Chapter 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: Tukey (1977))

As discussed in Chapter 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 7), the *validity* (Chapter 9) 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 (or any other R object, use the `str` function.

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

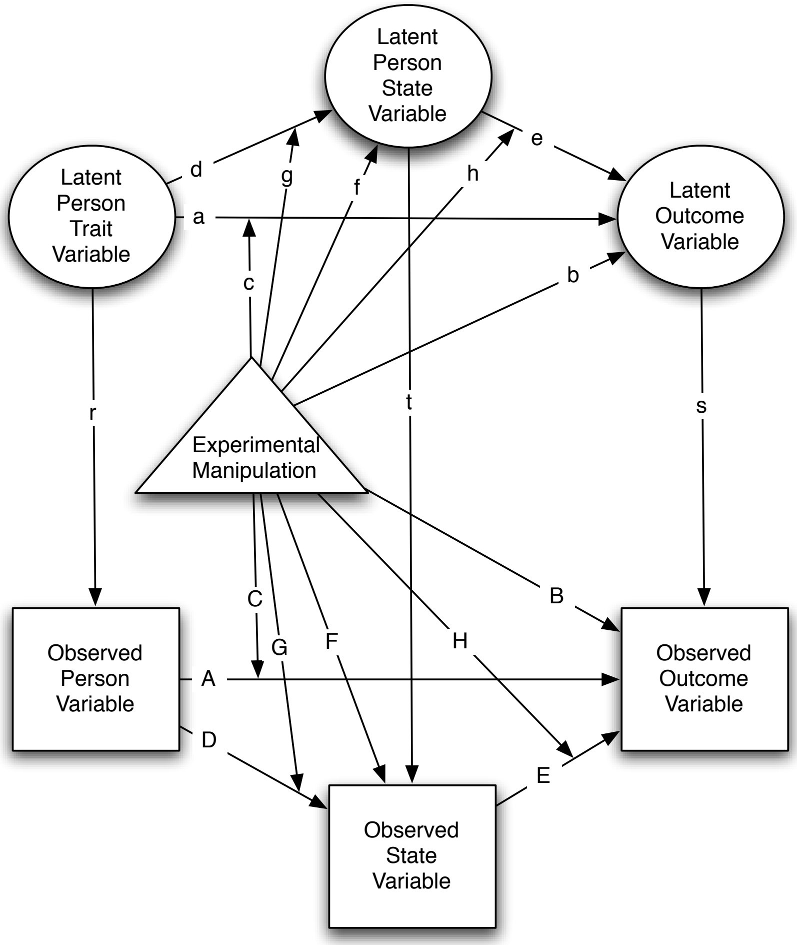
Latent Person State Variable
Latent Person Trait Variable
Latent Outcome Variable
d
g
f
h
e
a
c
b
r
t
s
Experimental Manipulation
C
G
F
H
B
Observed Person Variable
A
Observed Outcome Variable
D
E
Observed State Variable

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))

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)
'data.frame':        7 obs. of  7 variables:
```

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

```
$ 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
```

3.1 Four broad classes of scales

The association of numbers with data would seem to be easy but in fact is one of the most intractable problems in psychology. It would seem that associating a number to a data point is straight forward and it is ("Alice answered 18 questions correctly, Bob answered 12, Eric 15") but the inferences associated with these numbers differ depending what these numbers represent. In the mid-20th century, the assignment of numbers and use of the the term *measurement* applied to psychological phenomena led to a acrimonious debate between physicists and psychologists (Ferguson et al., 1940) that was left unresolved. To the physicists, measurement is "the assignment of numerals to things so as to represent facts of conventions about them" (p 340). Although this meaning is clearly what we think of when measuring mass or distance, it implicitly requires the ability to form ratios. (Something is twice as heavy as something else, something is three times further away). But the assignment of numbers to observations in psychology usually does not meet this requirement. In response to the Ferguson et al. (1940) report, Stevens (1946) proposed what has become the conventional way of treating numbers in psychology. That is, numbers can be seen as representing *nominal*, *ordinal*, *interval* or *ratio* levels of measurement (Table 3.2). Stevens was responding to the criticism that psychological scales were meaningless because they were not true measurement.

This controversy over what is a measurement continues to this day with some referring to the "pathological nature" of psychometrics (Michell, 2000) for ignoring the fundamental work in *measurement theory* (Falmagne, 1992; Krantz and Suppes, 1971) associated with *conjoint measurement* as advanced by Krantz and Tversky (1971) and others. Falmagne's (1992) review is a very nice introduction to the power of measurement theory. Other useful reviews include the history of measurement Díez (1997) which discusses the important work of Hölder (1901) in a translation by (Michell and Ernst, 1997).

Although foolhardy to summarize volumes of work in a paragraph, a core idea in measurement theory is that a variable may be said to be measured on an interval scale, u, if particular patterns of comparisons of order are maintained. Consider the dimensions, X and the operation $\geq$ on pairs of elements of X. Then $(a,b) \geq (c,d) \iff u(a) - u(b) \geq u(c) - u(d)$ which implies $(a,b) \geq (c,d) \iff (a,c) \geq (b,d)$ (Falmagne, 1992). Extending this idea to the way that variables may be combined and still be said to be measured on the same metric, is the basis of *conjoint measurement theory* (Krantz and Tversky, 1971). A fundamental conclusion is that u is an interval scale of a, b, p, and q, if $(a,p) \geq (b,p) \iff (a,q) \geq (b,q)$ for all p and q. (As is discussed in Chapter 8 it is a violation of this relationship that leads proponents of the *1PL Rasch model* to reject models with more parameters such as the *2PL* or *1PN*).

In psychometrics, some attention has been paid to measurement theory, and indeed the advantages of *item response theory* (Chapter 8) compared to *classical test theory* (Chapter 7) have been framed in terms of the measurement properties of scales developed with the two models (but see Cliff (1992) for a concern that not enough attention has been paid). Even within the *IRT* approach, the differences between one parameter *Rasch models* (8.1.1) and more complicated models are debated in terms of basic measurement properties (Cliff, 1992).

Proponents of measurement theory seem to suggest that unless psychologists use interval or ratio measures, they are not doing "real" science. But this seems to ignore examples of how careful observation, combined with theoretical sophistication but with measures no more complicated than counts and ordinal relationships has led to theories as diverse as evolution or plate tectonics.

Table 3.2 Four types of scales and their associated statistics (Rossi, 2007; Stevens, 1946) The statistics listed for a scale are invariant for that type of transformation. The Beaufort wind speed scale is interval with respect to the velocity of the wind, but only ordinal with respect to the effect of the wind. The Richter scale of earthquake intensity is a logarithmic scale of the energy released but linear measure of the deflection on a seismometer. *Note that Stevens lists rank correlations as requiring interval properties although they are insensitive to monotonic transformations.

Scale	Basic operations	Transformations	Invariant statistic	Examples
Nominal	equality $x_i = x_j$	Permutations	Counts Mode χ^2 and (ϕ) correlation	Detection Species classification Taxons
Ordinal	order $x_i > x_j$	Monotonic (homeomorphic) x' =f(x) f is monotonic	Median Percentiles Spearman correlations*	Mhos Hardness scale Beaufort Wind (intensity) Richter earthquake scale
Interval	differences $(x_i - x_j) > (x_k - x_l)$	Linear (Affine) x' = a + bx	Mean (μ) Standard Deviation (σ) Pearson correlation (r) Regression (β)	Temperature (°F, °C) Beaufort Wind (velocity)
Ratio	ratios $\frac{x_i}{x_j} > \frac{x_k}{x_l}$	Multiplication (Similiarity) x' = bx	Coefficient of variation $(\frac{\sigma}{\mu})$	Length, mass, time Temperature (°K) Heating degree days

3.1.1 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.

3.1.2 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 use more meaningful numbers, can we treat the Mhos scale of hardness values the same way we treat the Indentation hardness values (refer back to Table 2.6) or the Beaufort scale ratings with wind velocity (Table 2.7)? To answer these questions, it is useful to first consider how to summarize a set of numbers in terms of dispersion and central tendency.

3.2 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 Tukey (1977). 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.

A graphic representation of the Tukey 5 points is the “BoxPlot” drawn by the `boxplot` function (Figure 3.2) which includes two more numbers, the upper and lower “whiskers”, which are defined as the most extreme numbers that do not exceed 1.5 the *InterQuartileRange* (*IQR*) beyond the upper and lower quartiles. Why, you might ask, 1.5? The IQR is the distance from the 25th to 75th percentile. If the data are sampled from a normal distribution, the IQR corresponds to 1.35 z units. And 1.5 times that is 2.02. That is, the whiskers will be 2.7 z score units above and below the mean and median. For a normal, this corresponds to

Table 3.3 Basic summary statistics from the `summary` function include Tukey's "five numbers".

```
> s.df <- read.clipboard()
> summary(s.df)
> colnames(s.df)[4] <- "theta"
>boxplot(s.df[,4:7],main="Boxplot of data from Table 3.1")
```

```
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
              Gina :1
```

roughly the .005 region at each tail, and thus any point beyond the whiskers in either direction has a .01 chance of occurring. (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 (use the "notch" option in `boxplot`), and "violin" plots which give more graphical representations of the distributions within the distributions (see `vioplot` in the *vioplot* package).

3.2.1 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, $\boldsymbol{\theta}$, 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. The concept that "*Alabama need not come first*" (Ehrenberg, 1977; Wainer, 1978, 1983; Wainer and Thissen, 1981) is a basic rule in table construction and implies that sorting the data by meaningful variables rather than mere alphabetical or item order will frequently produce useful findings. Specifying that the new values of the `data.frame` are to be ordered by the the rank ordered values of the `order` function sorts the data frame.

3.3 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. However, it is possible to say that an unfortunately common answer, the mode, is perhaps the worst way of estimating the central tendency.

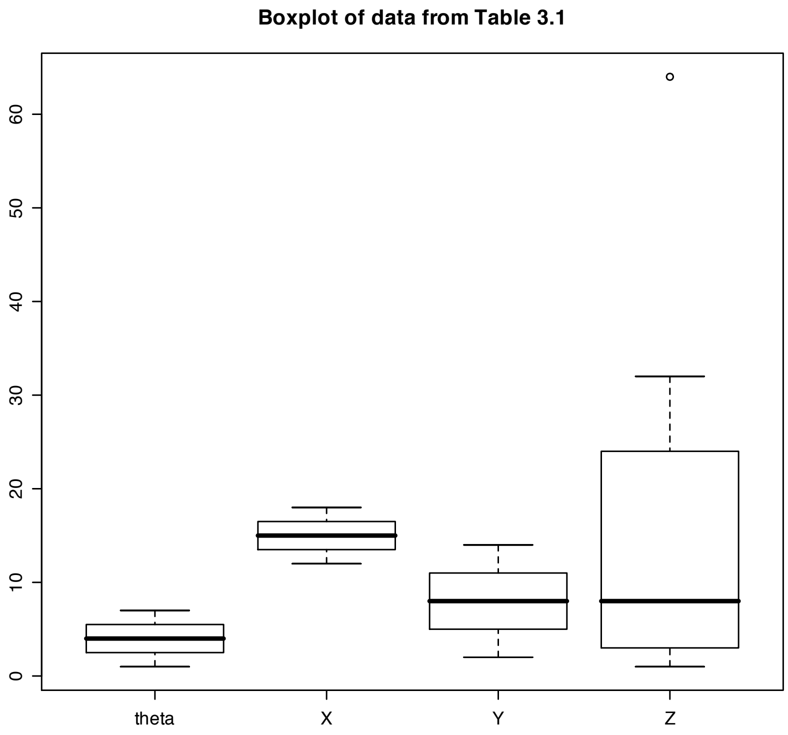

Fig. 3.2 The Tukey box and whiskers plot shows the minima, maxima, 25th and 75th percentiles, as well as the “whiskers” (either the lowest or highest observation or the most extreme value which is no more than 1.5 times the interquartile range from the box.) Note the outlier on the Z variable.

3.3.1 Mode: the most frequent

The mode or modal value represents the most frequently observed data point. This is perhaps useful for categorical data, but not as useful with ordinal or interval data, for 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 pseudo randomly generated from 1 to 100 from a uniform distribution using the `runif` function. (Alternatively,the `sample` could have been used to sample with replacement from a distribution ranging from 1-100). 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 (`round(x)`), `table` the results, and then `sort` that table, we find that the most frequent rounded observation was 39 or 48 which occurred 4 times. (The example code combines these three commands into one line.) This mode is different when we use the `stem` to produce a *stem and leaf* diagram Tukey (1977) which groups the data by the first decimal digits. The stem and leaf shows that there were just as many numbers in the 70s (14) as in the 30s. Breaking the data into 5 chunks instead of 10, leads to the most numbers being observed between 60 and 80. So, what is the mode?

Table 3.4 Sometimes, sorting the data shows relationships that are not obvious from the unsorted data. Two different sorts are shown, the first, sorting alphabetically by name is less useful than the second, sorting by variable 4. Note that the sort can either be based upon column number or by column name. Compare this organization to that of Table 3.1.

```
> n.df <- s.df[order(s.df[,2]),]     #create a new data frame, ordered by the 2nd variable of s.df
> s.df <- s.df[order(s.df$theta),]  #order the data frame by the fourth variable (theta)
> 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
```

```
> set.seed(1)      #to allow for the same solution each time
>  x <- runif(100,1,100)   #create 100 pseudo random numbers from 1 to 100.
> # x <- sample(100,100,replace=TRUE)
     #           Alternatively, take 100 samples from the integers 1 to 100
> sort(table(round(x)))
> stem(x)
> stem(x,scale=.5)
```

```
 2  3  8  9 11 12 15 18 19 22 30 32 33 38 40 49 52 53 56 58 60 61
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
63 69 70 71 76 81 82 83 84 86 89 90 94 95 96 99  7 13 34 41 42 44
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2
46 50 66 67 72 73 77 79 80 87 88 91 21 25 27 35 65 78 39 48
 2  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  4  4
> stem(x)

  0 | 237789
  1 | 1233589
  2 | 1112555777
  3 | 0234455589999
  4 | 01122446688889
  5 | 002368
  6 | 01355566779
  7 | 01223367788899
  8 | 001234677889
  9 | 0114569

> stem(x,.5)

  The decimal point is 1 digit(s) to the right of the |
  0 | 2377891233589
  2 | 111255557770234455589999
  4 | 01122446688889002368
  6 | 0135556677901223367788899
  8 | 00123467788890114569
```

The mode is a useful summary statistic for categorical data but should not be used to summarize characteristics of data that have at least ordinal properties.

3.3.2 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 median is half way between the two middle values. A robust estimate is one that has the property that slight changes in the distribution will lead to small changes in the estimate (Wilcox, 2005). The median is particularly robust in that monotonic changes in the values of all the numbers above it or below do not affect the median.

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 (see `summary`). 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 either by deleting the extreme scores or by converting all scores beyond a certain value to that value "(*winsorizing*" the mean, see `winsor`).

The median is perhaps the best single description of a set of numbers, for it is 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).

Galton (1874) was a great proponent of the median as an estimate of central tendency for the simple reason that it was easy to find when taking measurements in the field:

> Now suppose that I want to get the average height and "probable error" of a crowd [...]. Measuring them individually is out of the question; but it is not difficult to range them –roughly for the most part, but more carefully near the middle and one of the quarter points of the series. Then I pick out two men, and two only–the one as near near the middle as may be, and the other near the quarter point, and I measure them at leisure. The height of the first man is the average of the whole series, and the difference between him and the other man gives the probable error (Galton, 1874, p 343).

In addition to the technique of lining people up by height to quickly find the median height, Galton (1899) proposed a novel way of using normal theory to estimate both the median and the *interquartile range*:

> The problem is representative of a large class of much importance to anthropologists in the field, few of whom appear to be quick at arithmetic or acquainted even with the elements of algebra. They often desire to ascertain the physical characteristics of [people] who are too timourous or suspicious to be measured individually, but who could easily be dealt with by my method. Suppose it to be a question of strength, as measured by lifting power, and that it has been ascertained that a per cent. of them fail to lift a certain bag A of known weight, and that b per cent of them fail to lift an even heavier bag B. From these two data, the median strength can be

> determined by the simple method spoken of above, and not only it but also the distribution of strength among the people.

Unfortunately, when the data are grouped in only a few levels (say 4 or 5 response levels on a teacher rating scale, or by year in school for college students), the median does not give the resolution needed for useful descriptions of the data. It is more useful to consider that each number, x, represents the range from x - .5w to x +.5w, where w is the width of the interval represented by the number. If there are multiple observations with the same nominal value, they can be thought of as being uniformly distributed across that range. Thus, given the the two sets of numbers, x and y, with values ranging from 1 to 5 (Table 3.5) the simple median (the 17th number in these 33 item sets) is 3 in both cases, but the first "3" represents the lower range of 2.5-3.5 and the second "3" represents the highest part of the same range. Using linear interpolation and the `interp.median` function, the *interpolated medians* are 2.54 and 3.46 respectively. By comparing the results of the `summary` and `interp.quartiles` functions, the distinction is even clearer. The `summary` output fails to capture the difference between these two sets of data as well as does the interpolated quartiles results (See Figure 3.3 for another way of looking at the data.).Note the use of the `order` function to rearrange the data and the `print` function to specify the precision of the answer.

Table 3.5 Finding the median and other quantiles by interpolation gives more precision. Compare the 1st, 2nd and 3rd Quartiles from the `summary` function to those found by the `interp.quantiles` function. See Figure 3.3 for another perspective.

```
> x <-  c(1,1,2,2,2,3,3,3,3,4,5,1,1,1,2,2,3,3,3,3,4,5,1,1,1,2,2,3,3,3,3,4,2)
> y <-  c(1,2,3,3,3,3,4,4,4,5,5,1,2,3,3,3,3,4,4,5,5,5,1,5,3,3,3,3,4,4,4,5,5)
> x <-  x[order(x)]   #sort the data by ascending order to make it clearer
> y <- y[order(y)]
> data.df <- data.frame(x,y)
> summary(data.df)
> print(interp.quartiles(x),digits=3)   #use print with digits to make pretty output
>  print(interp.quartiles(y),digits=3)

> x
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 5 5
> y
 [1] 1 1 1 2 2 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5
       x               y        
 Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:3.000  
 Median :3.000   Median :3.000  
 Mean   :2.485   Mean   :3.485  
 3rd Qu.:3.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000  
> print(interp.quartiles(x),digits=3)   #use print with digits to make pretty output
[1] 1.50 2.54 3.23
>  print(interp.quartiles(y),digits=3)
[1] 2.77 3.46 4.50
```

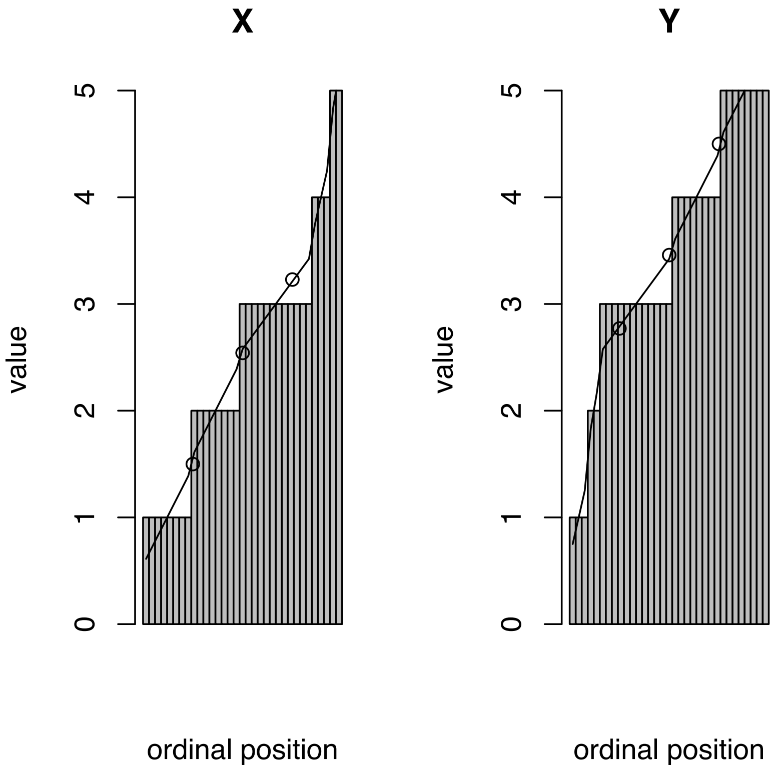

Fig. 3.3 When the data represent just a few response levels (e.g., emotion or personality items, or years of education), raw medians and quartile statistics fail to capture distinctions in the data. Using linear interpolation within each response level (`interp.quartiles`), finer distinctions may be made. Although both the X and Y data sets have equal medians the data are quite different. See Table 3.5 for the data.

3.3.3 3 forms of the mean

Even though most people think they know what is a *mean* there are at least three different forms seen in psychometrics and statistics. One, the *arithmetic average* is what is most commonly thought of as the mean.

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

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
```

Because the mean is very sensitive to outliers, it is sometimes recommended to "trim" the top and bottom n%. Trimming the top and bottom 20% of the data in Table 3.1 leads to very different estimates for one of the variables (Z). Another technique for reducing the effect of outliers is to find the "Winsorized" mean. This involves sorting the data and replacing all values less than the nth value with the nth value, and all values greater than the N-th value with the N-th value (Wilcox, 2005). Several packages have functions to calculate the Winsorized mean, including `winsor` in the **psych** package.

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

```
theta     X     Y     Z
  4.0  15.0   8.0  12.4
  4.00 15.00  8.00 13.71
```

Another way to find a mean is the *geometric mean* which is the nth root of the n products of X_i:

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

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, na.rm=TRUE) { 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 type of mean, the *harmonic mean*, is the reciprocal of the arithmetic average of the reciprocals and we can create the function `harmonic.mean` to calculate it:

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

```
> harmonic.mean <- function(x,na.rm=TRUE) { 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:

$$\bar{X}_{geometric} = e^{(\sum_{i=1}^{N} log(X_i))/N}. \tag{3.4}$$

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 $= \frac{4}{1/10+1/20+1/20+1/30} = \frac{4}{.2333} = 17.14$ rather than the 20/cell if they were distributed equally. Harmonic means are also used when averaging resistances in electric circuits or the amount of insulation in a combination of windows.

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. For distributions that are log normally distributed, the geometric mean is a better indicator of the central tendency of the distribution than is the arithmetic mean. Unfortunately, if any of the values are 0, the geometric mean is 0, and the harmonic mean is undefined.

3.3.4 Comparing variables or groups by their central tendency

Returning to the data in Table 3.1, the five estimates of central tendency give strikingly different estimates of which variable is "on the average greater" (Table 3.6). X has the greatest median, geometric and harmonic mean, while Z has the greatest arithmetic mean, but not the greatest trimmed mean. Z is a particularly troublesome variable, with the greatest arithmetic mean and the next to smallest harmonic mean.

Table 3.6 Six 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
Trimmed     4.00 15.00 8.00 12.40
Winsorized  4.00 15.00 8.00 13.71
Geometric   3.38 14.87 6.76  8.00
Harmonic    2.70 14.73 5.40  3.53
```

3.4 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 3.7. 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 3.4). 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.

Table 3.7 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

3.4.1 Circular Means

An even more drastic transformation of the data that requires yet another way of estimating central tendency is when the units represent angles and thus can be represented as locations on a circle. The appropriate central tendency is not the arithmetic mean but rather the *circular mean* (Jammalamadaka and Lund, 2006). A typical example in psychology is the measurement of mood over the day. *Energetic arousal*, *EA*, as measured by such self report items as being alert, wide awake, and not sleepy or tired Rafaeli and Revelle (2006); Thayer (1989) shows a marked *diurnal* or *circadian rhythm*. That is, EA is low but rising in the morning, peaks sometimes during the early to late afternoon, and then declines at night. Another example of a phasic rhythm that shows marked individual differences is body temperature Baehr et al. (2000). Consider the hypothetical data in table 3.8 showing the time of day that each of four scales achieved its maximum for six subjects. (This could be found by examining the data within each subject for multiple times of day and then finding the maximum for the scale. A technique for finding the phase angle associated with the maximum will

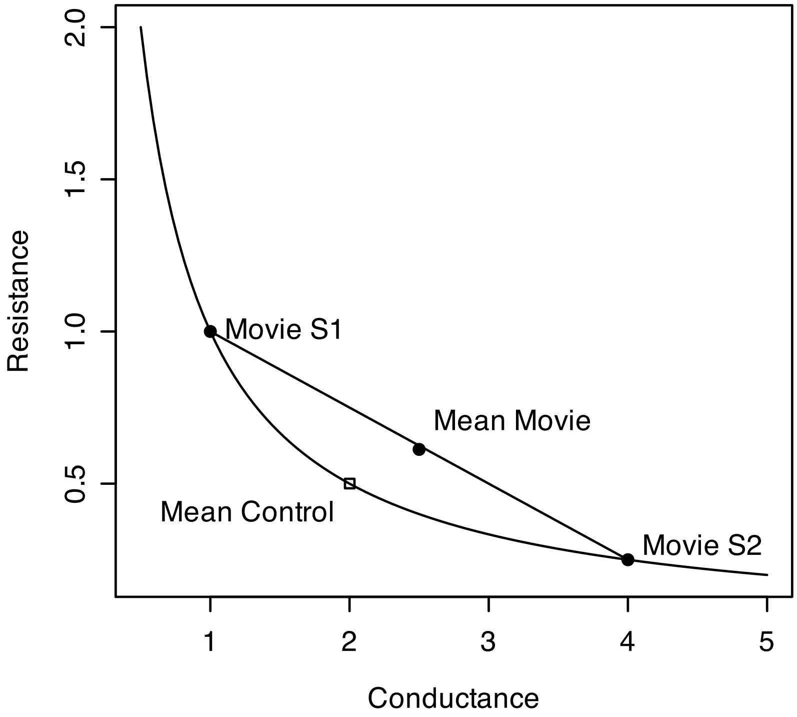

Fig. 3.4 The effect of non-linearity and variability on estimates of central tendency. The movie condition increases the variability of the arousal measures. The "real effect" of the movie is to increase variability which is mistakenly interpreted as an increase/decrease in arousal.

be discussed later (5.4.3). What is the central tendency for each scale? The simple arithmetic mean suggests that Tense Arousal achieves its maximum at 12 noon and Negative Affect has an average maximum at 9 am. But examining the data suggests that midnight and 5 am are better measures of the central tendency. Using `mean.circular` from the **circular** package or `circadian.mean` from the **psych** package converts the angles (expressed in radians for `mean.circular` or hours for `circadian.mean`) to two dimensional vectors (representing the sin and cosine of the angle), finds the averages for each dimension, and then translates the average vector back into angles. Note how for the sample mood data in table 3.8, the circular means correctly capture the change in phase angles between the four moods.

Table 3.8 Hypothetical mood data from six subjects for four mood variables. The values reflect the time of day that each scale achieves its maximum value for each subject. Each mood variable is just the previous one shifted by 5 hours. Note how this structure is preserved for the *circular mean* but not for the arithmetic mean.

Subject	Energetic Arousal	Positive Affect	Tense Arousal	Negative Affect
1	9	14	19	24
2	11	16	21	2
3	13	18	23	4
4	15	20	1	6
5	17	22	3	8
6	19	24	5	10
Arithmetic Mean	14	19	12	9
Circular Mean	14	19	24	5

3.5 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.

3.5.1 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)/20 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). Situations where the participants are *self weighted* lead to this problem. The average velocity of tortoises and hares passing by an observer will be weighted towards the velocity of hares as more of those pass by, even though the overall velocity of both tortoises and hares is much less.

3.5.2 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 3.9 Average class size depends upon point of view. For the faculty members, the median of 10 is very appealing. From the Dean's perspective, the faculty members teach an average of 50 students per calls.

Faculty Member	Freshman/ Sophmore	Junior	Senior	Graduate	Mean	Median
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 3.10 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

3.6 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, 2008). 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 Winter and McClelland (1978). 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 Winter and McClelland (1978) consider the left panel of Figure 3.5. 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. 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 3.5, 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. Consider equation 2.18 which is reproduced here:

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

Let the ability parameter, θ, take on different values for the three colleges, (JC = -1, TC = 0, IC = 1), let ability increase 1 unit for every year of schooling, and set the difficulty for the writing at 4 and for the math at 0. Then equation 3.5 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 is real, but it is at the observed score level, not at the latent level. For the different slopes of the lines reflect not an interaction of change in ability as a function of college, for

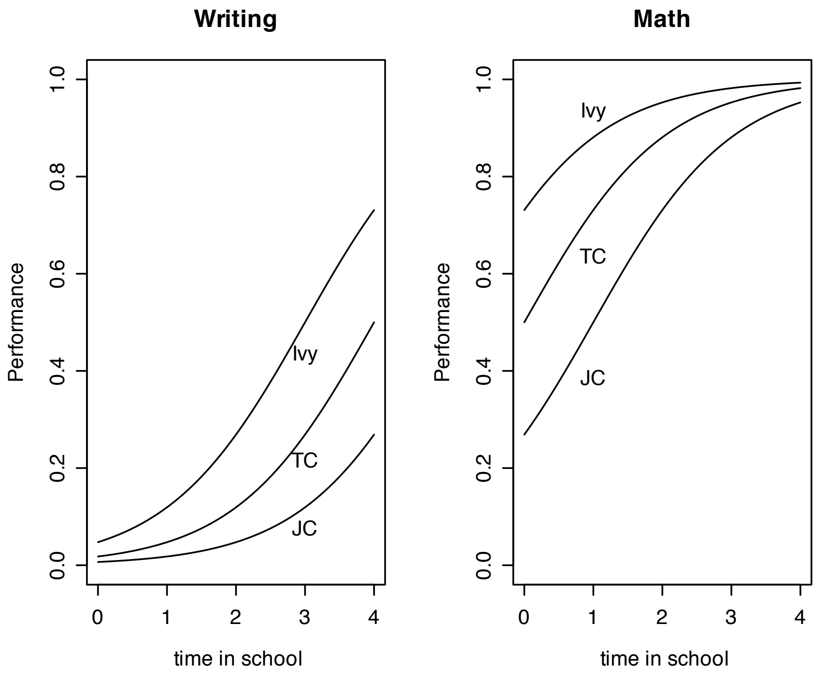

Fig. 3.5 The effect of four years of schooling upon writing and mathematics performance. More selective colleges produce greater change in writing performance than do teacher colleges or junior colleges, but have a smaller effect on improvement in math performance.

at the latent, $\boldsymbol{\theta}$, 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). The moral of this example is that when seeing fan-fold interactions such as in Figure 3.5, 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 3.4) and interactions (Figure 3.5) are open to measurement artifact interpretations (Revelle, 2007).

3.6.1 Linearity, non-linearity and the properties of measurement

It is these problems in interpretation of mean differences that has been the focus of work on the fundamentals of measurement (Krantz and Suppes, 1971) and that was the basis of the controversy in the 1930s (Ferguson et al., 1940). Stevens (1946) proposal to consider four levels of psychological measures suggested that to compare means it was necessary to have at least interval levels of measurement and that without such measurement qualities, we are restricted to comparisons of medians. The problem of interpretation considered in Figure 3.5 does not occur if the discussion is in terms of medians, for in that case, the effect of a year in schooling is a monotonic increase for all three institutions and there is no possibility of saying that one group changed more than another group.

Comparisons using scales developed using the *Rasch model* have been claimed by some (Bond and Fox, 2007; Borsboom, 2005; Borsboom and Scholten, 2008) to offer the interval quality of measurement required for the comparisons of means using the principles of *conjoint measurement* (Krantz and Tversky, 1971) although others strongly disagree (Kyngdon, 2008; Michell, 2000, 2004) and yet others remain strongly undecided (Reise and Waller, 2009). The pragmatic advice is to be very careful about interpreting ordinal interactions or any effect that can go away with a monotonic transformation and to look for disordinal interactions or effects that remain even after extreme but monotonic transformations.

3.7 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.

3.7.1 Measures of range

Perhaps the most obvious 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 probability of observing at least one rare (extreme) event will increase as well (The probabiity of the extreme event has not changed, but given more observations, the probability of observing at least one increases.) This is shown in the left panel of Figure 3.6 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. Thus, finding the `max` and `min` of the data is useful, but normally just as a way of checking errors.

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 3.6, the interquartile range barely varies with sample above about 32. Here the range is expressed in raw score units. 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.

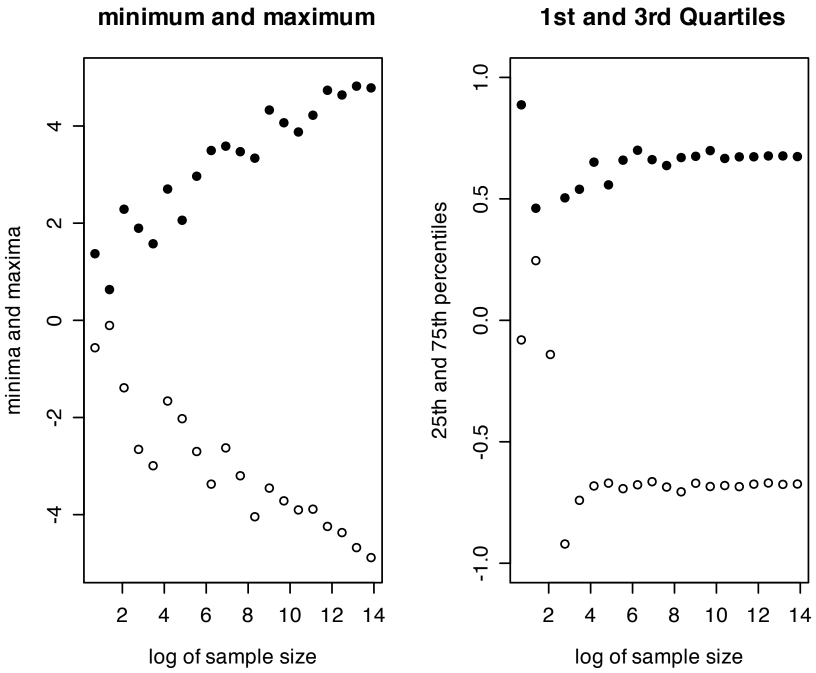

Fig. 3.6 Left hand panel: The minimum and maximum of a sample will generally get further apart as the sample size increases. Right hand panel: The distance between the 25th and 75th percentile (the interquartile range) barely changes as sample size increases. Data are taken from random normal distributions of sample sizes of 2 to 2^{20}. Sample size is log transformed.

In that 50% of the observations will be between the lower and upper quartile, Galton (1888) took 1/2 of the interquartile range as a measure of the *probable error*. That is, for any set of numbers with median, M, the interval M - .5 * IQR to M + .5 IQR will include half of the numbers.

> This unit is known by the uncouth and not easily justified names of 'probable error,' which I suppose is intended to express the fact that the number of deviations or 'Errors' in the two outer fourths of the series is the same as those in the middle two fourths; and therefore the probabilty is equal that an unknown error will fall into either of these two great halves, the outer or the inner. (Galton, 1908, Chapter XX. Heredity)

3.7.2 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.

3.7.2.1 Median absolute deviation from the median

When using medians as estimates of central tendencies, it is common to also consider the median absolute distance from the median `mad`. That is, `median(abs(X-median(X))`. 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 (see below) 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. If, however, the data are not normal, but contain some particularly unusual data points (outliers), the `mad` will be much less than the `sd` (see the discussion of robust estimators of dispersion at 3.14).

3.7.2.2 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 for two dimensions $c^2 = a^2 + b^2$ or $c = \sqrt{(}a^2 + b^2))$.

To find the sums of squares of a vector X we 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])
> SS <- diag(t(X)%*% X)

> 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
```

3.7.3 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*. The variance is the second moment around the mean:

$$\sigma^2 = \sum_{i=1}^{n}(x_i^2)/(n-1) = x^T x/(n-1) \tag{3.8}$$

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. The term *variance* was introduced as the squared standard deviation by William Sealy Gossett publishing under the name of "Student" (Pearson, 1923).

```
> 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
   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 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. sd returns the standard deviation of a vector or each column of a data frame, var returns the variance and covariances of each column of a data frame or of a matrix.

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

3.7.4 Coefficient of variation

Particularly when using values that have the appearance of ratio measurement (e.g., dollars, reaction times, micro liters of a biological assay) an index of how much variation there is compared to the mean level is the *coefficient of variation*. This is simply the ratio of the standard deviation to the sample mean. Although not commonly seen in psychometrics, the

CV will be seen in biological journals (reporting the error of the assay), financial reports as well as manufacturing process control situations.

$$CV = \frac{\sigma_x}{\bar{X}} \tag{3.9}$$

3.8 Geometric interpretations of Variance and Covariance

It is sometimes useful to think of data geometrically. A set of n scores on a single variable may be thought of geometrically as representing a vector in n-dimensional space where the dimensions of the space represent the individual scores. Center this vector on the grand mean (i.e., convert the scores to deviation scores). Then the length of this vector is the square root of the sums of squares and the average length of this vector across all dimensions is the standard deviation.

Another measure of dispersion is the *average squared distance* between the n data points. This is found by finding all n^2 pairwise distances, squaring them, and then dividing by n^2. But since the diagonal of that matrix is necessarily zero, it is more appropriate to divide by n*(n-1). This value is, it turns out, twice the variance. Remembering that standard deviation is the square root of the variance, we find that the *average distance* between any two data points is $\sigma_x\sqrt{2}$.

Why is this? Consider the matrix of distances between pairs of data points:

$$\begin{pmatrix} 0 & X_1 - X_2 & ... & X_1 - X_n \\ X_2 - X_1 & 0 & ... & X_n - X_2 \\ ... & ... & 0 & ... \\ X_n - X_1 & X_n - X_2 & ... & 0 \end{pmatrix}$$

Square each element:

$$\begin{pmatrix} 0 & X_1^2 + X_2^2 - 2X_1X_2 & ... & X_1^2 + X_n^2 - 2X_1X_n \\ X_1^2 + X_2^2 - 2X_1X_2 & 0 & ... & X_2^2 + X_n^2 - 2X_2X_n \\ ... & ... & 0 & ... \\ X_1^2 + X_n^2 - 2X_1X_n & X_2^2 + X_n^2 - 2X_2X_n & ... & 0 \end{pmatrix}.$$

Sum all of these elements to obtain

$$\sum_{i=1}^{n} d_i^2 = 2n\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j \tag{3.10}$$

The average squared distance may be obtained by dividing the total squared distance by n^2 (to obtain a population variance) or by $n(n-1)$ to obtain the sample estimate of the variance.

$$\bar{d^2} = 2(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j)/n)/(n-1) \tag{3.11}$$

But this is just the same as

$$2(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i X_.)/(n-1) = 2(\sum_{i=1}^{n} X_i^2 - nX_.^2)/(n-1) \tag{3.12}$$

which is twice the variance:

$$\sigma^2 = x^T x/(n-1) = (X - X_.)^T (X - X_.)/(n-1) = \sum_{i=1}^{n} (X_i - X_.)^2/(n-1) = (\sum_{i=1}^{n} X_i^2 - nX_.^2)/(n-1) \tag{3.13}$$

That is, the average distance between any two data points will be $\sigma_x\sqrt{2}$. Knowing the standard deviation allows us to judge not just how likely a point is deviate from the mean, but also how likely two points are to differ by a particular amount.

integrate this section with previous section

3.9 Variance, Covariance, and Distance

There are a variety of ways to conceptualize variance and covariance. Algebraically, for a vector X with elements X_i, variance is the average of the sum of squared distances from the mean ,$X_.$, (Equation 3.13) or alternatively, 1/2 of the average squared distance between any two points (Equation 3.12). For two vectors, X_1 and X_2, the covariance between them may be similarly defined as the average product of deviation scores:

$$Cov_{12} = \sum_{i=1}^{n} x_{1i}x_{2i}/n = \sum_{i=1}^{n} (X_{1i} - X_{1.})(X_{2i} - X_{2.})/n = \{\sum_{i=1}^{n} X_{1i}X_{2i} - \sum_{i=1}^{n} X_{1i} \sum_{i=1}^{n} X_{2i}/n\}/n. \tag{3.14}$$

A spatial interpretation of covariance may be expressed in terms of the average distance between the corresponding points in X_1 and X_2. For simplicity, express each vector in terms of deviations from the respective means: $x_{1i} = X_{1i} - X_{1.}$

$$dist_{12}^2 = \frac{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}{n} = \frac{\sum_{i=1}^{n} (x_{1i}^2 + x_{2i}^2 - 2x_{1i}x_{2i})}{n} = Var_1 + Var_2 - 2Cov_{12} \tag{3.15}$$

That is, the covariance is the difference between the average of the variances of each vector (which are themselves just twice the average squared distances between each point on a vector) and half the average squared distance between the corresponding pair of points on each vector.

$$Cov_{12} = \frac{\{Var_1 + Var_2 - dist_{12}^2\}}{2} = \frac{\{Var_1 + Var_2\}}{2} - \frac{dist_{12}^2}{2}. \tag{3.16}$$

If each element of X_1 is the same as each element of X_2, then the pairwise distances are zero, the two variance are identical, and the covariance is the same as the variance.

3.10 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 to ignore the units of measurement and to convert deviation scores into *standard scores*. That is, to divide deviation scores by the standard deviation:

$$z_i = x_i/\sigma_x = (X - X.)/\sqrt{Var_X} \tag{3.17}$$

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 the z_i scores by a constant and then adding a different constant (Table 3.11). The `rescale` function does this by using the `scale` function to first convert the data to z scores, and then multiplies by the desired standard deviation and adds the desired mean (see Figure 3.8).

Table 3.11 Raw scores (X_i) are typically converted into deviation scores (x_i) or standard scores (z_i). These are, in turn, the transformed into "public" scores for communication to laypeople.

	Transformation	Mean	Standard Deviation
Raw Data		$X. = \sum(X_i)/n$	$s_x = \sqrt{\sum(X_i - X.)^2/(n-1)}$
deviation score	$x_i = X_i - X.$	0	$s_x = \sqrt{\sum(x_i)^2/(n-1)}$
standard score	$z_i = x_i/s_x$	0	1
"IQ"	z_i *15+100	100	15
"SAT"	z_i*100+500	500	100
"ACT"	z_i*6+18	18	6
"T-score"	z_i*10+ 50	50	10
"Stanine"	z_i*2.0+5	5	2.0

3.11 Assessing the higher order moments of the normal and other distributions

The *central limit theorem* shows that the distribution of the means of independent identically distributed samples with finite means and variances will tend asymptotically towards the *normal distribution* originally described by DeMoivre in 1733, by Laplace in 1774 and Gauss in 1809 and named by Galton (1877) and others discussed by Stigler (1986). The equation for the *normal curve* expressed in terms of the mean and standard deviation is

$$f(x,\mu,\sigma) = N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{3.18}$$

Three normal curves, differing in their mean and standard deviation (i.e, $N(0,1), N(0,2)$ and $N(1,2)$) are shown in Figure 3.7. Although typically shown in terms of $N(0,1)$, alternative scalings of the normal seen in psychology and psychometrics have different values of the mean and standard deviation (Table 3.11 and Figure 3.8) partly in order to facilitate communication with non-statisticians, and partly to obscure the meaning of the scores.

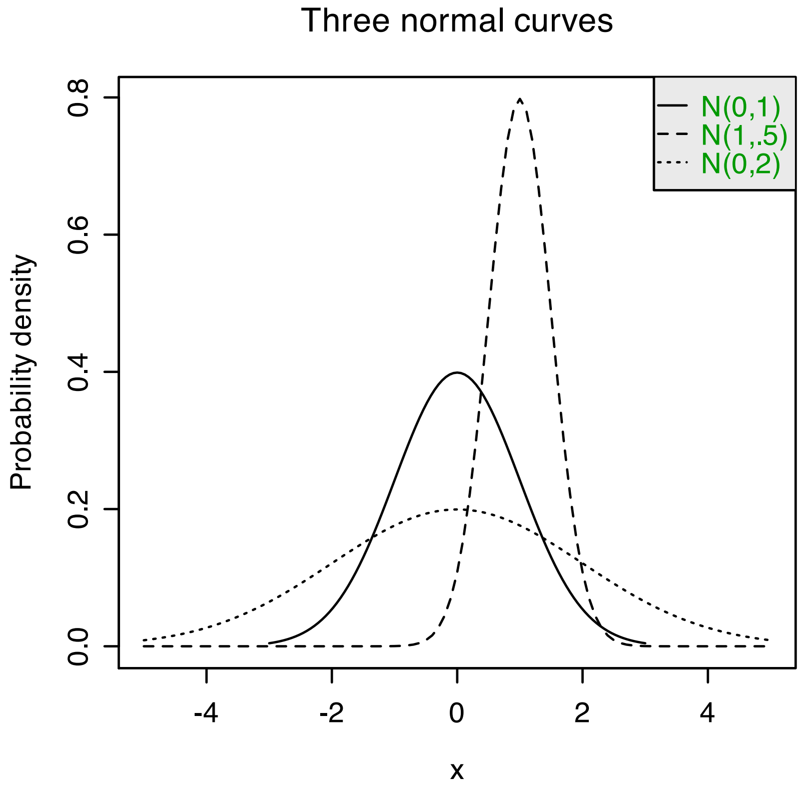

Fig. 3.7 Normal curves can differ in their location (mean) as well as width (standard deviation). Shown are normals with means of 0 or 1 and standard deviations of 1 or 2.

In addition to its mathematical simplicity, the normal distribution is seen in many settings where the accumulation of errors is random (e.g., astronomical observations) or made up of many small sources of variance (the distribution of height among Belgian soldiers as described by Quetelet in 1837 (Stigler, 1999). Unfortunately, real data rarely are so easily described. Karl Pearson (1905) made this distinction quite clearly:

> The chief physical differences between actual frequency distributions and the Gaussian theoretical distributions are:

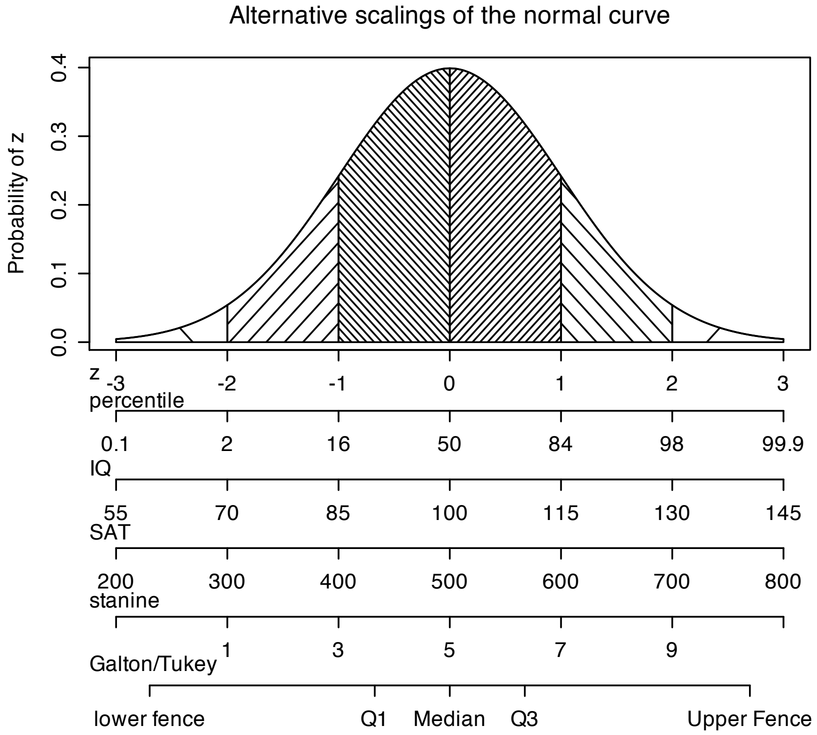

Fig. 3.8 The normal curve may be expressed in standard (z) units with a mean of 0 and a standard deviation of 1. Alternative scalings of the normal include “percentiles” (a non linear transformation of the z scores), “IQ” scores with a mean of 100, and a standard deviation of 15, “SAT/GRE” scores with a mean of 500 and a standard deviation of 100, “ACT” scores with a mean of 18 and a standard deviation of 6, or “standardized nines - stanines” with a mean of 5 and a standard deviation of 2. Note that for stanines, each separate score refer to the range from -.5 to +.5 from that score. Thus, the 9th stanine includes the z-score region from 1.75z and above and has 4% of the normal population. The 5 numbers of the box plot correspond to the lower whisker, 25th, 50th and 75th percentiles, and the upper whisker.

> (i) The significant separation between the mode of position of maximum frequency and the average or mean character.
>
> (ii) The ratio of this separation between mean and mode to the variability of the character–a quantity I have termed the *skewness*.
>
> (iii) A degree of flat-toppedness which is greater or less than that of the normal curve. Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat-topped than the normal curve. If more flat-topped I term them *platykurtic*, if less flat-topped *leptokurtic*, and if equally flat-topped *mesokurtic*. A frequency distribution may be symmetrical, satisfying both the first two conditions for normality, but it may fail to be *mesokurtic*, and thus the Gaussian curve cannot describe it. (Pearson, 1905, p 173).

Just as the variance is the second moment around the mean and describes the width of the distribution, so does the *skew* (the third moment) describe the shape and the *kurotosis* (the fourth moment) the peakedness versus flatness of the distribution. Pearson (1905)

$$skew = \gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\sqrt{n}\sum_{i=1}^{n} x_i)^3}{(\sum_{i=1}^{n} (x_i^2)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n} (X_i - X.)^3}{(\sum_{i=1}^{n} (X_i - X.)^2)^{3/2}} \tag{3.19}$$

The *standard error of skew* is

$$\sigma_{\gamma_1} = \sqrt{\frac{6}{N}} \tag{3.20}$$

Distributions with positive skew have long right tails while those with negative skew have long left tails. Examples of positively skewed distribution are common in psychological measures such as reaction time Ratcliff (1993) or measures of negative affect Rafaeli and Revelle (2006). As we shall see later (Chapter 4), differences in skew are particularly important in the effect they have on correlations. Positively skewed reaction time data are sometimes modeled as *log normal distributions* or sometimes as *Weibull distribtions*. Just as the normal represents the sum of Independently and Identically Distributed random variables (*IID*s), so does the log normal represent the product of *IID*s. Such a positively skewed distribution that is commonly seen in economics is the *log normal distribution* which can reflect a normal distribution of multiplicative growth rates (Figure 3.9) and is seen in the distribution of income in the United States. That is, if the percentage raise given employees is normally distributed, the resulting income distribution after several years of such raises will be log normal. Cognitive processes operating in a cascade can also be thought of in terms of the log normal distribution. Estimating the central tendency of skewed distributions is particularly problematic, for the various estimates discussed earlier will differ drastically. Consider the curve generated using the `dlnorm` function set with a log mean of 10.5 and a sd of .8. These values were chosen to give a rough example of the distribution of family income in the US which in 2008 had a median of $50,302, a mean of $68,204 and a trimmed mean of $56,720. (See the `income` data set for the data). An even more drastic curve is the power law ($f(n) = K/n^a$) summarizing the distribution of publications of Ph.Ds. with a mode of 0 and an upper range in the 1,000s (Anderson et al., 2008; Lotka, 1926; Vinkler, 2007).

Platykurtic distributions (kurtosis > 0) have more of their density in the center of the distribution than would be expected given the magnitude of their standard deviations. *Leptokurtic distributions*, on the other hand, have "fatter tails" than would be expected given their standard deviation Pearson (1905). ("Student" introduced the mnemonic that a platypus has a short tail and that kangaroos, who are known for "lepping" have long tails Student (1927)).

$$kurtosis = \gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{(\sum_{i=1}^{n} x_i)^4}{\sum_{i=1}^{n} x_i^4} - 3 = \frac{(\sum_{i=1}^{n} (X_i - X.))^4}{\sum_{i=1}^{n} (X_i - X.)^4} - 3 \tag{3.21}$$

Given the standard error of the skew (Equation 3.20) and the *standard error of kurtosis* (Equation 3.22), it is possible to test whether a particular distribution has excess skew or kurtosis.

$$\sigma_{\gamma_2} = \sqrt{\frac{24}{N}} \tag{3.22}$$

Although it is frequently reported that in positively skewed distribution, the mode will be less than the mean which will be less than the median (e.g., Figure 3.9 , this is not always the case. von Hippel (2005) discusses a number of counter examples.

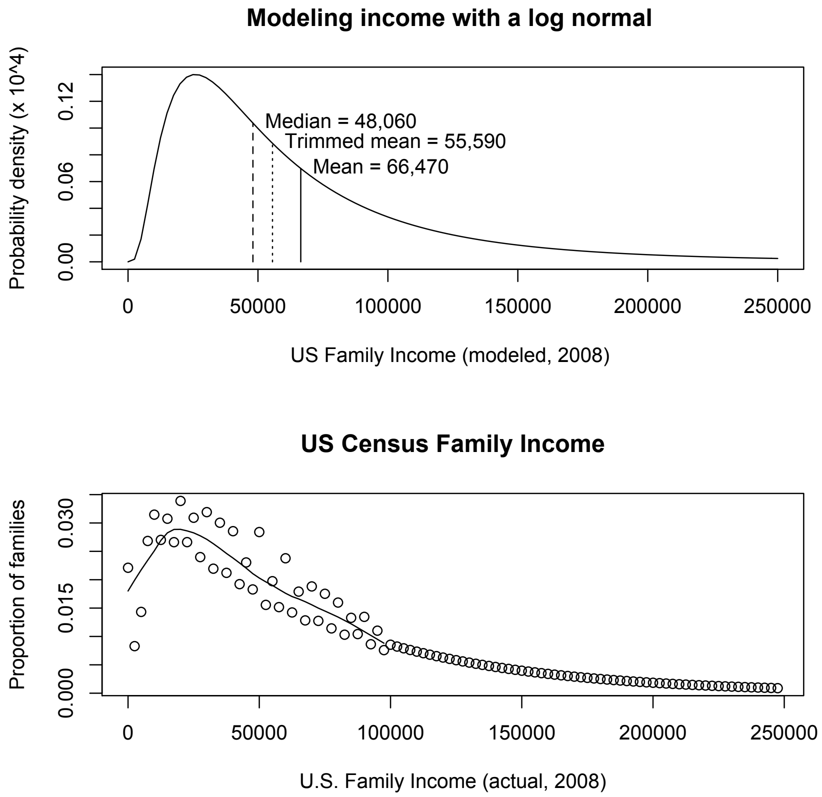

Fig. 3.9 A log normal distribution is skewed to the right and represents the distribution of normally distributed multiplicative processes. An example of US. family income distributions adapted from the US Census (2008) is shown with a mean of $66,570, trimmed mean of $55,590, and median of $48,060. Means, the median, skew and kurtosis were found from simulating 10,000 cases from the log normal with a log mean of 10.5 and sd of .8. Curve drawn with the `curve` function plotting the `dlnorm` function with mean of 10.5 and sd of .8: curve(dlnorm(x, 10.8, .8), x = c(0,250000)). The top panel shows the modeled data, the lower panel, the actual data. Values for income above 100,000 are inferred from the census data categories of 100-150, 150-200, 200-250 and fit with a negative exponential. See `income` for the US census data set on family income. The smooth curve for the numbers less than 100,000 is generated using the `lowess` function. The sawtooth alternation of the actual data suggests that people are reporting their income to the nearest $5,000.

It is helpful to consider the distributions generated from several different families of distributions to realize that just because a distribution is symmetric and peaks in the middle does not tell us much about the length of the tails. Consider the four distributions shown in Figure 3.10. The top and bottom curves are normal, one with standard deviation 1, one with standard deviation 2. Both of these have 0 skew and 0 kurtosis. However the other two, the logistic and the Cauchy are definitely not normal. In fact, the Cauchy has infinite variance and kurtosis!

The *Cauchy distribution* is frequently used as a counter example to those who want to generalize the central limit theorem to all distributions, for the means of observations from the Cauchy distribution are not distributed normally, but rather remain distributed as before. The distribution is sometimes referred to as the "*witch of Agnesi*" (Stigler, 1999). The function is

$$f(x) = \frac{1}{\pi(1+x^2)} \tag{3.23}$$

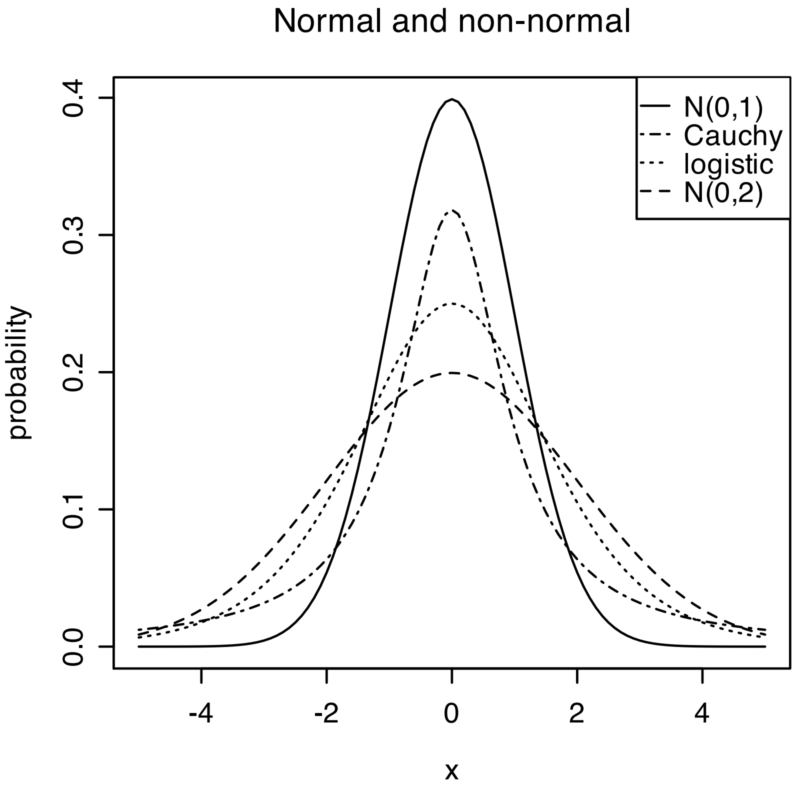

Fig. 3.10 Symmetric and single peaked is not the same as being a normal distribution. Two of these distributions are normal, differing only in their standard deviations, one, the logistic has slightly more kurtosis, and one (the Cauchy) has infinite variance and kurtosis.

3.12 Generating commonly observed distributions

Many statistics books include tables of the t or F or χ^2 distribution. By using R this is unnecessary since these and many more distributions can be obtained directly. Consider the normal distribution as an example. `dnorm(x, mean=mu, sd=sigma)` will give the probability density of observing that x in a distribution with mean=mu and standard deviation= sigma. `pnorm(q,mean=0,sd=1)` will give the probability of observing the value q or less. `qnorm(p, mean=0, sd=1)` will give the quantile value of a value with probability p. `rnorm(n,mean,sd)` will generate n random observations sampled from the normal distribution with specified mean and standard deviation. Thus, to find out what z value has a .05 probability we ask for `qnorm(.05)`. Or, to evaluate the probability of observing a z value of 2.5, specify `pnorm(2.5)`. (These last two examples are one side p values).

Applying these prefixes (d,p,q, r) to the various distributions available in R allows us to evaluate or simulate many different distributions (Table 3.12).

Table 3.12 Some of the most useful distributions for psychometrics that are available as functions. To obtain the density, prefix with *d*, probability with *p*, quantiles with *q* and to generate random values with *r*. (e.g., the normal distribution may be chosen by using dnorm, pnorm, qnorm, or rnorm.) Each function has specific parameters, some of which take default values, some of which require being specified. Use *help* for each function for details.

Distribution	base name	Parameter 1	Parameter 2	Parameter 3	example application
Normal	norm	mean	sigma		Most data
Multivariate normal	mvnorm	mean	r	sigma	Most data
Log Normal	lnorm	log mean	log sigma		income or reaction time
Uniform	unif	min	max		rectangular distributions
Binomial	binom	size	prob		Bernuilli trials (e.g. coin flips)
Student's t	t	df		non-centrality	Finding significance of a t-test
Multivariate t	mvt	df	corr	non-centrality	Multivariate applications
Fisher's F	f	df1	df2	non-centrality	Testing for significance of F test
χ^2	chisq	df		non-centrality	Testing for significance of χ^2
Beta	beta	shape1	shape2	non-centrality	distribution theory
Cauchy	cauchy	location	scale		Infinite variance distribution
Exponential	exp	rate			Exponential decay
Gamma	gamma	shape	rate	scale	distribution theoryh
Hypergeometric	hyper	m	n	k	
Logistic	logis	location	scale		Item Response Theory
Poisson	pois	lambda			Count data
Weibull	weibull	shape	scale		Reaction time distributions

3.13 Mixed distributions

The standard deviation and its associated transformations are useful if the data are normal. But what if they are not? Consider the case of participants assessed on some measure. 90% of these participants are sampled from a normal population with a mean of 0 and a standard deviation (and variance) of 1. But if 10% of the participants are sampled from a population with the same mean but 100 times as much variance (sd = 10), the pooled variance of the sample will be .9 * 1 + .1 * 100 or 10.9 and the standard deviation will be 3.3. Although it would seem obvious that these two distributions would appear to be very different, this is not the case (see Figure 3.11 and Table 3.13). As discussed by Wilcox (2005), even if the contaminated distribution is a mixture of 90% N(0,1) and 10% N(3,40), the plots of the uncontaminated and contaminated distributions look very similar.

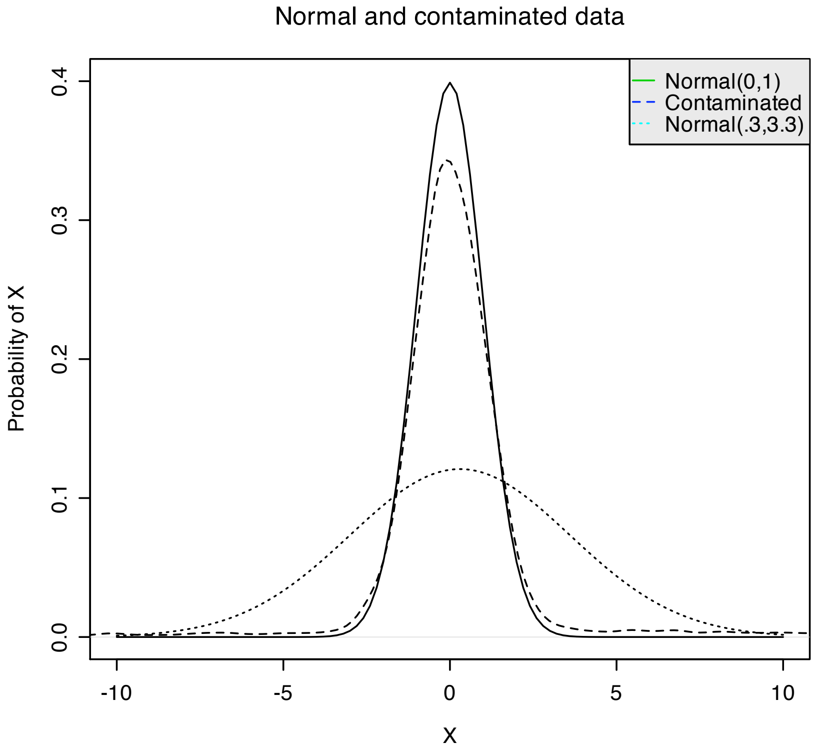

Fig. 3.11 Probability density distributions for a normal distribution with mean 0 and standard deviation 1, a contaminated distribution (dashed line) formed from combining a N(0,1) with a N(3,10), and a normal with the same mean and standard deviation as the contaminated N(.3,3.3) (dotted line). Adapted from Wilcox, 2005.

3.14 Robust measures of dispersion

Estimates of central tendency and of dispersion that are less sensitive to contamination and outliers are said to be *robust estimators*. Just as the median and trimmed mean are less sensitive to contamination, so is the median absolute difference from the median (`mad`). Consider the following seven data sets: The first one (x) is simply a normal distribution with mean 0 and sd of 1. Noise10, Noise20, and Noise40 are normals with means of 3 and standard deviations of 10, 20, and 40 respectively. Mixed10 is made up of a mixture of 90% sampled from x and 10% sampled from Noise10. Mixed20 has the same sampling frequencies, but noise is sampled from Noise20. Similarly for Mixed40. X and the noise samples are created using the `rnorm` function to create random data with a normal distribution with a specified mean and standard deviation. The mixtures are formed by combining (using `c`) random samples (using `sample`) of the X and noise distributions. Descriptive statistics are found by `describe`.

The first four variables (X, Noise10, Noise20, and Noise40) are normally distributed, and the means, trimmed means, and medians are almost identical. Similarly, the standard deviations and median absolute deviations from the medians (MAD) are almost identical. But this is not the case for the contaminated scores. Although the simple arithmetic means of the mixed distributions reflect the contamination, the trimmed means (trimmed by dropping the top and bottom 10%) and medians are very close to that of the uncontaminated distribution. Similarly the MAD of the contaminated scores are barely affected (1.14 versus .99) even though the standard deviations are drastically larger (12.58 versus 1.0).

Table 3.13 Generating distributions of normal and contaminated normal data. A normal distribution with N(0,1) is 10% contaminated with N(3,10), N(3,20) or N(3,40). Although these mixtures are formed from two normals, they are not normal, but rather have very heavy tails. Observe how the median and trimmed mean are not affected by the contamination. Figure 3.11 shows a plot of the probability density of the original and the mixed10 contaminated distribution. The contamination may be detected by examining the difference between the standard deviation and the median absolute deviation from the median or the kurtosis.

```
> n <- 10000
> frac  <-  .1
> m  <-  3
> x <- rnorm(n)
> noise10 <- rnorm(n,m,sd=10)
> mixed10 <- c(sample(x,n * (1-frac),replace=TRUE),sample(noise10,n*frac,replace=TRUE))
> dmixed <- density(mixed,bw=.3,kernel="gaussian")
> noise20 <- rnorm(n,m,sd=20)
> noise40 <- rnorm(n,m,sd=40)
> mixed20 <- c(sample(x,n * (1-frac),replace=TRUE),sample(noise20,n*frac,replace=TRUE))
> mixed40 <- c(sample(x,n * (1-frac),replace=TRUE),sample(noise40,n*frac,replace=TRUE))
> data.df <- data.frame(x,noise10,noise20,noise40,mixed10,mixed20,mixed40)
> describe(data.df)
```

```
        var     n mean    sd median trimmed   mad     min    max  range skew kurtosis   se
x         1 10000 0.01  1.00   0.01    0.01  0.99   -4.28   3.71   7.99 0.02    -0.04 0.01
noise10   2 10000 2.89  9.91   2.94    2.89  9.95  -32.96  46.39  79.35 0.01     0.02 0.10
noise20   3 10000 3.14 20.21   3.10    3.12 20.25  -74.10  85.71 159.81 0.01    -0.02 0.20
noise40   4 10000 3.49 40.41   3.05    3.23 40.51 -149.29 169.40 318.69 0.06    -0.07 0.40
mixed10   5 10000 0.31  3.38   0.05    0.06  1.10  -30.01  46.39  76.40 2.29    24.49 0.03
mixed20   6 10000 0.27  6.41   0.02    0.02  1.14  -66.22  64.16 130.38 1.16    26.01 0.06
mixed40   7 10000 0.25 12.58   0.00    0.00  1.14 -117.71 169.40 287.11 1.01    27.64 0.13
```

Robust estimates of central tendency and of dispersion are important to consider when estimating experimental effects, for although conventional tests such as the *t-test* and *F-test* are not overly sensitive to *Type I errors* when the distributions are not normal, they are very sensitive to *Type II errors*. That is to say, if the data are not normal due to contamination as seen is Figure 3.11, true differences of central tendencies will not be detected by conventional tests (Wilcox, 1987; Wilcox and Keselman, 2003; Wilcox, 2005). Robust estimates of central tendency and robust equivalents of the t and F tests are slightly less powerful when the data are truly normal, but much more powerful in cases of non-normality Wilcox and Keselman

(2003); Wilcox (2005). Functions to do robust analysis are available in multiple packages, including **MASS**, **robust**, and **robustbase** as well as from the web pages of various investigators (e.g. Rand Wilcox at the University of Southern California).

3.15 Monotonic transformations of data and "Tukey's ladder"

If the data are non-normal or if the relations are non-linear, what should we do? John Tukey (1977) suggested a *ladder of tranformations* that can be applied to the data (Table 3.14, Figure 3.12). These transformations have the effect of emphasizing different aspects of the data. If the data are skewed heavily to the right (e.g., for reaction times or incomes), taking logs or reciprocals deemphasizes the largest numbers and makes distinctions between the smaller numbers easier to see. Similarly, taking squares or cubes of the data can make some relationships much clearer. Consider the advantage of treating distance travelled as a function of squared time when study the effects of acceleration. Similarly, when examining the damaging effects of wind intensity upon houses, squaring the wind velocity leads to a better understanding of the effects. The appropriate use of the ladder of transformations is look at the data, look at the distributions of the data, and then look at the bivariate plots of the data. Try alternative transforms until these exploratory plots look better. Data analysis is detective work and requires considering many alternative hypotheses about the best way to treat the data.

Table 3.14 Tukey's ladder of transformations. One goes up and down the ladder until the relationships desired are roughly linear or the distribution is less skewed. The effect of taking powers of the numbers is to emphasize the larger numbers, the effect of taking roots, logs, or reciprocals is to emphasize the smaller numbers.

Transformation	effect	
x^3	emphasize large numbers	reduce negative skew
x^2	emphasize large numbers	reduce negative skew
x	the basic data	
$\sqrt{x}$	emphasize smaller numbers	reduce positive skew
-1/x	emphasize smaller numbers	reduce positive skew
log(x)	emphasize smaller numbers	reduce positive skew
$-1/x^2$	emphasize smaller numbers	reduce positive skew
$-1/x^3$	emphasize smaller numbers	reduce positive skew

mention Box Cox?

3.16 What is the fundamental scale?

It would be nice to be able to answer this question with a simple statement, but the answer is really that it all depends. It depends upon what is being measured and what inferences we are trying to draw from the data. We have recognized for centuries that money, whether expressed in dollars, ducats, Euros, Renminbi, or Yen is measured in a linear, ratio scale but has a negatively accelerated effect upon happiness (Bernoulli, 1738). (That is, the utility

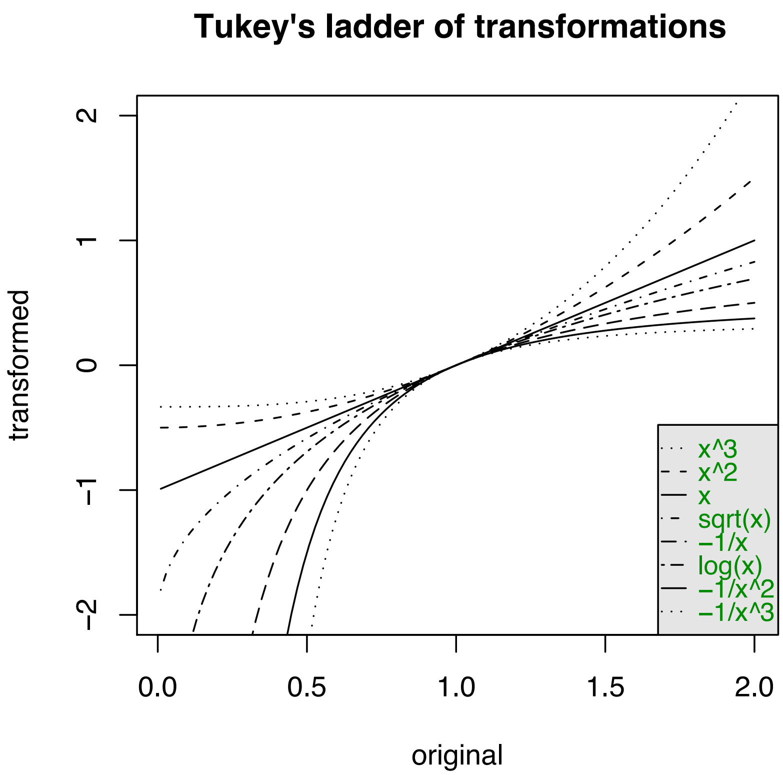

Fig. 3.12 Tukey (1977) suggested a number of transformations of data that allow relationships to be seen more easily. Ranging from the cube to the reciprocal of the cube, these transformations emphasize different parts of the distribution.

of money is negatively accelerated.) The perceived intensity of a stimulus is a logarithmic function of the physical intensity (Weber, 1834b). The probability of giving a correct answer on a test is an increasing but non-linear function of the normal way we think of ability (Embretson and Hershberger, 1999; McDonald, 1999). The amount of energy used to heat a house is a negative but linear function of the outside temperature. The time it takes to fall a particular distance is a function of the square root of that distance. The gravitational attraction between two masses is a function of the inverse of the squared distance. The hull speed of a sailboat is function of the square root of the length of the boat. Sound intensity in decibels is expressed in logarithmic units of the ratio of the power of the observed sound to the the power of a reference sound. The units of the pH scale in chemistry are (negative) logarithmic units of the concentration of hydrogen ions.

The conclusion from these examples is that the appropriate scale is one that makes the relationships between our observed variables and manipulations easier to understand and to communicate. The scales of our observed variables are reflections of the values of our latent variables (Figure 3.1) and are most useful when they allow us to simplify our inferences about the the relationships between the latent variables. By not considering the scaling properties of our observations it is easy to draw incorrect conclusions about the the underlying processes (consider the example discussed in section 3.6). By searching for the transformations that allow us to best represent the data perhaps we are able to better understand the latent processes involved.