Exploratory Data Analysis

2906 Multivariate Statistical Techniques
Session 1
January 24, 2019

Data Entry + Treatment

- **Take control**
  - Think through your output in advance
  - Study experimental software output files
  - Stay flexible (e.g., per trial vs. per person; flipped)
- **Minimize error**
  - Double check all steps from measurement to analysis
  - Stay raw (keep multiple copies when you recode)
  - Record value labels (and especially any changes)

Exploratory Data Analysis

- **Understand what you have in front of you**
- **Active and future-directed**; an attempt to understand the process that generates the data and improve future data collections.
- Sometimes meant to generate hypotheses that can be tested preliminarily in the same data set;
  - Then need replication in a different data set.

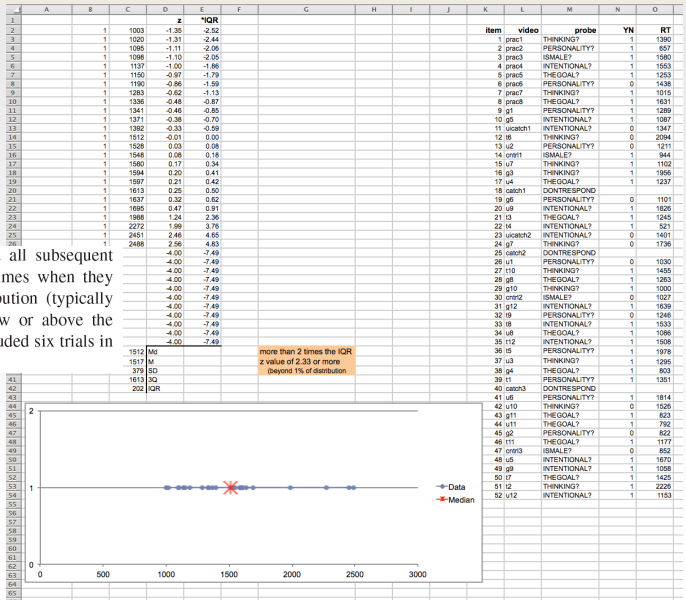
Data Screening

- **Accuracy of data entry**
  - Order manipulations?
  - Variables within their logical value range?
  - Label values (e.g., for GENDER: 1 = ‘female’ 2 = ‘male’)
- **Outliers**
  - Within an acceptable range + frequency? (e.g., z = 2.56)
  - Clarify *why* you have outliers
  - In a particular subgroup? (e.g., experimental condition)
  - w/s vs. b/s outliers
- **Multivariate outliers?**

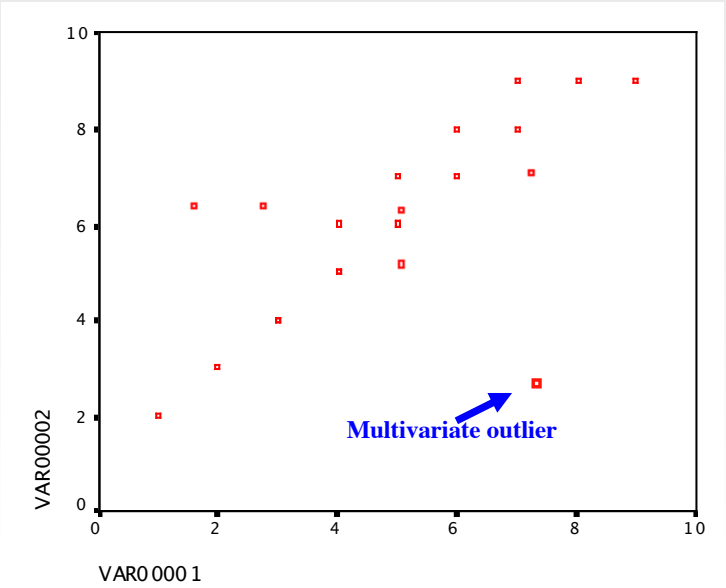
Detecting and Handling Outliers

Example from Malle & Holbrook (2012):

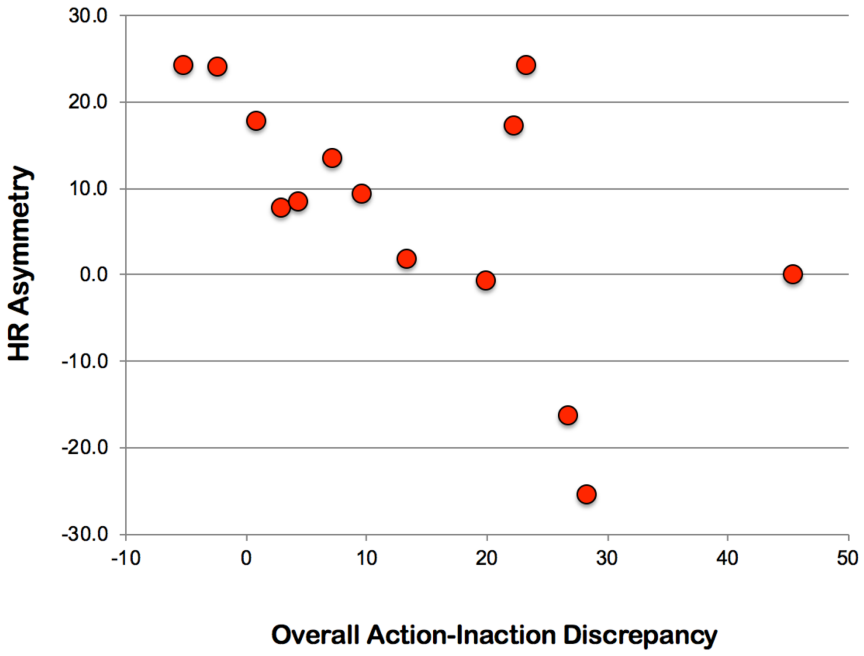
Invalid and corrected scores. In this and all subsequent studies, we excluded extremely short reaction times when they were located far outside the overall data distribution (typically below 400 ms) or 2.5 standard deviations below or above the particular participant's mean. In Study 1, we excluded six trials in this way (<1% out of 1,296).

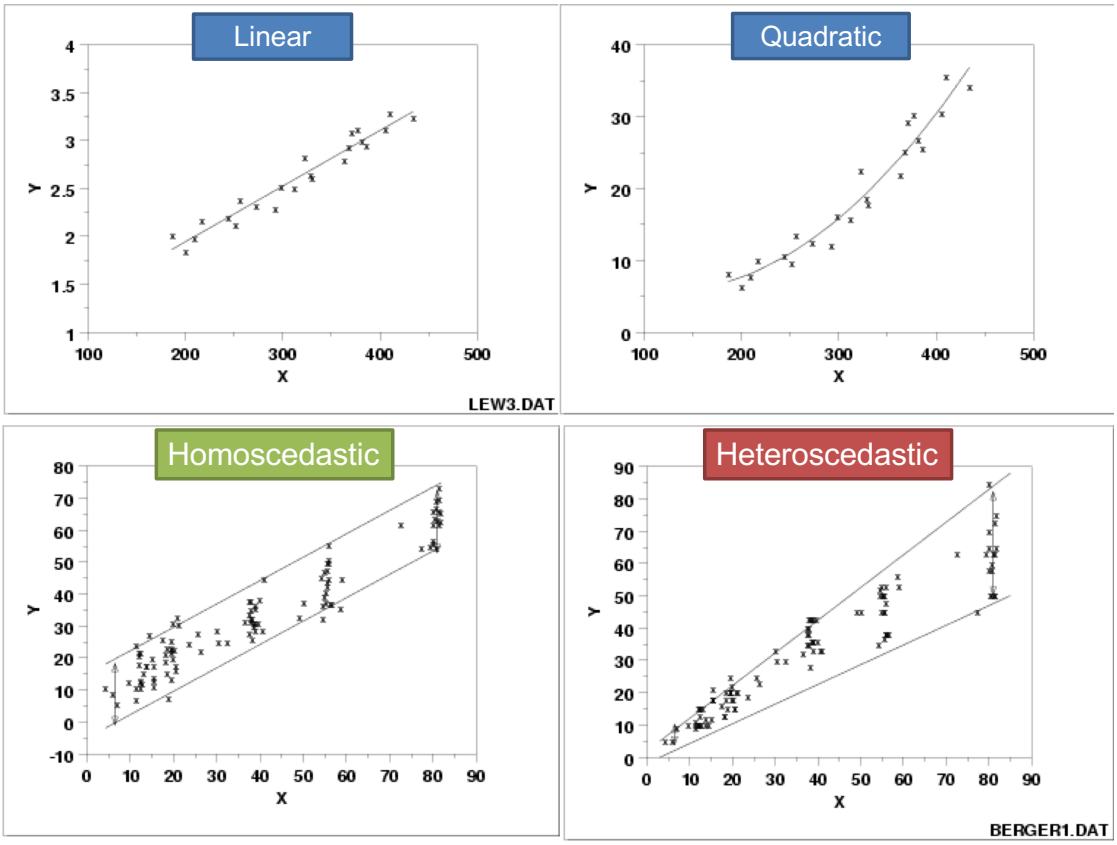


Scatterplot



Excel


30.0
20.0
10.0
0.0
-10.0
-20.0
-30.0
HR Asymmetry
-10
0
10
20
30
40
50
Overall Action-Inaction Discrepancy


Linear
Y
X
LEW3.DAT
Quadratic
Y
X
Homoscedastic
Y
X
Heteroscedastic
Y
X
BERGER1.DAT

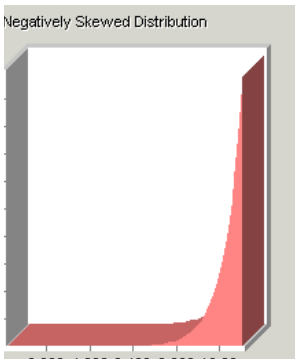
Distributional Properties

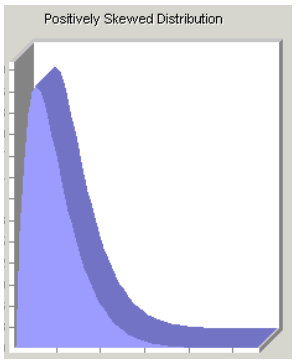
Mean: $\frac{1}{N}\sum_{1}^{N}(X_i)^1$

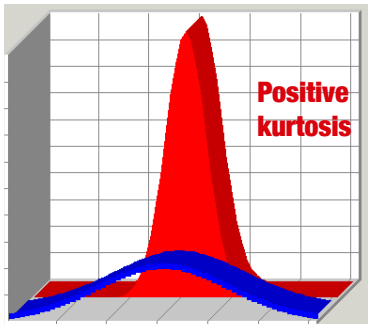
Variance: $\frac{1}{N-1}\sum_{1}^{N}(X_i-\bar{X})^2$

Skewness: $\frac{1}{N}\sum_{1}^{N}\left(\frac{X_i-\bar{X}}{s}\right)^3$

Kurtosis: $\frac{1}{N}\sum_{1}^{N}\left(\frac{X_i-\bar{X}}{s}\right)^4-3$







Negative kurtosis

Distributional Properties

```
EXAMINE
  VARIABLES=G_GOA_RTY
  /PLOT  STEMLEAF NPPLOT
  /STATISTICS EXTREME (5)
  /CINTERVAL 95.
```

IQR (Inter-Quartile Range) =$_{df}$ value at 75th – value at 25th percentile
Extremes =$_{df}$ values that lie 2*IQR outside the 25th/75th percentile
Outliers =$_{df}$ values that lie 1.5*IQR outside the 25th/75th percentile, but less then 2*IQR

Descriptives

		Statistic	Std. Error
G_GOA_RTY Mean		1573.0429	68.75588
95% Confidence Interval for Mean	Lower Bound	1433.3141	
	Upper Bound	1712.7716	
5% Trimmed Mean		1553.9841	
Median		1528.6667	
Variance		165457.992	
Std. Deviation		406.76528	
Minimum		973.00	
Maximum		2516.00	
Range		1543.00	
Interquartile Range	(IQR)	497.33	
Skewness		.699	.398
Kurtosis		.202	.778

Extreme Values

		Case Number	Value
G_GOA_RTY	Highest 1	20	2516.00
	2	13	2513.33
	3	32	2325.33
	4	9	2201.50
	5	7	2124.67
	Lowest 1	23	973.00
	2	12	976.00
	3	11	988.33
	4	19	1099.33
	5	2	1128.00

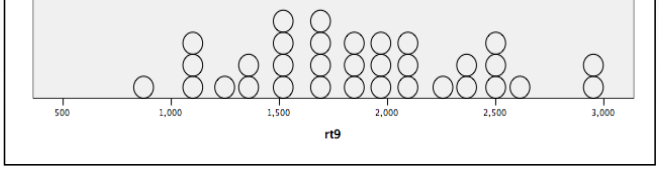
G_GOA_RTY Stem-and-Leaf Plot

```
 Frequency     Stem &  Leaf

     3.00        0 .  999
    13.00        1 .  0111223333444
    14.00        1 .  55555666777788
     3.00        2 .  123
     2.00 Extremes    (>=2513)

Stem width:   1000.00
Each leaf:       1 case(s)
```

Dotplot

```
XGRAPH CHART=yvars BY xvars BY zvars

  /BIN START={AUTO**} SIZE={AUTO**   }
            {x     }      {WIDTH (x)}
                          {COUNT (n)}

  /DISPLAY DOT={ASYMMETRIC**}
              {SYMMETRIC   }
              {FLAT        }

  /DISTRIBUTION TYPE=NORMAL

  /COORDINATE SPLIT={NO**}
                    {YES }

  /ERRORBAR {CI     {(95)}}
                    {(n) }
            {STDDEV {(2) }}
                    {(n) }
            {SE     {(2) }}
                    {(n) }

  /MISSING USE={LISTWISE**  } REPORT={NO**}
              {VARIABLEWISE}        {YES }

  /PANEL COLVAR=varlist COLOP={CROSS**} ROWVAR=varlist ROWOP={CROSS**}
                              {NEST   }                     {NEST   }
```

```
XGRAPH
  CHART=[POINT] BY rt9[s]
  /DISPLAY DOT=ASYMMETRIC.
```



Stem-and-Leaf Plot

```
Frequency     Stem &  Leaf

     5.00      -5 .  00000
     5.00      -4 .  00000
    15.00      -3 .  000000000000000
     7.00      -2 .  0000000
     9.00      -1 .  000000000
    46.00       0 .  0000000000000000000000000000000000000000000000
     6.00       1 .  000000
    17.00       2 .  00000000000000000
     9.00       3 .  000000000
     6.00       4 .  000000

     2.00       5 .  00

Stem width:     1
Each leaf:        1 case(s)
```

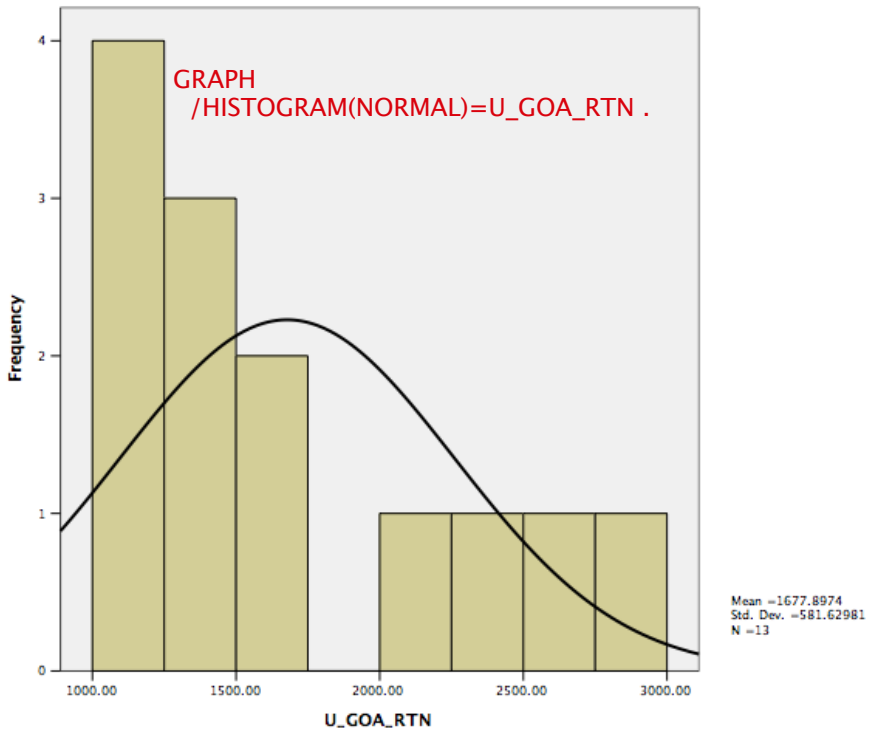
Histogram

```
FREQUENCIES VARIABLES=U_GOA_RTN
  /HISTOGRAM  NORMAL.
```

U_GOA_RTN

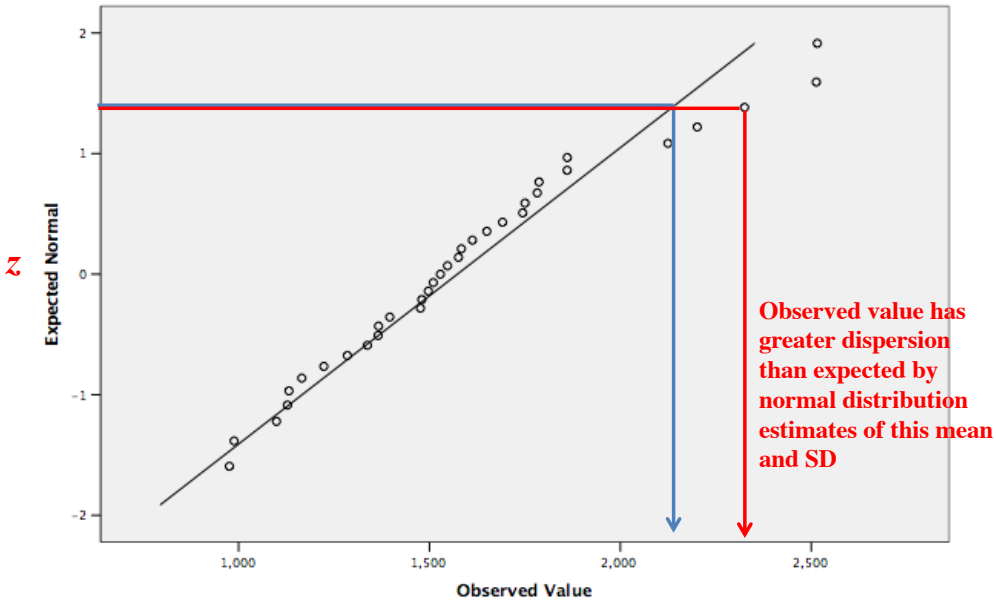
		Freq	Percent	Valid Percent	Cumulative Percent
Valid	1087.00	1	2.8	7.7	7.7
	1141.00	1	2.8	7.7	15.4
	1180.67	1	2.8	7.7	23.1
	1228.00	1	2.8	7.7	30.8
	1275.00	1	2.8	7.7	38.5
	1343.00	1	2.8	7.7	46.2
	1449.00	1	2.8	7.7	53.8
	1552.50	1	2.8	7.7	61.5
	1703.50	1	2.8	7.7	69.2
	2238.00	1	2.8	7.7	76.9
	2336.00	1	2.8	7.7	84.6
	2511.00	1	2.8	7.7	92.3
	2768.00	1	2.8	7.7	100.0
	Total	13	36.1	100.0	
Missing	System	23	63.9		
Total		36	100.0		

```
GRAPH
  /HISTOGRAM(NORMAL)=U_GOA_RTN .
```

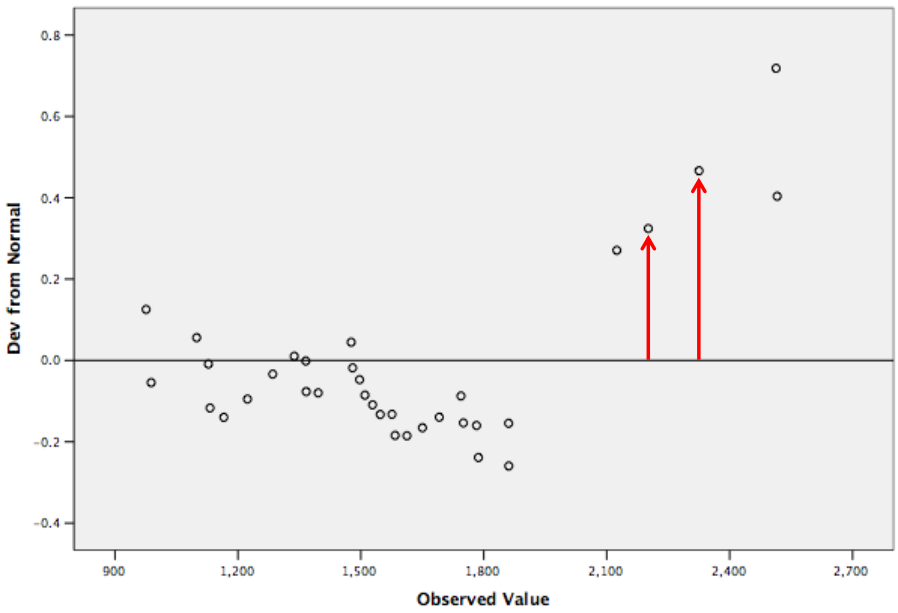


```
EXAMINE
  VARIABLES=G_GOA_RTY
  /PLOT  STEMLEAF NPPLOT
  /STATISTICS EXTREME (5)
  /CINTERVAL 95.
```

Normal P-P Plot of G_GOA_RTY



Detrended Normal Q-Q Plot of G_GOA_RTY



Dealing with Skewness and Kurtosis

- **Deviations from normality assumption**
  - Significance test: $parameter/s_e = +/- 2 \Rightarrow p < .05$
  - If significant, think about its meaning
  - Consider aggregation, transformation, non-parametrics
    - *aggregation* ~ averaging of variables
    - helpful because skew and kurtosis of individual variables can cancel out
    - aggregated variables must be on the same scale
- **In SPSS, aggregating can be done this way:**

```
COMPUTE newvar = mean.2(oldvar1, oldvar2, oldvar3).
```

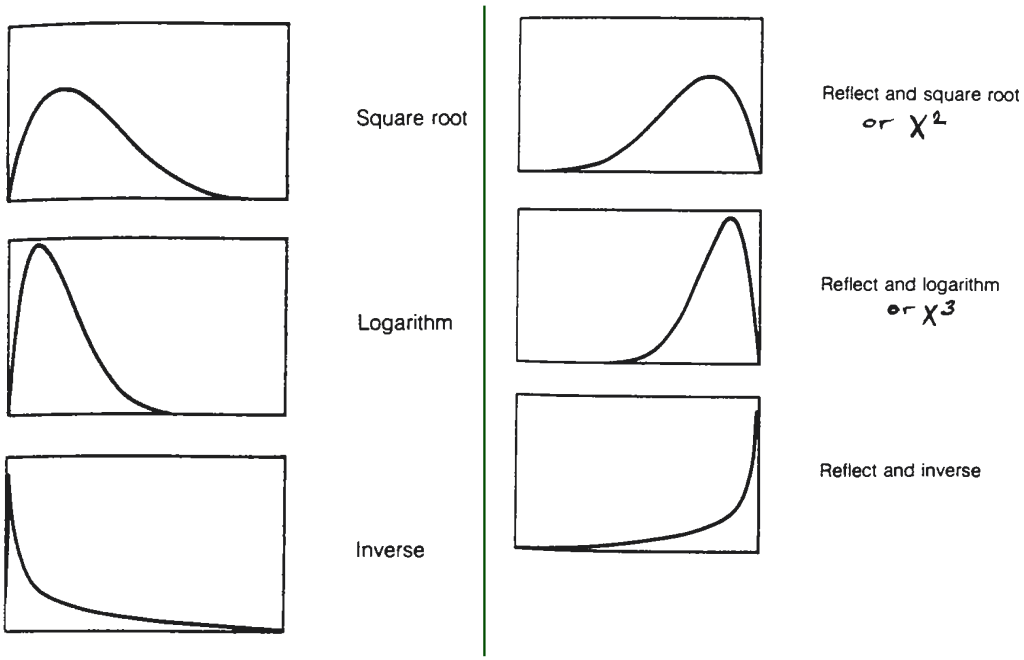
Transformations?

TABLE 4.2 CONTROL LANGUAGE FOR COMMON DATA TRANSFORMATIONS

	BMDP /TRANSFORM	SPSS[x] COMPUTE	SAS DATA Procedure	SYSTAT Data Module
Moderate positive skewness	NEWX = SQRT(X).	NEWX = SQRT(X)	NEWX = SQRT(X)	LET NEWX = SQR(X)
Substantial positive skewness	NEWX = LOG(X).	NEWX = LG10(X)	NEWX = LOG10(X)	LET NEWX = LOG(X)/LOG(10)
With zero	NEWX = LOG(X + C).	NEWX = LG10(X + C)	NEWX = LOG10(X + C)	LET NEWX = LOG(X + C)/LOG(10)
Severe positive skewness, L-shaped	NEWX = 1/X.	NEWX = 1/X	NEWX = 1/X	LET NEWX = 1/X
With zero	NEWX = 1/(X + C).	NEWX = 1/(X + C)	NEWX = 1/(X + C)	LET NEWX = 1/(X + C)
Moderate negative skewness	NEWX = SQRT(K − X).	NEWX = SQRT(K − X)	NEWX = SQRT(K − X)	LET NEWX = SQR(K − X)
Substantial negative skewness	NEWX = LOG(K − X).	NEWX = LG10(K − X)	NEWX = LOG10(K − X)	LET NEWX = LOG(K − X)/LOG(10)
Severe negative skewness, J-shaped	NEWX = 1/(K − X).	NEWX = 1/(K − X)	NEWX = 1/(K − X)	LET NEWX = 1/(K − X)

C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

From Tabachnick & Fidell and *Handout* 1



Missing Data

- **Are they random or meaningful?** (e.g., part of experimental group, indicating misunderstanding)
- **Ways of dealing with missing data**
  - Exclude the entire case or a specific value.
  - Substitution (e.g., by M/Md of the entire sample; by M/Md of the subgroup of which the case is a member; by regressing).
- **Golden Rule: Compare analyses with or without outliers, before and after transformation, etc.**
  - Create dummy variable for cases with substituted values.

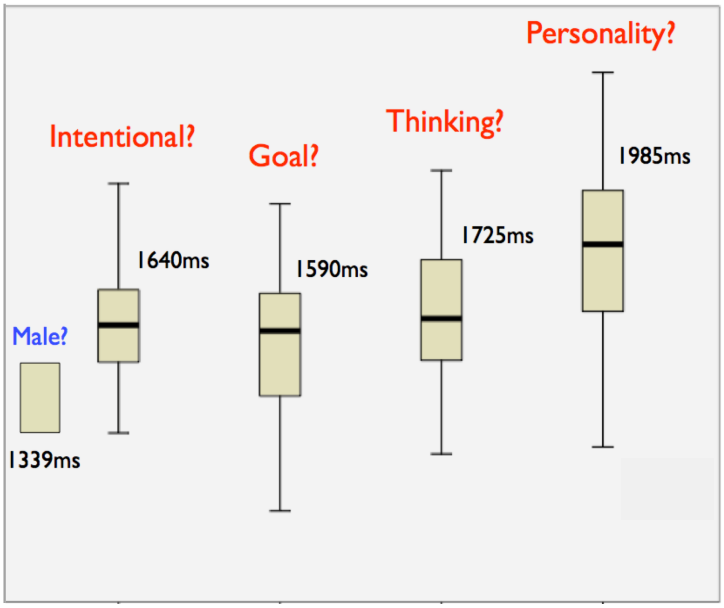
Example from Malle & Holbrook (2012):

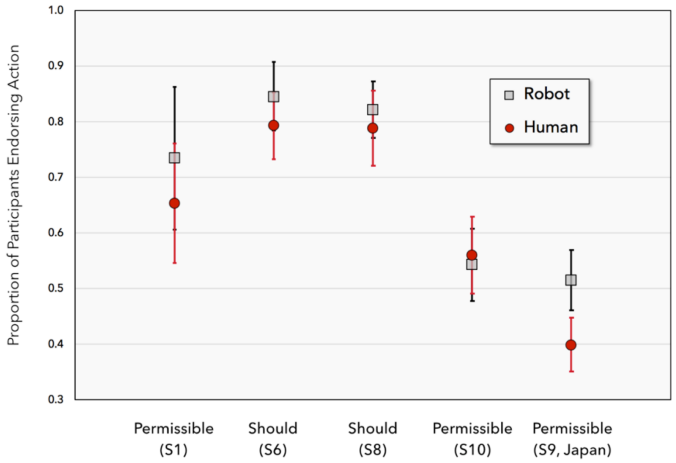
Missing value replacements. Some replacements for missing reaction times were necessary when individuals who had valid reaction times for most of the 12 design cells would have been omitted from the within-subject ANOVA because of one or more cells in which they had not provided any *Yes* responses and, therefore, had no score of inference speed. To retain 36 participants, we used cell-based sample means to replace 57 missing values within the matrix of 432 potential values (36 participants × 12 cells). To retain the same 36 participants in the analysis of inference likelihood, four entries were mean replaced. It should be noted that these replacements do not change the cell means of the design but somewhat lower their standard deviations.[5] (See supplementary material for all studies' means and standard deviations.)

BOXPLOT. *Vertical boxplot*. The boundaries of the box are Tukey's hinges. The median is identified by a line inside the box. The length of the box is the interquartile range (IQR) computed from Tukey's hinges. Values more than three IQR's from the end of a box are labeled as extreme, denoted with an asterisk (*). Values more than 1.5 IQR's but less than 3 IQR's from the end of the box are labeled as outliers (o).

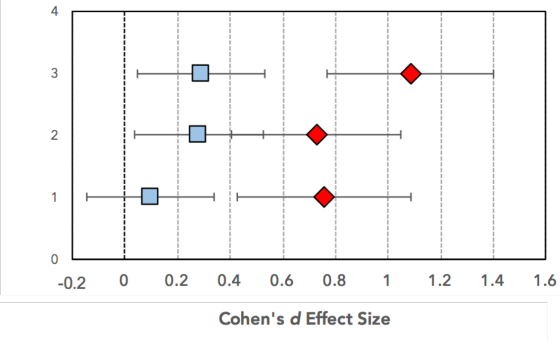
Box height = Inter-quartile range (IQR) = range between 25th and 75th percentile

The **arms** designate the range of actual values that are *not yet* outliers. Because outliers are defined as 1.5*IQR beyond the 25th/75th percentile (i.e., beyond the vertical box edge), the arm range is *up to* 1.5*IQR, but it can be less if there are no values that go all the way to 1.5*IQR.

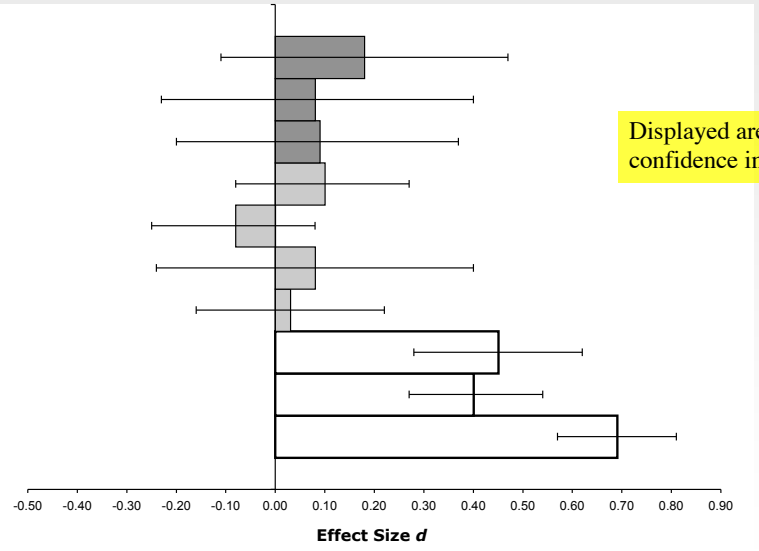




Plots of Means+ SEs or Means + CIs



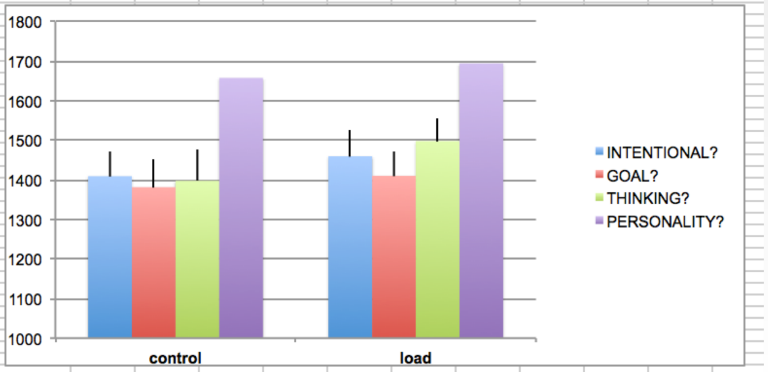
Expanded Histogram



Displayed are 95% confidence intervals

Getting Those Error Bars

	control	load	control	load	control	load
	M	M	SD	SD	**se**	**se**
INTENTIONAL?	**1409**	**1459**	394	430	63	69
GOAL?	**1381**	**1410**	429	380	69	61
THINKING?	**1397**	**1498**	498	370	80	59
PERSONALITY?	**1657**	**1694**	467	456	75	73
N =	39					



Other Helpful Commands for EDA

- **CROSSTABS**
- **CORRELATION (PARTIAL CORR)**
- **MEANS TABLES**

Graphing software options:

- SPSS, with editing
- Excel, with patience (and Keynote or PPT on top of it)
- JMP, R, many others (check what's available on Brown's software server)