# Transformations & Data Cleaning

- Linear & non-linear transformations
- 2-kinds of Z-scores
- Identifying Outliers & Influential Cases
- Univariate Outlier Analyses -- trimming vs. Winsorizing
- Outlier Analyses for r & ANOVA
- Outlier Analyses for Multiple Regression

## Transformations

### Linear Transformations

- transformations that involve only +, - , * and /
- used to "re-express" data to enhance communication
  e.g., using % correct instead of # correct
- do not causes changes in NHST, CI or Effect Size results
  - r, t & F results will be same before and after transformation
  - but if doing t/F, be sure to transform all scores around the overall mean, not to transform each group's scores around their own mean

### Nonlinear Transformations

- transformations involving other operations (e.g., ^2, √ & log)
- used to "symmetrize" data distributions to improve their fit to assumptions of statistical analysis (i.e., Normal Distribution assumption of r, t & F)
- may change r, t & F results -- hope is that transformed results will be more accurate, because data will better fit assumptions

## Effect of Linear Transformations on the Mean and Std

We can anticipate the effect on the mean and std of adding, subtracting, multiplying or dividing each value in a data set by a constant.

| operation | effect on mean | effect on std |
|---|---|---|
| + ? | Mean + ? | No change |
| - ? | Mean - ? | No Change |
| * ? | Mean * ? | Std * ? |
| / ? | Mean / ? | Std / ? |

Commonly Used Linear Transformations & one to watch for ...

Z-score $\quad Z = \dfrac{x - mean}{std} \quad$ (m = 0, s = 1) $\quad$ linear

T-score $\quad T = (Z * 10) + 50 \quad$ (m = 50, s = 10) $\quad$ linear

Standard Test $\quad S = (Z * 100) + 500 \quad$ (m = 500, s = 100) linear

y' $\qquad y' = (b * x) + a \qquad$ (m=m of y, s ≈ s of y) linear

% $\qquad p\_y = (y / max) * 100 \qquad\qquad$ linear

change score $\quad \Delta =$ "post" score - "pre" score $\qquad$ nonlinear

---

A quick word about Z-scores…
There are "two kinds" based on ...
   • mean and std of that sample (M s)
   • mean and std of the represented population ($\mu$ $\sigma$)

$$Z_{sample} = \dfrac{X - M}{s} \qquad\qquad Z_{pop} = \dfrac{X - \mu}{\sigma}$$

• mean of Z-scores always 0

• std of Z-scores always 1

• translates relative scores into easily interpreted values

• mean > 0 if sample "better" than pop

• mean < 0 if sample "poorer" than pop

• std < 1 if sample s < $\sigma$

• std > 1 if sample s > $\sigma$

• provides ready comparison of sample mean and std to "population values"

Similarly, you can compose T and Standard scores using $\mu$ and $\sigma$ to get sample-population comparisons using these score-types.

---

Non-linear transformations  -- to "symmetrize" data distributions
the "transformation needed" is related to the extent & direction of skewing

Skewness $\qquad$ Suggested Transformation

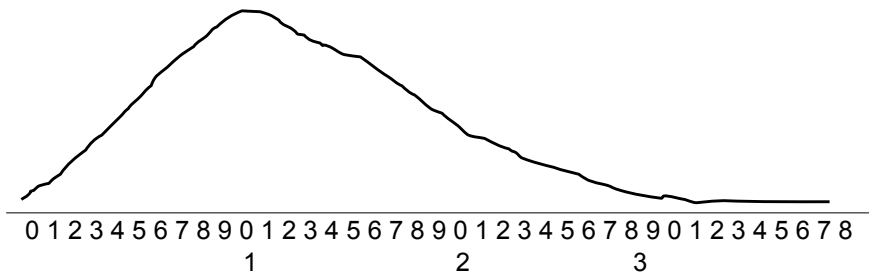  <  +/- .80 $\qquad$ unlikely to disrupt common statistical analyses

**Most transformations are directly applied to positively skewed distributions.**

  +.8  to +1.5 $\quad$ square root transformation (0 & positive values only)

  +1.5  to +3.0 $\quad$ $\log_{10}$ transformation  (positive values only)

  +3.0 or greater $\quad$ inverse transformation 1 / #  (positive values only)

Transformation of negatively skewed distributions first require "reflection", which involves subtracting all values from the largest value+1.
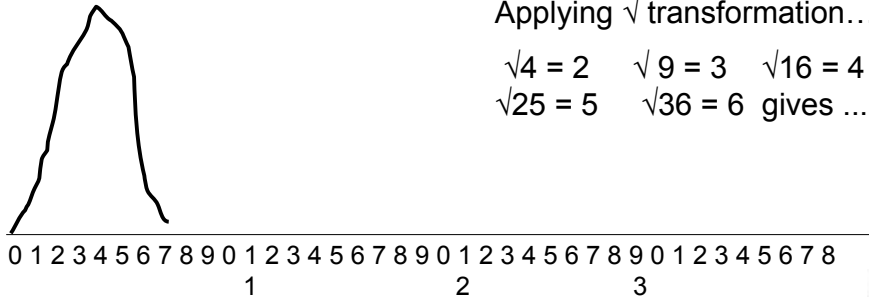
  - .80 to - 1.5 $\quad$ reflect & square root transformation

  -1.5  to -3.0 $\quad$ reflect & $\log_{10}$ transformation  (positive values only)

  -3.0 or greater $\quad$ reflect & inverse transformation 1 / #  (positive values only)

How "symmetrizing" works…



0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
                 1              2            3

Applying $\sqrt{}$ transformation…

$\sqrt{4} = 2$    $\sqrt{9} = 3$    $\sqrt{16} = 4$
$\sqrt{25} = 5$    $\sqrt{36} = 6$    gives ...

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
                 1              2            3

## Influential Cases & Outlier Analysis

The purpose of a sample is to represent a specific population

• the better the sample represents the population, the more accurate will be the inferential statistics and the results of any inferential statistical tests

• sample members are useful only to the extent that they aid the representation

• influential cases are sample members with "extreme" or "non-representative" scores that influence the inferential stats

• outliers are cases with values that are "too extreme" to have come from the same population as the rest of the cases

e.g., the ages of the college sample were   21,  62, 22,  19 &  20
    the "62" is an influential case
      -- will radically increase the mean and std of this sample
      -- is "too large" to have come from the same pop as rest

## How outliers influence univariate statistics

• outliers can lead to too-high, too-low or nearly correct estimates of the population mean, depending upon the number and location of the outliers (asymmetrical vs. symmetrical patterns)

• outliers always lead to overestimates of the population std
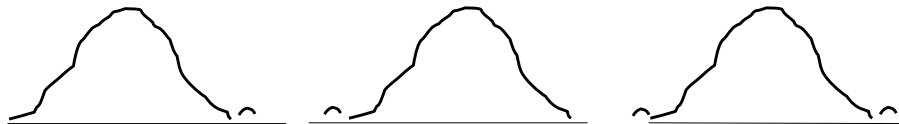


Mean estimate is "too high" & std is overestimated

Mean estimate is "too low" & std is overestimated

Mean estimate is "right" & std is overestimated

Identifying Outliers for removal

The preferred technique is to be able to identify participants who are not likely to be members of the population of interest.

However, often the only indication we have that a participant doesn't come from the population of interest is that they have an "outlying score".

So, we will operate under the assumption that "outlying scores" mean that:

1) a participant is probably not a member of the target population (or that something "bad" happened to their score)

2) if included the data as is would produce a biased estimate of the population value (e.g., mean, std, r, F, etc.) and so, should not be included in the data analysis.

Key -- application of the approach must be "hypothesis blind"

---

Statistical Identification of Outliers

One common procedure was to identify as outliers any data point with a Z-value greater than 2.0 (2.5 & 3.0 were also common suggestions)

• this means that some values that really do belong to the distribution were incorrectly identified as outliers, but simulation and real-data research suggested that statistical estimates were better and more replicable when these procedures were used.

• one problem with this approach is that outliers influence the std (making it larger than it should be) → leading us to miss outliers (yep – outliers can "hide themselves" when this approach is used)

The most common approach used currently is to base outlier analyses on rank-order statistics, which are much less influenced by the outliers they are trying to identify.

---

Outlier Identification formulas

The formula is very simple (easier using SPSS)…

   Lower bound value = Lower hinge – 1.5 (hinge spread)

   Upper bound value = Upper hinge + 1.5 (hinge spread)

Any data value smaller than the lower bound or larger than the upper bound is deemed an outlier – a value that probably doesn't come from the same population as the other values in the sample.

What we do with identified outliers

• Trimming – eliminating the data value from the analyses
   • leads to loss of data, statistical power, etc.)

• Winsorizing – recode the data value to the "nearest acceptable" value (upper or lower bound value)

   • large/small values are still present, but won't unduly influence statistical estimates & maintains sample size

## Outlier Analysis & Nonlinear Transformations – What Order?

• are both used to help "prepare" data to more closely conform to the assumptions of the statistical models we'll use
• so a natural question is "which to do first" ???
• not surprising "it depends" ...



These data would have outliers and substantial + skewing

the skewing probably "produced" the outliers -- so transform first & recheck for outliers



These data would also have outliers and substantial + skewing

the outliers probably "produced" the skewing -- so remove outliers first & recheck for skewness

Certainly it isn't always this clear, but usually one order works better than the other -- nicer distribution and fewer "outliers".
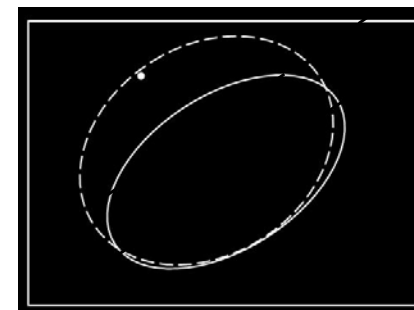
## Applying outlier analyses to correlation & regression

• univariate outliers can lead to over- or underestimates of r & b

• there are also "bivariate outliers" – data points that are outliers for neither "X" nor "Y", but change the shape of the envelope, and so the estimate of r, b and/or a

•this applies when the predictor is binary, as well

"dual univariate outlier" -- doesn't bias b or a, but r is overestimated
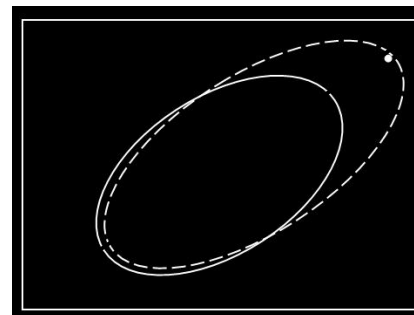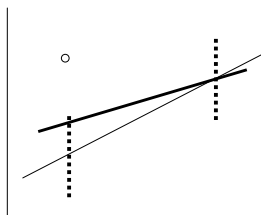
"bivarite outlier" -- biases a, but not b and r is underestimated



## Applying outlier analyses to binary predictors

• outliers can lead to too-high or too-low estimates of each group mean
• and so, can lead to poor estimations of r, b and a
• it is important to do outlier analysis of Y for each value of predictor



An extreme example – the sort that results from miscoding of the binary variable
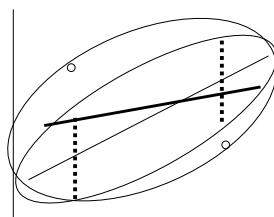
Both of these are "bivariate outliers" – neither would be revealed by univariate outlier analyses

## Applying outlier analyses to multiple regression

• remember that R and the bs are influenced by the criterion-predictor correlations (sometimes called "validities") and the predictor-predictor correlations (collinearities)

• so, any outlier that influences any correlation can influence the multiple regression model

• in addition to univariate & bivariate outlier analyses (including all the predictor-predictor variables) there are also multivariate outlier analyses, including…

> • leverage – influence a case can have upon the position of the regression line (distance from centroid along line)

> • discrepancy – position relative to the envelope (distance from line at perpendicular)

> • influence – the combination of leverage & discrepancy

Now that you have "cleaned data" …

Remember – the purpose of cleaning is to improve …
- the representativeness of the sample stats to the population
- the fit of the sample distributions to the assumptions of the statistical analysis models used

Need to tell readers about what you've done, why, the result, etc.

Reporting formats vary greatly – look at work where you intend to publish or present.

- Be sure to present criteria for identification & procedure used
  - Skewness – tell original skew, the transformation and resulting skew
  - Outliers – 1.5 +/- Tukey's Hinge & trim or Winsorize
  - Bivariate outliers – scatterplot identification & trim
  - MR influence – specific stat, criterion (1.0), & trim
- Rarely do you present the pre- and post-cleaning statistical results (even univariate) – usually only the post