An introduction to Psychometric Theory
Issues in Scaling

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2019

Outline

Four types of scales and their associated statistics

Table: 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.

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

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

Graphical and tabular summaries of data

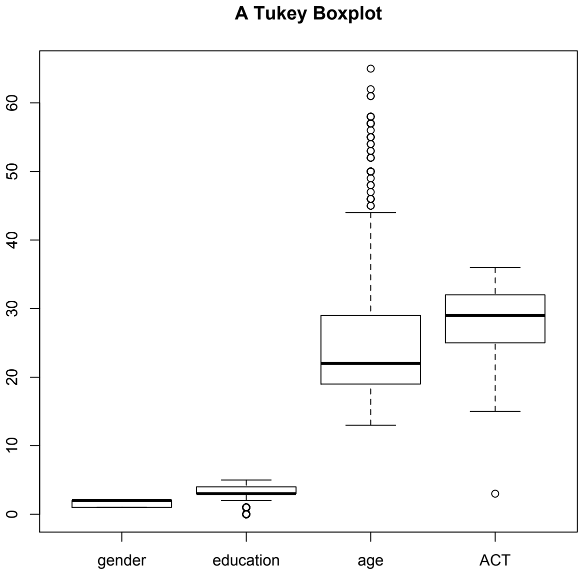
1. The Tukey 5 number summary shows the important characteristics of a set of numbers
   - Maximum
   - 75th percentile
   - Median (50th percentile)
   - 25th percentile
   - Minimum
2. Graphically, this is the box plot
   - Variations on the box plot include confidence intervals for the median

The summary command gives the Tukey 5 numbers

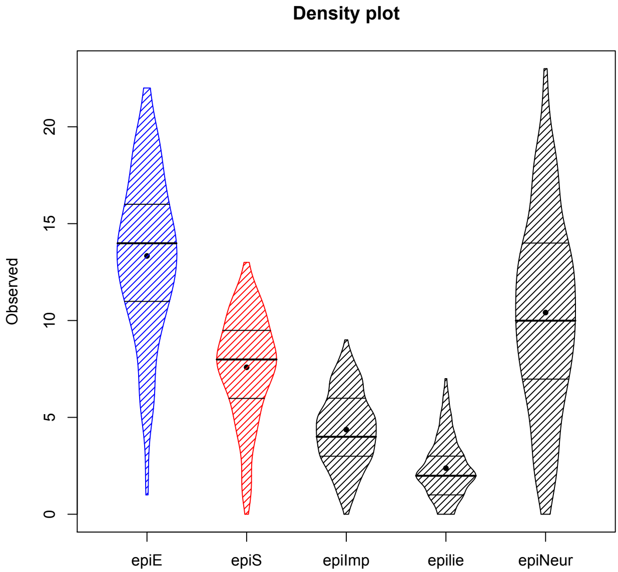
```
> summary(sat.act)

    gender        education          age             ACT             SATV
 Min.   :1.000   Min.   :0.000   Min.   :13.00   Min.   : 3.00   Min.   :200.0
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.:19.00   1st Qu.:25.00   1st Qu.:550.0
 Median :2.000   Median :3.000   Median :22.00   Median :29.00   Median :620.0
 Mean   :1.647   Mean   :3.164   Mean   :25.59   Mean   :28.55   Mean   :612.2
 3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:29.00   3rd Qu.:32.00   3rd Qu.:700.0
 Max.   :2.000   Max.   :5.000   Max.   :65.00   Max.   :36.00   Max.   :800.0
```

A box plot of the first 4 sat.act variables



A violin or density plot of the first 5 epi.bfi variables



The `describe` function gives more descriptive statistics

```
> describe(sat.act)

          vars   n   mean     sd median trimmed    mad min max range  skew kurtosis   se
gender       1 700   1.65   0.48      2    1.68   0.00   1   2     1 -0.61    -1.62 0.02
education    2 700   3.16   1.43      3    3.31   1.48   0   5     5 -0.68    -0.07 0.05
age          3 700  25.59   9.50     22   23.86   5.93  13  65    52  1.64     2.42 0.36
ACT          4 700  28.55   4.82     29   28.84   4.45   3  36    33 -0.66     0.53 0.18
SATV         5 700 612.23 112.90    620  619.45 118.61 200 800   600 -0.64     0.33 4.27
SATQ         6 687 610.22 115.64    620  617.25 118.61 200 800   600 -0.59    -0.02 4.41
```

Multiple measures of central tendency

mode The most frequent observation. Not a very stable measure, depends upon grouping. Can be used for categorical data.

median The number with 50% above and 50% below. A powerful, if underused, measure. Not sensitive to transforms of the shape of the distribution, nor outliers. Appropriate for ordinal data, and useful for interval data.

mean One of at least seven measures that assume interval properties of the data.

Multiple ways to estimate the mean

Arithmetic mean $\bar{X} = X_{.} = (\sum_{i=1}^{N} X_i)/N$ `mean(x)`

Trimmed mean throws away the top and bottom t% of observations. This follows the principle that all data are normal at the middle. `mean(x,trim=.1)`

Winsorized mean Find the arithmetic mean after replacing the n lowest observations with the nth value, and the N largest values with the Nth largest. `winsor(x,trim=.2)`

Geometric Mean $\bar{X}_{geometric} = \sqrt[N]{\prod_{i=1}^{N} X_i} = e^{\Sigma(ln(x))/N}$ (The anti-log of the mean log score). `geometric.mean(x)`

Harmonic Mean $\bar{X}_{harmonic} = \frac{N}{\sum_{i=1}^{N} 1/X_i}$ (The reciprocal of the mean reciprocal). `harmonic.mean(x)`

Circular Mean $\bar{x}_{circular} = tan^{-1}\left(\frac{\sum cos(x)}{\sum sin(x)}\right)$ `circular.mean(x)` (where x is in radians)

circadian.mean `circular.mean(x)` (where x is in hours)

Circular statistics

Table: 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

Some hypothetical data stored in a data.frame

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

```
> s.df <- read.clipboard()
> dim(s.df)  #how many elements are in each dimension
[1] 7 7
> str(s.df) #show the structure
'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
```

ga

Saving the data.frame in a readable form

The previous slide is readable by humans, but harder to read by computer. PDFs are formatted in a rather weird way. We can share data on slides by using the dput function. Copy this output to your clipboard from the slide, and then get it into Rdirectly.

```
> dput(sf.df)
```

```
structure(list(ID = 1:7, Name = structure(c(2L, 4L, 1L, 7L, 5L,
6L, 3L), .Label = c("Alice", "Bob", "Chuck", "Debby", "Eric",
"Fred", "Gina"), class = "factor"), gender = structure(c(2L,
1L, 1L, 1L, 2L, 2L, 2L), .Label = c("Female", "Male"), class = "factor"),
    theta = c(1L, 3L, 7L, 6L, 4L, 5L, 2L), X = c(12L, 14L, 18L,
    17L, 15L, 16L, 13L), Y = c(2L, 6L, 14L, 12L, 8L, 10L, 4L),
    Z = c(1L, 4L, 64L, 32L, 8L, 16L, 2L)), .Names = c("ID", "Name",
"gender", "theta", "X", "Y", "Z"), class = "data.frame", row.names = c(NA,
-7L))
```

```
my.data <- structure(list(ID = 1:7, Name = structure(c(2L, 4L, 1L, 7L, 5L,
6L, 3L), .Label = c("Alice", "Bob", "Chuck", "Debby", "Eric",
"Fred", "Gina"), class = "factor"), gender = structure(c(2L,
1L, 1L, 1L, 2L, 2L, 2L), .Label = c("Female", "Male"), class = "factor"),
    theta = c(1L, 3L, 7L, 6L, 4L, 5L, 2L), X = c(12L, 14L, 18L,
    17L, 15L, 16L, 13L), Y = c(2L, 6L, 14L, 12L, 8L, 10L, 4L),
    Z = c(1L, 4L, 64L, 32L, 8L, 16L, 2L)), .Names = c("ID", "Name",
"gender", "theta", "X", "Y", "Z"), class = "data.frame", row.names = c(NA,
-7L))
```

Sorting the data can display certain features

We use the `order` function applied to the "Names" column and then to the 4th column.

```
> my.data.alpha <-
     my.data[order(my.data[,"Name"]),]
> my.data.alpha

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

```
> my.data.theta <-
    my.data[order(my.data[,4]),]
> my.data.theta

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

It was harder to see the perfect relationship between θ and X, Y, and Z with the original data.

Multiple estimates of the central tendency using the apply function

```
> apply(my.data[4:7],2,mean)
   theta        X        Y        Z
 4.00000 15.00000  8.00000 18.14286
> apply(my.data[4:7],2,mean,trim=.2)
theta     X     Y     Z
  4.0  15.0   8.0  12.4
> apply(my.data[4:7],2,winsor.mean,trim=.2)
   theta        X        Y        Z
 4.00000 15.00000  8.00000 12.91429
> apply(my.data[4:7],2,harmonic.mean)
    theta         X         Y         Z
 2.699725 14.729687  5.399449  3.527559
> apply(my.data[4:7],2,geometric.mean)
    theta         X         Y         Z
 3.380015 14.865151  6.760030  8.000000
```

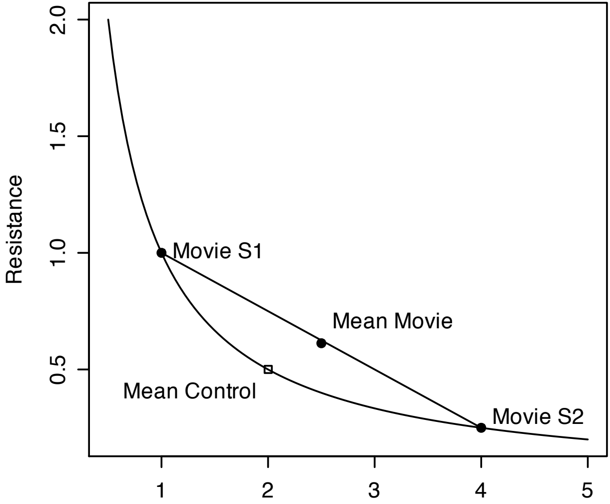
1. The basic mean is applied to columns 4 - 7
2. Then do this, but trim the top and bottom 20%
3. Now, don't trim, but winsorize
4. Compare with the harmonic mean
5. Compare with geometric mean.

Effect of reciprocal transformation upon means

Table: Hypothetical study of arousal using an exciting movie. The post test shows greater arousal if measured using skin conductance (higher skin conductance means more arousal), but less arousal if measured using skin resistance (higher skin conductance means less arousal)

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

Non linearity can influence means if the variances differ.



What is the "average" class size?

Table: 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. But what about the students?

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

Class size from the students' point of view.

Table: 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

Time in therapy

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.

Does teaching effect learning?

1. A leading research team in motivational and educational psychology was interested in the effect that different teaching techniques at various colleges and universities have upon their students. They were particularly interested in the effect upon writing performance of attending a very selective university, a less selective university, or a two year junior college.
2. A writing test was given to the entering students at three institutions in the Boston area. After one year, a similar writing test was given again. Although there was some attrition from each sample, the researchers report data only for those who finished one year. The pre and post test scores as well as the change scores were as shown below:

Types of teaching affect student outcomes?

Table: Three types of teaching and their effect on student outcomes

School	Pretest	Posttest	Change
Junior College	1	5	4
Non-selective university	5	27	22
Selective university	27	73	45

From these data, the researchers concluded that the quality of teaching at the selective university was much better than that of the less selective university or the junior college and that the students learned a great deal more. They proposed to study the techniques used there in order to apply them to the other institutions.

Teaching and math performance

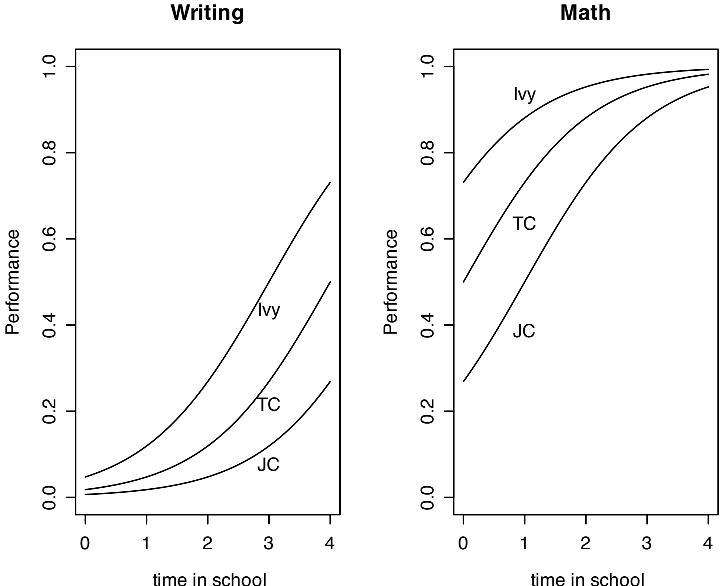
Another research team in motivational and educational psychology was interested in the effect that different teaching at various colleges and universities affect math performance. They used the same schools as the previous example with the same design.

Table: Three types of teaching and their effect on student outcomes

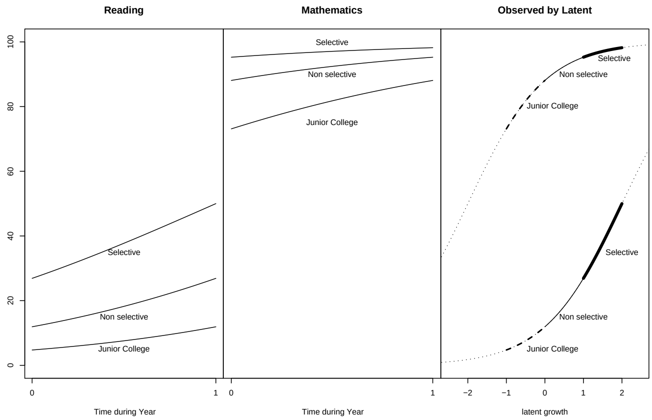
School	Pretest	Posttest	Change
Junior College	27	73	45
Non-selective university	73	95	22
Selective university	95	99	4

They concluded that the teaching at the junior college was far superior to that of the select university. What is wrong with this conclusion?

Effect of teaching, effect of students, or just scaling?



The effect of scaling upon the latent variable - observed variable relationship

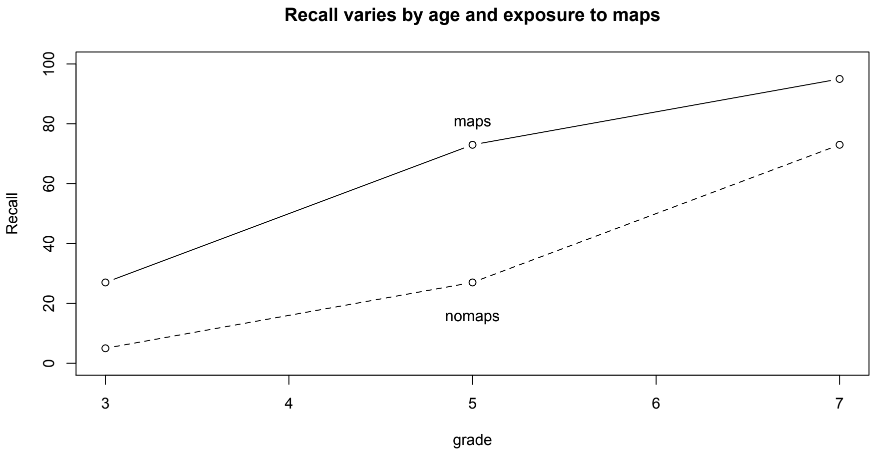


The problem of scaling is ubiquitous

1. A leading cognitive developmentalist believed that there is a critiical stage for learning spatial representations using maps. Children younger than this stage are not helped by maps, nor are children older than this stage.
2. He randomly assigned 3rd, 5th, and 7th grade students into two conditions (nested within grade), control, and map use. Performance was measures on a task of spatial recall (children were shown toys at particular locations in a set of rooms and then asked to find them again later.) Half the children were shown a map of the rooms before doing the task.
3. Their scores were

	No Map	Maps	Effect	
3rd grade	5	27	22	Too young
5th grade	27	73	46	Critical period
7th grade	73	95	22	Too old

Map use is most effective at a particular developmental stage



R code for the prior figure

R code

```
mapuse <- matrix(c( 3,5,27,5,27,73,7, 73,95),ncol=3,byrow=TRUE)
colnames(mapuse) <- c("grade","nomaps","maps")
rownames(mapuse) <- c("3rd","5th","7th")
maps.df <- data.frame(mapuse)
maps.df
with(maps.df,plot(maps~grade,ylab="Recall",ylim=c(0,100),
typ="b", main="Recall varies by age and exposure to maps"))
with(maps.df,points(nomaps~grade,ylab="Recall",
          ylim=c(0,100),typ="b",lty="dashed"))
> text(5,80,"maps")   #add line labels
> text(5,15,"nomaps")
```

```
    grade nomaps maps
3rd     3      5   27
5th     5     27   73
7th     7     73   95
```

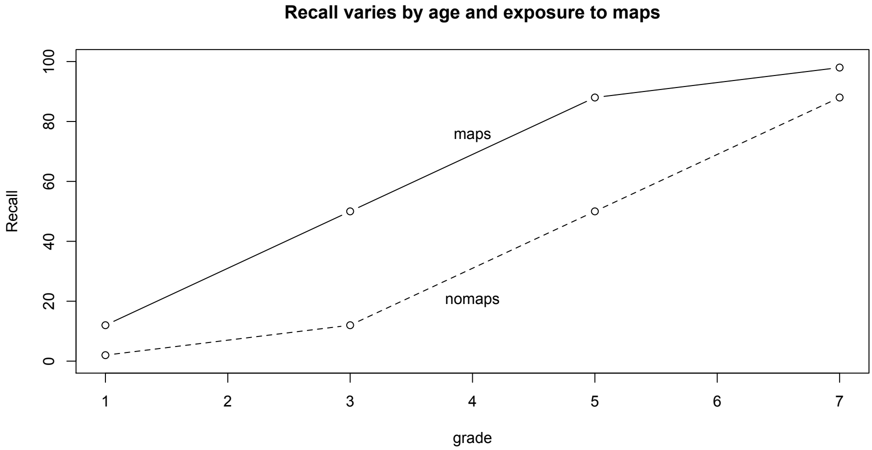
Yet another developmentalist

Another cognitive developmentalist believed that there is a critical stage but that it appears earlier than previously thought. Children younger than this stage are not helped by maps, nor are children older than this stage. He randomly assigned 1st, 3rd, 5th and 7th grade students into two conditions (nested within grade), control and mapa use. Performance was measured on a task of spatial recall (children were shown toys at particular locations in a set of rooms and then asked to find them again later. Half the children were shown a map of the room before doing the task.

The scores were

	No Map	Maps	Effect	
1st grade	2	12	10	Too young
3rd grade	12	50	38	
5th grade	50	88	38	Critical period
7th grade	88	98	10	Too old

A critical period in developmental?



R code for the prior figure

R code

```
mapuse <- matrix(c( 1,2,12,10,3,12,50,38,5,50,88,38,7,88,98,10),ncol=4,byrow=TR
colnames(mapuse) <- c("grade","nomaps","maps","Diff")
rownames(mapuse) <- c("1st" ,"3rd","5th","7th")
maps.df <- data.frame(mapuse)
maps.df
with(maps.df,plot(maps~grade,ylab="Recall",ylim=c(0,100),
typ="b", main="Recall varies by age and exposure to maps"))
with(maps.df,points(nomaps~grade,ylab="Recall",
          ylim=c(0,100),typ="b",lty="dashed"))
text(4,75,"maps")   #add line labels
text(4,20,"nomaps")
```

```
     grade nomaps maps
3rd      3      5   27
5th      5     27   73
7th      7     73   95
```

Traditional levels of measurement

Nominal Categories: X, Y, W, V

Ordinal Ranks ($X > Y > W > V$)

Interval Equal Differences ($X - Y > W - V$)

Ratio Equal intervals with a zero point ($X/Y > W/V$)

Types of scales and types of inference

1. Nominal allow us to say whether groups differ in frequency
2. Ordinal allows to compare rank orders of the data, is one score greater than another score. Any monotonic transformation will preserve rank order.
3. Interval is the claim that we can compare the magnitude of intervals. Only linear transformations will preserve interval information (i.e. we can add and subtract the numbers and preserve interval information. item Ratio scales preserve absolute magnitude differences.

Ordinal scales

1. Any monotonic transformation will preserve order
2. Inferences from observed to latent variable are restricted to rank orders
3. Statistics: Medians, Quartiles, Percentiles

Interval scales

1. Possible to infer the magnitude of differences between points on the latent variable given differences on the observed variable?X is as much greater than Y as Z is from W
2. Linear transformations preserve interval information
3. Allowable statistics: Means, Variances
4. Although our data are actually probably just ordinal, we tend to use interval assumptions.

Ratio Scales

1. Interval scales with a zero point
2. Possible to compare ratios of magnitudes (X is twice as long as Y)
3. Are there any psychological examples?

The search for an appropriate scale

1. Is today colder than yesterday? (ranks) Is the amount that today is colder than yesterday more than the amount that yesterday was colder than the day before? (intervals)
    - $50F - 39F < 68F - 50F$
    - $10C - 4C < 20C - 10C$
    - $283K - 277K < 293K - 283K$
2. How much colder is today than yesterday?
    - (Degree days as measure of energy use) is almost ratio
    - K as measure of molecular energy

Measurement confusions – arousal

1. Arousal is a fundamental concept in many psychological theories. It is thought to reflect basic levels of alertness and preparedness. Typical indices of arousal are measures of the amount of palmer sweating.
2. This may be indexed by the amount of electricity that is conducted by the fingertips.
3. Alternatively, it may be indexed (negatively) by the amount of skin resistance of the finger tips. The Galvanic Skin Response (GSR) reflects moment to moment changes, SC and SR reflect longer term, basal levels.
4. High skin conductance (low skin resistance) is thought to reflect high arousal.

Arousal and anxiety

1. Anxiety is thought to be related to arousal. The following data were collected by two different experimenters. One collected

	Resistance ,	conductance data.
low anxiety	1, 5	1, .2
high anxiety	2, 2	.5, .5

The means were therefore:

	Resistance ,	conductance data.
low anxiety	3	.6
high anxiety	2	.5,

2. That is, the low anxiety participants had higher skin resistance and thus were more relaxed, but they also had higher skin conductance, and thus were more aroused.
3. How can this be?

Multiple measures of dispersion

Range (highest - lowest) is sensitive to the number of observations, but is a very good way to detect errors in data entry.

MAD (Median Absolute Deviation from the Median) applied ordinal statistics to interval measures

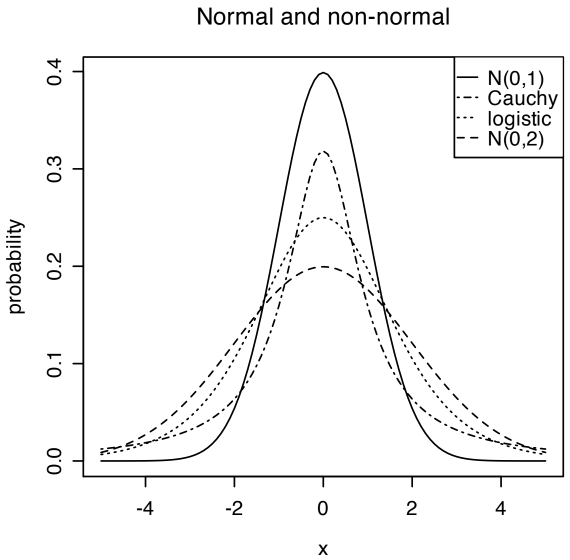
Variance (σ^2) is the Mean Square deviation (implies interval data)

Standard Deviation (σ) is the Root Mean Square deviation.

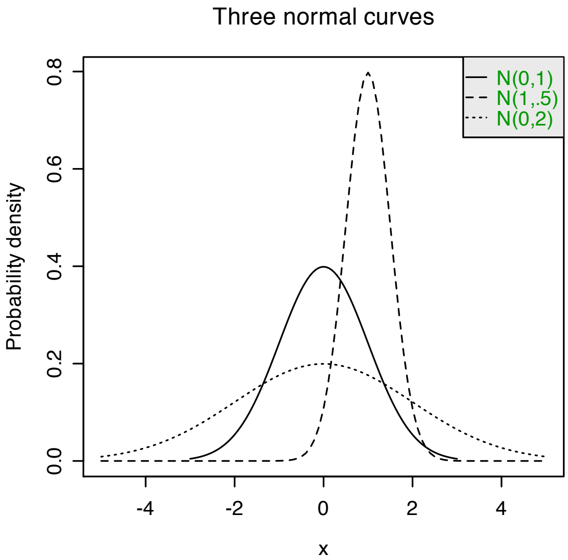
Coefficient of Variation $\frac{\sigma_x}{\mu_x}$

Average difference $\sigma_x\sqrt{2}$

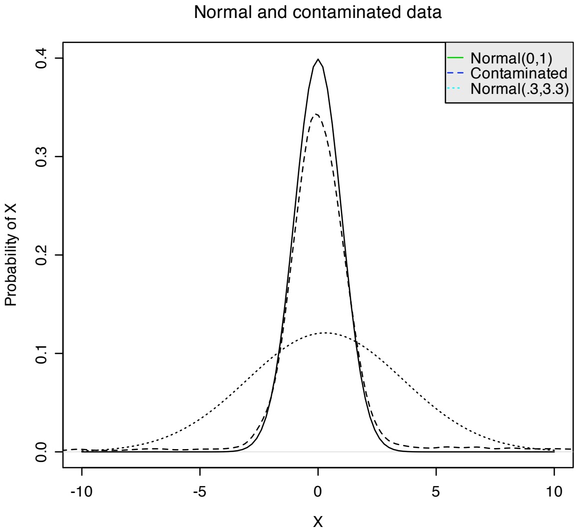
Normal and non-normal curves



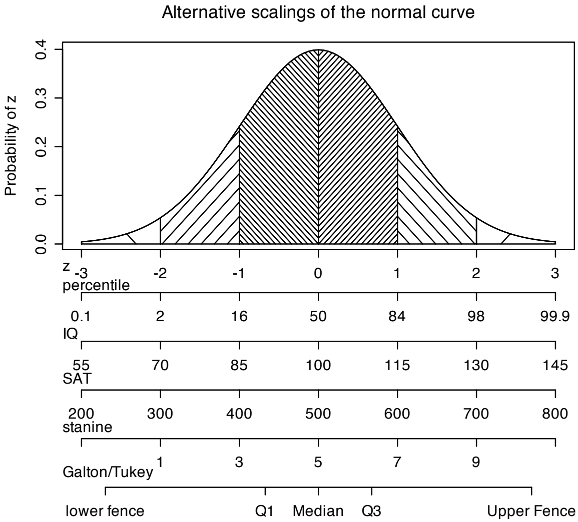
Three normal curves



Seriously contaminated data



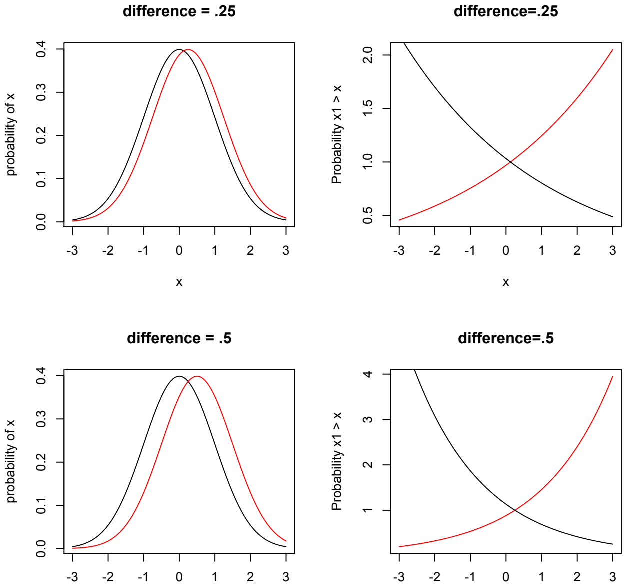
The normal curve and its frequent transforms



Decision making and the benefit of extreme selection ratios

1. Typical traits are approximated by a normal distribution.
2. Small differences in means or variances can lead to large differences in relative odds at the tails
3. Accuracy of decision/prediction is higher for extreme values.
4. Do we infer trait mean differences from observing differences of extreme values?

The effect of small mean differences at the tails of a distribution



The effect of small differences in variance at the tails of a distribution

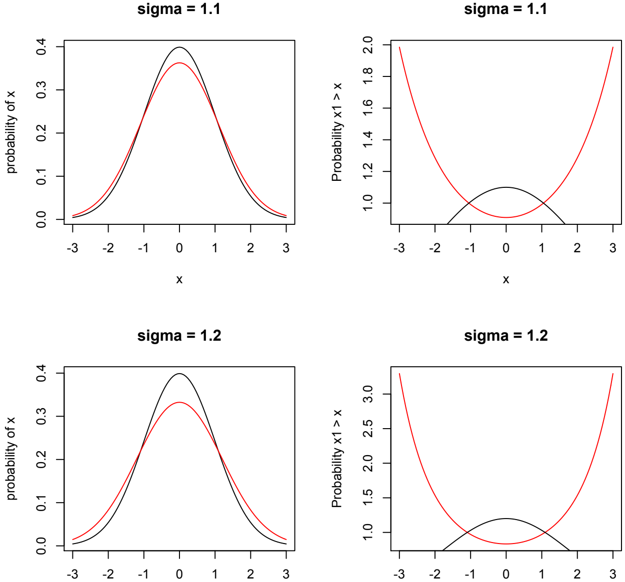
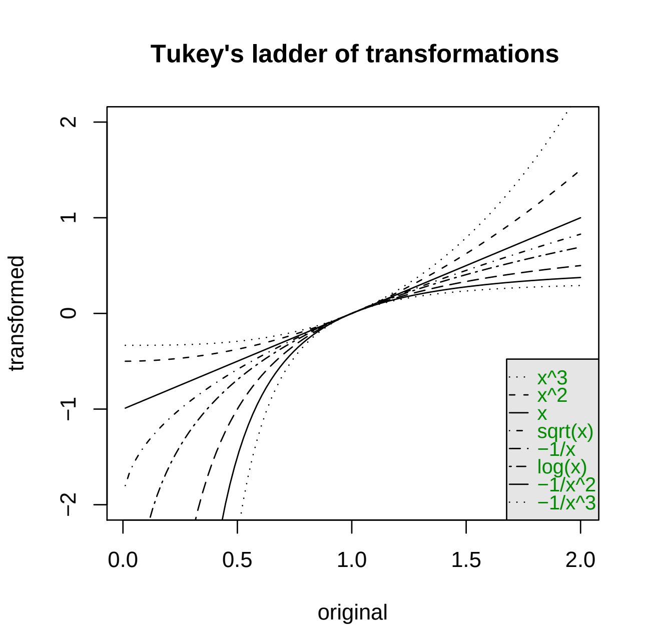


Table: Tukey's ladder of transformations. One goes up and down the ladder until the relationships desired are roughly linear or the distribution is less skewed. The effect of taking powers of the numbers is to emphasize the larger numbers, the effect of taking roots, logs, or reciprocals is to emphasize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


Tukey's ladder of transformations
transformed
original
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
x^3
x^2
x
sqrt(x)
−1/x
log(x)
−1/x^2
−1/x^3

The best scale is the one that works best

1. Money is linear but negatively accelerated with utility.
2. Perceived intensity is a log function of physical intensity.
3. Probabilty of being correct is a logistic or cumulative normal function of ability.
4. Energy used to heat a house is linear function of outdoor temperature.
5. Time to fall a particular distance varies as the square root of the distance ($s = at^2 <=> t = \sqrt{\frac{s}{a}}$)
6. Gravitational attraction varies as $1/distance^2$ ($F = G\frac{m_1 m_2}{d^2}$)
7. Hull speed of sailboat varies as square root of length of boat.
8. Sound intensity in db is log(observed/reference)
9. pH of solutions is -log(concentration of hydrogen ions)

Rossi, G. B. (2007). Measurability. *Measurement*, *40*(6), 545 – 562.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley Pub. Co.