

# Psych 524

Andrew Ainsworth

Data Screening 2

# Transformation

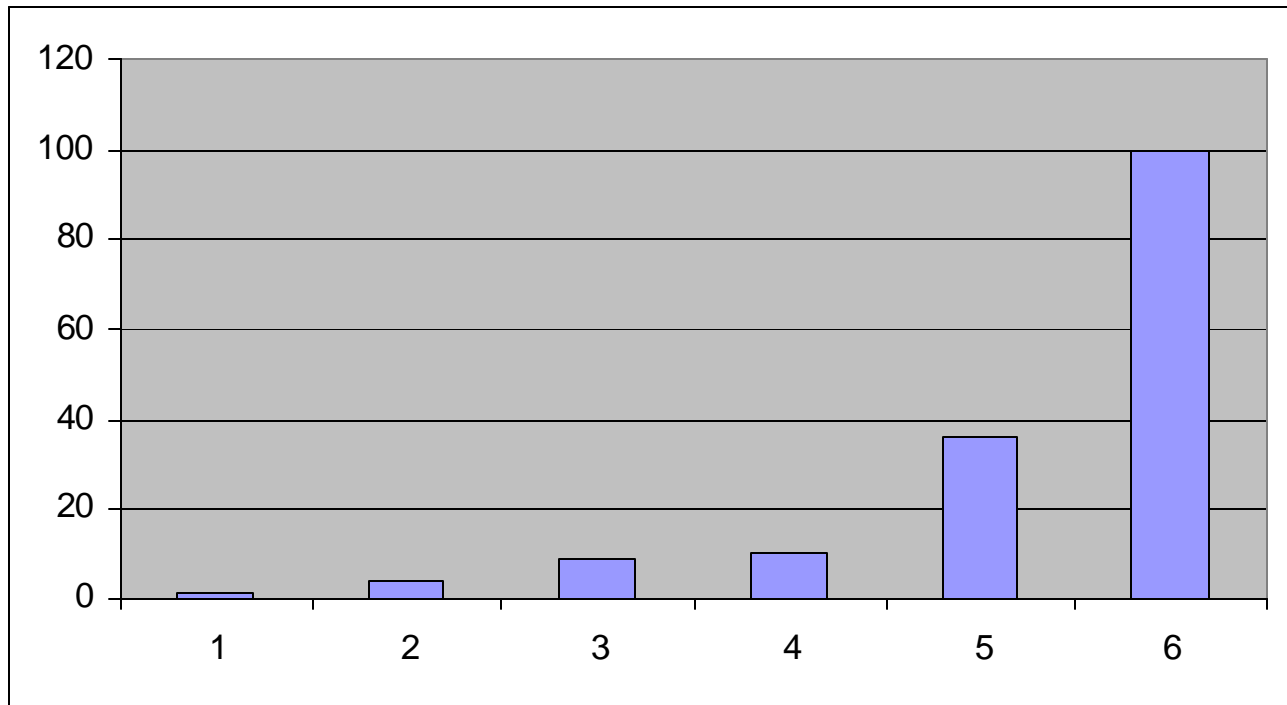
- allows for the correction of non-normality caused by skewness, kurtosis, or other problems (lack of linearity)
- Shouldn't be done if values represent meaningful scale
- Square root – moderate violations, LOG – severe, and inverse for severe violation

# Transformation

- For positively skewed data square root and log keep data in the original order but bring in the spread, while inverse flips the order of the data
- For negatively skewed data the reverse is true; without adjustment square root and log reverse order and inverse keeps the same order

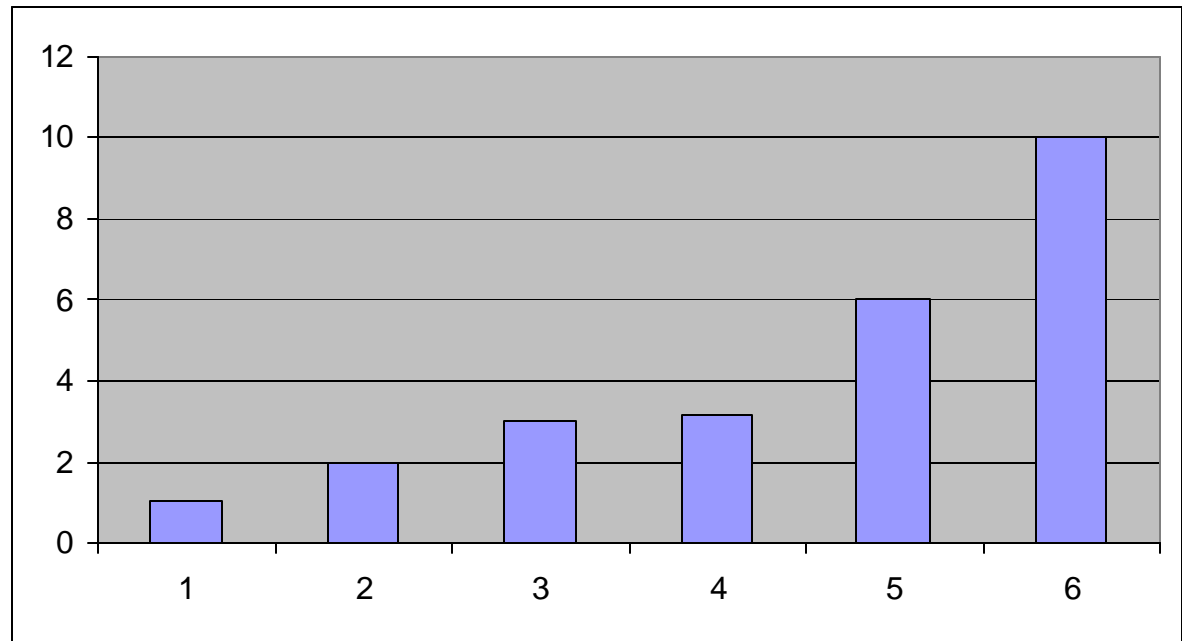
# Original Data

1  
4  
9  
10  
36  
100



# Square Root Transform

1  
2  
3  
3.16227  
8  
6  
10



# Log Transform

0

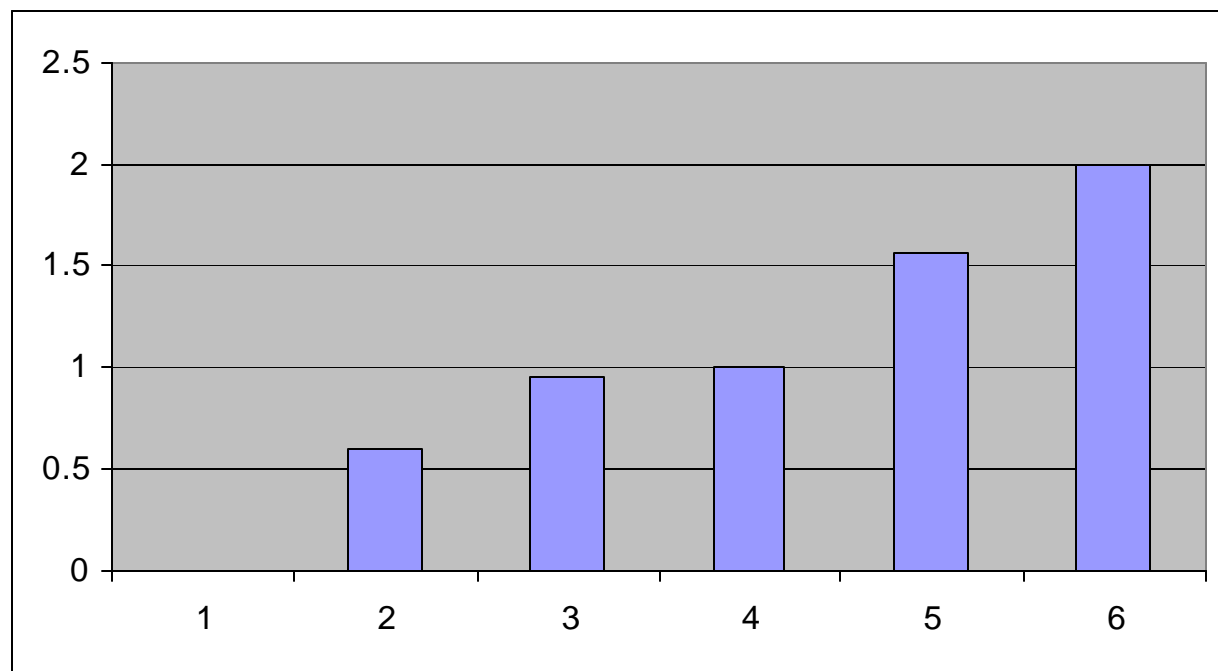
0.60206

0.954243

1

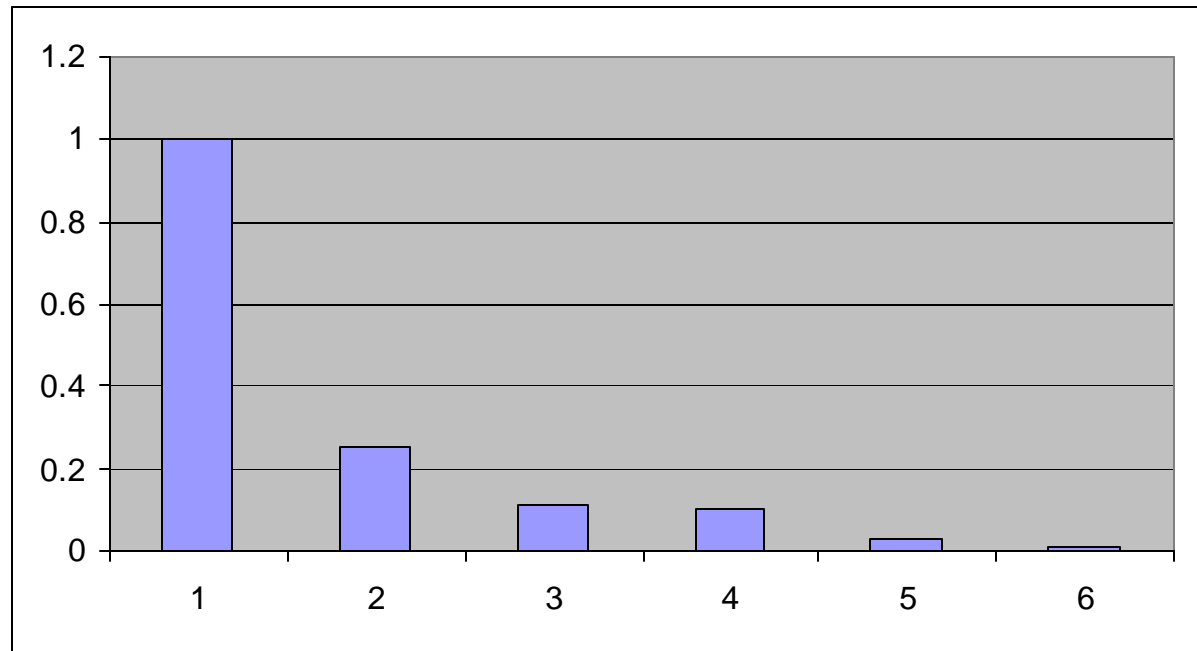
1.556303

2



# Inverse Transform

1  
0.25  
0.111111  
0.1  
0.027778  
0.01



# Dealing with Missing Data

- The default in many programs (e.g. SPSS) is to do a complete cases analysis (listwise deletion)
  - simple and easy
  - but many concerns (e.g. percent of missing, pattern of missing) because doing complete cases analysis assumes missing at random



# Missing Completely at Random

- MCAR – means that the patterns of missing on any one variable is not related to another variable.
  - Example of non-MCAR: Measures of IQ and Income – subjects below a certain level of IQ (e.g. cutoff for “retardation”) may not have any income because they are under guardian care, so they leave the income variable blank

# Complete Cases Analysis

	$x_1$	$x_2$	$x_3$
$S_1$	$x_{11}$	$x_{12}$	$x_{13}$
$S_2$	$x_{21}$	—	$x_{23}$
$S_3$	—	$x_{32}$	$x_{33}$



Only this case  
is used

# Missing Value Correlation Matrix

- Create a correlation matrix using complete cases for each pair of variables
  - For each correlation estimate you are using the most data possible
  - But each estimate is based on a different number of subjects
  - “Delete cases pairwise” in SPSS

# Dealing with Missing Data

- Imputation (replacing missing data)
  - Variable Mean insert – doesn't effect the mean estimation be restricts the variance
  - Group mean insert – if you have grouped data then replace missing values with the mean of the group the subject belonged to.
  - Regression – predicting a subject's missing value on one variable by scores on other variables. Could be used iteratively. Iterative means the process is repeated until the estimated value stabilizes

# Dealing with Missing Data

- Imputation (replacing missing data)
  - Estimation maximization (EM) algorithm – this is a maximum likelihood iterative estimation method.
  - Multiple Imputation – use multiple methods from above (and others in the book) and compute average estimate.
    - This is nice because it also gives you a standard error estimate for the estimation