mistakenly interpreted as an increase/decrease in arousal.

Table 3.4. Hypothetical study of arousal using an exciting movie. The post test shows greater arousal if measured using skin conductance, but less arousal if measured using skin resistance.

Condition	Subject	Skin Conductance	Skin Resistance
Pretest	1	2	.50
	2	2	.50
Average		2	.50
Posttest	1	1	1.00
	2	4	.25
Average		2.5	.61

3.4 Whose mean? The problem of point of view

Even if the arithmetic average is used, finding the central tendency is not as easy as just adding up the observations and dividing by the total number of observations (Equation 3.1). For it is important to think about what is being averaged. Incorrectly finding an average can lead to very serious inferential mistakes. Consider two examples, the first is how long people are in psychotherapy, the second is what is the average class size in particular department.

3.4.1 Average length of time in psychotherapy

A psychotherapist is asked what is the average length of time that a patient is in therapy. This seems to be an easy question, for of the 20 patients, 19 have been in therapy for between 6 and 18 months (with a median of 12) and one has just started. Thus, the median client is in therapy for 52 weeks with an average (in weeks) 1 * 1 + 19 * 52 or 49.4.

However, a more careful analysis examines the case load over a year and discovers that indeed, 19 patients have a median time in treatment of 52 weeks, but that each week the therapist is also seeing a new client for just one session. That is, over the year, the therapist sees 52 patients for 1 week and 19 for a median of 52 weeks. Thus, the median client is in therapy for 1 week and the average client is in therapy of (52 * 1 + 19 * 52)/(52+19) = 14.6 weeks.

A similar problem of taking cross sectional statistics to estimate long term duration have been shown in measuring the average length of time people are on welfare (a social worker's case load at any one time reflects mainly long term clients, but most clients are on welfare for only a short period of time.

3.4.2 Average class size

Consider the problem of a department chairman who wants to recruit faculty by emphasizing the smallness of class size but also report to a dean how effective the department is at meeting its teaching requirements. Suppose there are 20 classes taught by a total of five different faculty members. 12 of the classes are of size 10, 4 of size 20, 2 of 100, one of 200, and one of 400. The median class size from the faculty member point of view is 10, but the mean class size to report to the dean is 50!

But what seems like a great experience for students, with a median class size of 10, is actually much larger from the students' point of view, for 400 of the 1,000 students are in a class of 400, 200 are in a class of 200, 200 are in classes of 100, and only 80 are in classes of 20, and 120 are in class sizes of 10. That is, the median class size from the students' perspective is 200, with an average class size of (10*120+ 20*80 + 200*100 + 200*200 + 400* 400)/1000 = 222.8.

Table 3.5. Average class size depends upon point of view. For the faculty members, the median of 10 is very appealing.

Faculty	Freshman/	Junior	Senior	Graduate	Mean	Median
Member	Sophmore					
A	20	10	10	10	12.5	10
B	20	10	10	10	12.5	10
C	20	10	10	10	12.5	10
D	20	100	10	10	35.0	15
E	200	100	400	10	177.5	150
Total						
Mean	56	46	110	10	50.0	39
Median	20	10	10	10	12.5	10

Table 3.6. Class size from the students' point of view. Most students are in large classes; the median class size is 200 with a mean of 223 .

Class size	Number of classes	number of students
10	12	120
20	4	80
100	2	200
200	1	200
400	1	400

3.5 Non-linearity and interpretation of experimental effects

Many experiments examining the effects of various manipulations or interventions on subjects differing in some way are attempts at showing that manipulation X interacts with personality dimension Y such that X has a bigger effect upon people with one value of Y than another [Revelle, 2007], [Revelle and Oehleberg, 2007]. Unfortunately, without random assignment of subjects to conditions, preexisting differences between the subjects in combination with non-linearity of the observed score-latent score relationship can lead to interactions at the observed score level that do not reflect interactions at the latent score level.

In a brave attempt to measure the effect of a liberal arts education, Winter and McClelland developed a new measure said to assess the " the ability to form and articulate complex concepts and then the use of these concepts in drawing contrasts among examples and instances in the real world" (p 9). Their measure was to have students analyze the differences between two thematic apperception protocols. Winter and McClelland compared freshman and senior students at a "high-quality, high prestige 4 year liberal arts college located in New England" (referred to as "Ivy College") with those of "Teachers College", which was a "4-year state supported institution, relatively nonselective, and enrolling mostly lower-middle-class commuter students who are preparing for specific vocations such as teaching". They also included students from a "Community College" sample with students similar to those of "Teachers Colllege". Taking raw difference scores from freshman year to senior year, they found much greater improvement for the students at "Ivy College" and concluded that "The liberal education of Ivy College improved the ability to form and articulate concepts, sharpened the accuracy of concepts, and tended to fuse these two component skills together" (p 15). That is, that the students learned much more at the more prestigious (and expensive) school [**?**]. While the conclusions of this study are perhaps dear to all faculty members at such prestigious institutions, they suffer from a serious problem.

Rather than reproducing the data from [**?**] consider the left panel of Figure 3.4. The students at "Ivy College" improved more than did their colleagues at "Teachers College" or the "Junior College. When shown these data, most faculty members explain them by pointing out that well paid faculty at prestigious institutions are better teachers.

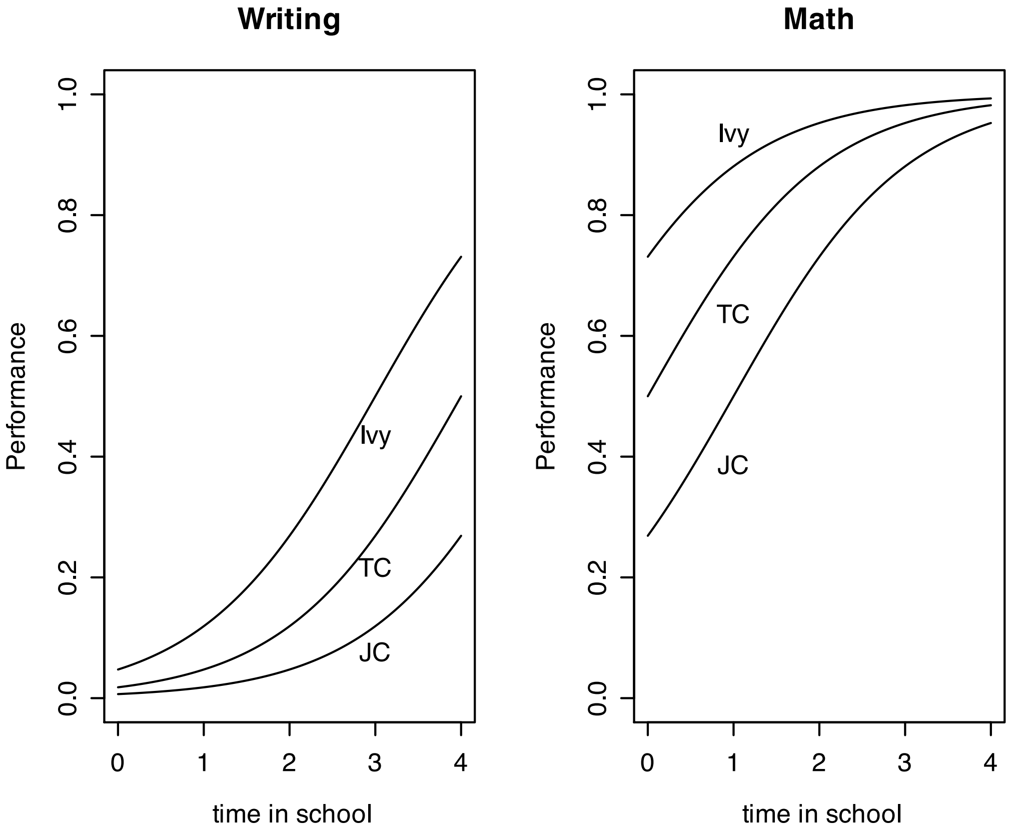

Fig. 3.4. The effect of four years of schooling upon writing and mathematics performance. More selective colleges produce greater change in writing performance than do teacher colleges or junior colleges, but have a smaller effect on improvement in math performance.

Most students explain these results as differences in ability (the “rich get richer” hypothesis) or bright students are more able to learn complex material than are less able students.

However,when given a hypothetical conceptual replication of the study, but involving mathematics performance, yielding the results shown in the right hand panel of Figure 3.4, both students and faculty members immediately point out that there is a ceiling effect on the math performance. That is, the bright students could not show as much change as the less able students because their scores were too close to the maximum.

What is interesting for psychometricians, of course, is that both panels are generated from the exact same monotonic curve, but with items of different difficulties. Using equation 2.10 which is reproduced here:

$$prob(correct|\theta,\delta) = \frac{1}{1+e^{\delta-\theta}} \tag{3.5}$$

and letting the ability parameter, θ, take on different values for the three colleges, (JC=-1, TC=0, IC=1), letting ability increase 1 unit for every year of schooling, and setting the difficulty for the writing at 4 and for the math at 0, is able to produce both the left panel (a hard task) or the right panel (an easy task). The appearance of an interaction in both panels reflects not an interaction of change in ability as a function of college, for at the latent,

θ, level, one year of schooling had an equal effect upon ability (an increase of 1 point) for students at all three colleges and for either the writing or the math test.

This example is important to consider for it reflects an interpretive bias that is all to easy to have: if the data fit one's hypothesis (e.g., that smart students learn more), interpret that result as confirming the hypothesis, but if the results go against the hypothesis (smart students learn less), interpret the results as an artifact of scaling (in this case, a ceiling effect). When seeing fan-fold interactions such as in Figure 3.4, do not interpret them as showing an interaction at the latent level unless further evidence allows one to reject the hypothesis of non-linearity.

Other examples of supposed interactions that could easily be scaling artifacts include stage models of development (children at a particular stage learn much more than children below or above that stage; the effect of hippocampal damage on short term versus long term memory performance, and the interactive effect on vigilance performance of time on task with the personality dimension of impulsivity. In general, without demonstrating a linear correspondence between the latent and observed score, main effects (Figure 3.3) and interactions (Figure 3.4) are open to measurement artifact interpretations.

3.6 Measures of dispersion

In addition to describing a data set with a measure of central tendency, it is important to have some idea of the amount of dispersion around that central value.

3.6.1 Measures of range

An easy measure of dispersion is the range from the highest to the lowest. Unfortunately, range partly reflects the size of a sample, for as the sample size increases, the range will increase as well. This is shown in the left panel of Figure 3.5 for samples of size 2 to 10^6. The range (the difference between the maximum and minimum values) increases dramatically with sample size. One important use of the range is detect data entry errors. For if the largest possible value should be 9 and an occasional 99 is discovered, it is likely that a mistake has occurred.

A more useful measure of range is the interquartile range, that is the range from the 25th percentile to the 75th percentile. As seen in the right panel of Figure 3.5, the interquartile range barely varies with sample above about 32. The *IQR* function can be used to find the interquartile range. For normal data, the IQR should be the twice the normal score of the 75th percentile = 2 * qnorm(.75) = 1.348980.

3.6.2 Average distance from the central tendency

Given some estimate of the "average' observation (where the average could be the median, the arithmetic mean, the geometric mean, or the harmonic mean), how far away is the average participant? Once again, there are multiple ways of answering this question.

Median absolute distance from the median

When using medians as estimates of central tendencies, it is common to also consider the median absolute distance from the median. The *mad* function returns the appropriate value. For consistency with normal data, by default the mad function is adjusted for the fact that it is systematically smaller than the standard deviation by a factor of 1/qnorm(.75). Thus, the default is to return the median absolute deviation * 1.4826. With this adjustment, if the data are normal, then the mad and sd function will return almost identical values.

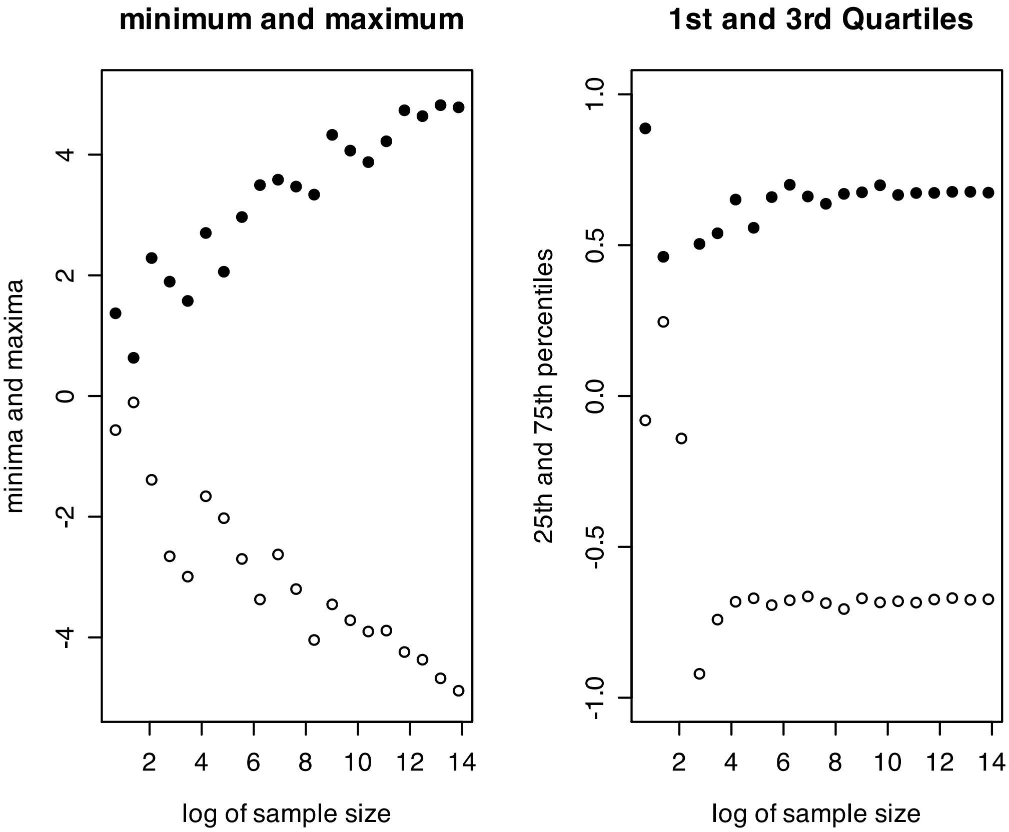

Fig. 3.5. Left hand panel: The minimum and maximum of a sample will generally get further apart as the sample size increases. Right hand panel: The distance between the 25th and 75th percentile (twice the interquartile range) barely changes as sample size increases. Data are taken from random normal distributions of sample sizes of 2 to 2^{20}. Sample size is log transformed.

Sums of squares and Euclidean distance

A vector X with n elements can be thought of as a line in n dimensional space. Generalizing Pythagorus to n dimensions, the length of that line in Euclidean space will be the square root of the sum of the squared distances along each of the n dimensions (remember that Pythagorus showed that $c^2 = a^2 + b^2$).

To find the sums of squares of a vector is to multiply the transpose of the vector (X^T) times the vector (X):

$$SS = SumSquares = \sum_{i=1}^{n}(X_i^2) = X^T X \tag{3.6}$$

If X is a matrix, then the Sum Squares will be the diagonal of the $X^T X$ matrix product. Letting X be the matrix formed from the last 4 variables from Table 3.1:

```
X <- as.matrix(s.df[,4:7])
X
```

```
SS <- diag(t(X)%*% X)
SS
```

```
> X
  theta  X  Y  Z
1     1 12  2  1
7     2 13  4  2
2     3 14  6  4
5     4 15  8  8
6     5 16 10 16
4     6 17 12 32
3     7 18 14 64
> SS
theta     X     Y     Z
  140  1603   560  5461
```

3.6.3 Deviation scores and the standard deviation

Rather than considering the raw data (X), it is more common to transform the data by subtracting the mean from all data points.

$$deviationscore_i = x_i = X_i - X. = X_i - \sum_{i=1}^{n}(X_i)/n \tag{3.7}$$

Finding the Sums of Squares or length of this vector is done by using equation 3.6, and for a data matrix, the SS of deviation scores will be $X^T X$. If the SS is scaled by the number of observations (n) or by the number of observations -1 (n-1), it becomes a Mean Square, or variance. Taking the square root of the Variance converts the numbers in the original units and is a measure of the length of the vector of deviations in n-dimensional space.

```
X <- as.matrix(s.df[,4:7])
c.means <- colMeans(X)
X.mean <- matrix(rep(c.means,7),byrow=TRUE,nrow=7)
x  <- X - X.mean
SS <- diag(t(x)%*% x)
x.var <- SS/(dim(x)[1]-1)
x.sd <- sqrt(x.var)
SS
x.var
x.sd
```

```
 SS
   theta        X        Y        Z
  28.000   28.000  112.000 3156.857
> x.var
     theta          X          Y          Z
  4.666667   4.666667  18.666667 526.142857
> x.sd
    theta         X         Y         Z
 2.160247  2.160247  4.320494 22.937804
```

As would be expected, because the operation of finding the sums of squares of a deviations from the mean is so common, rather than doing the matrix operations shown above, functions for the standard deviation and the variance are basic functions in R.

Deviation scores are in the same units as the original variables, but sum to zero.

3.6.4 Standard scores as unit free measures

In some fields, the unit of measurement is most important. In economics, a basic unit could be the dollar or the logarithm of the dollar. In education the basic unit might be years of schooling. In cognitive psychology the unit might be the millesecond. A tradition in much of individual differences psychology is ignore the units of measurement and to convert deviation scores into standard scores. That is, to divide deviation scores by the standard deviation:

$$z_i = x_i/\sigma_x = (X - X.)/sqrt(Var_X) \tag{3.8}$$

One particularly attractive feature of standard scores is that they have mean of 0 and standard deviation and variance of 1. This makes some derivations easier to do because variances or standard deviations drop out of the equations. A disadvantage of standard scores is communicating the scores to lay people. To be told that someone's son or daughter has a score of -1 is particularly discouraging. To avoid this problem (and to avoid the problem of decimals and negative numbers in general, a number of transformations of standard scores are used when communicating to the public. They are all of the form of multiplying by a constant and then adding a different constant (Table 3.7).

Table 3.7. Raw scores are typically converted into deviation scores or standard scores. These are, in turn, the transformed into "public" scores for communication to laypeople.

	Mean	Standard Deviation
Raw Data	$X. = \sum(X_i)/n$	$\sqrt{\sum(X_i - X.)^2/(n-1)}$ $s_x = \sqrt{\sum(x_i)^2/(n-1)}$
deviation score	0	s_x
standard score	0	1
"IQ"	100	15
"SAT"	500	100
"ACT"	18	6
"T-score"	50	10
"Stanine"	5	1.5