Athabasca University
Canada's Open University

www.athabascau.ca

Planning Your Analysis II: How And Why To Clean Your Data And Check Your Analysis Assumptions

Shawn N Fraser, PhD

Faculty of Health Disciplines
Athabasca University

Athabasca University

Outline

- Proviso
- Why to clean
- How to clean
- Examples
- Checking assumptions
- Examples

Athabasca University

Provisos

- Oversimplified
- Overly prescriptive
- Applied focus
- Brief overview of a few topics

Athabasca University

Provisos

- Power of method
  - Survey design
  - Participant burden
  - Ensuring accurate measurement
  - Inclusion/exclusion criteria
  - Independent observations

Athabasca University

Cleaning your data – Why

- Outliers
- Incorrect values
- Missing data
- Assumptions of tests

(Osborne & Overbay, 2004; Tabachnick & Fidell, 2007)

Athabasca University

Cleaning your data – Data effects

- Reduced power
  - due to increase error variance
- Decrease normality
  - Skewing distribution
- Inflate or deflate estimates
  - Biased picture of reality (even though it may be 'real')
- May violate assumptions

(Osborne & Overbay, 2004; Tabachnick & Fidell, 2007)

Athabasca University

Cleaning your data – data check

- Know your data:
  - Range of possible values for each variable
  - Plausible and likely distribution
  - Variable types
- 1st step is examining descriptive statistics and plots to check data
  - Are values within normal range?
    - E.g., age range 18- 65 years, item score from 1-5, height 48 to 78 inches
  - Are values plausible?
    - E,g, age = 124 years, item score = 7, height 84 inches
  - Is data missing?

Athabasca University

Cleaning your data

- Data entry error
  - Participant's or researcher's
- Data file error
  - 'missing value' codes
- Failure to enforce Inclusion/exclusion criteria
  - Age, disease status, medicated, …
- May be 'real' data
- Response 'depends'
  - If an error, fix it
  - If real, what is the influence?
  - Cooks D, DFITS, etc. can identify 'influence' of the outlier(s)

Cleaning your data – example 1

Athabasca University

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
student height in inches	50	692.00	7.00	699.00	77.3440	90.57381
gender of student	50	2	0	2	1.44	.577
marital status	49	2	1	3	1.82	.782
age group	50	2	1	3	1.96	.807
does subject have children	50	1	0	1	.52	.505
amount of tv watched per week	50	21	4	25	11.98	6.096
television shows-sitcoms	50	1	0	1	.64	.485
television shows-movies	50	1	0	1	.36	.485
television shows-sports	50	1	0	1	.52	.505
television shows-news shows	50	1	0	1	.46	.503
hours of study per week	48	36	2	38	15.60	8.444
student's current gpa	50	34.6	2.4	37.0	4.450	6.4209
positive evaluation, institution	50	3	2	5	3.38	.945
positive evaluation, major	49	4	1	5	3.27	.953
positive evaluation, facilities	50	4	1	5	2.76	1.061
positive eval, social life	50	4	1	5	3.10	1.182
hours per week spent working	49	50	0	50	26.12	14.857
Valid N (listwise)	45					

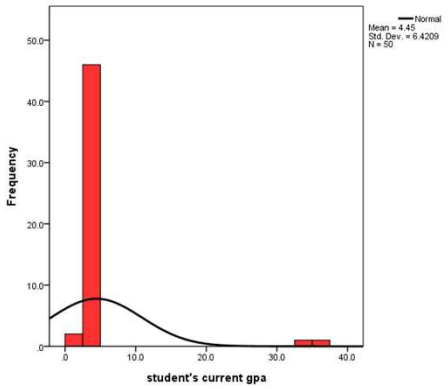
Cleaning your data – data check

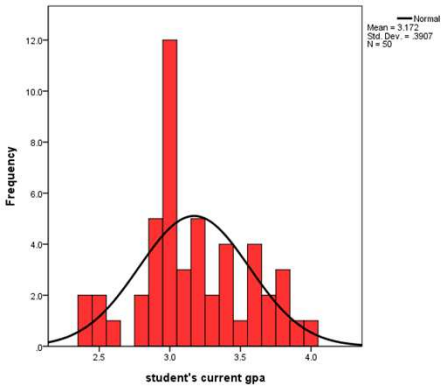
Athabasca University

- Plots
  - Plausible distribution
  - Unusual number of low or high values
  - Identify specific high or low values – Outliers

Athabasca University

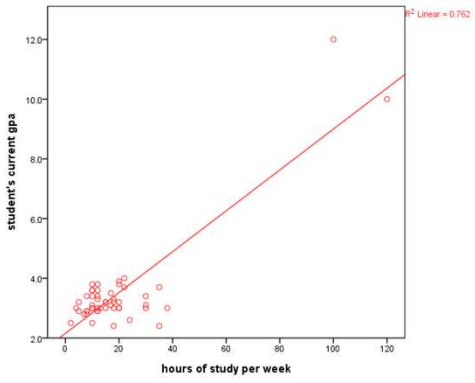
Cleaning your data – example 2

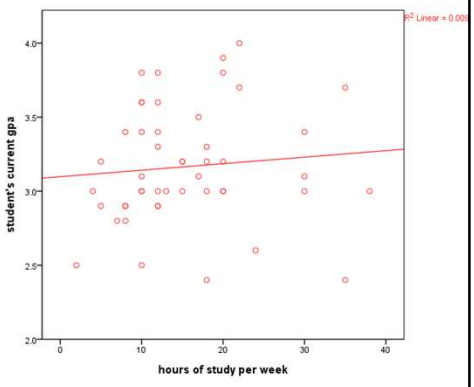




Athabasca University

Cleaning your data – example 3



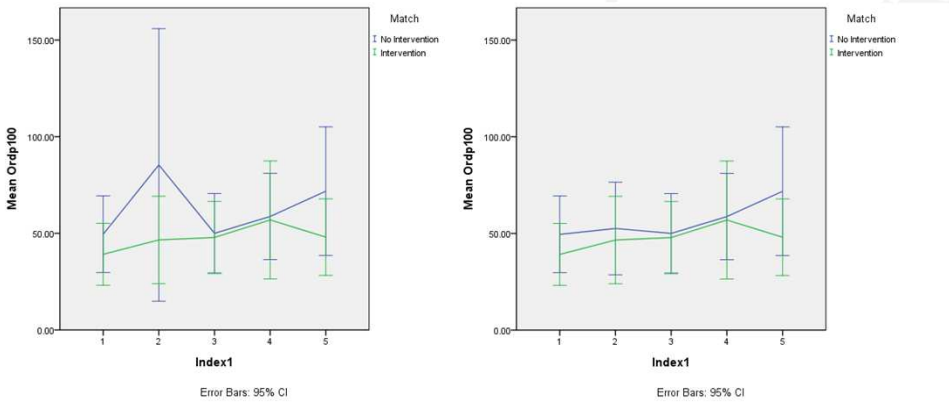


r = .87

r = .09

Athabasca University

Cleaning your data – example 4



Athabasca University

Cleaning your data – Summary

- Data file will contain errors
- Some errors can be prevented, some cannot
- Know your data by examining:
  - Descriptive statistics
  - Plots
  - Correlations
- Remedial actions
  - Replace/correct data errors, missing values
  - Delete errors when unsure
  - Replace or delete influential outliers

Athabasca University

Background - assumptions

- Statistics is a science like any other
- Statistical tests require evidence to support their use under a variety of conditions
  - Missing data
  - Distributional properties of data
  - Type of variable – ordinal, nominal, etc.
  - Sample size
  - Homogeneity of variance
- Research shows effects of varying conditions on statistical tests

Athabasca University

Background - assumptions

- Tests make assumptions about these conditions
  - E.g., normally distributed sample/population data, equal variances between groups, independence of observations, etc.
- Body of evidence shows how well certain tests 'perform' under varying conditions
  - The extent to which the assumptions are violated is the key issue
  - Many tests are 'robust' with respect to all but major violations
  - Some tests are sensitive to violations
  - For some tests/analyses it is unclear (HLM, SEM)

Typical assumptions

- T-test
  - Normal distribution in each group
  - Equal group variances
  - Independent or dependent observations
    - Different tests
- ANOVA
  - Independent observations
  - DV normally distributed/interval level
  - Equal group variances

Athabasca University

Typical assumptions

- Correlation & Regression
  - Linear relationships*
  - Variance is the same for all values of X*
  - Y and Residual error are normally distributed*
  - Perfect measurement
  - Independent observations
  - Model is 'correct'

  - *Extremely robust except for severe violations

(Berry, 1993; Cohen et al., 2003; Kleinbaum et al., 1988; Osborne & Waters, 2002)

Violations

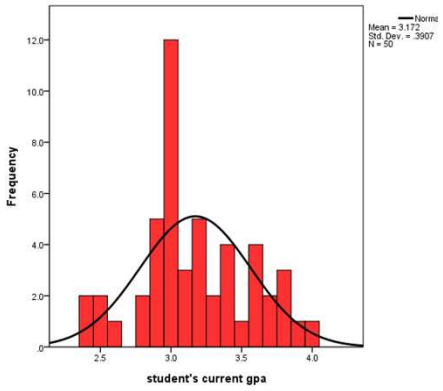
- Result could be:
  - Increase error and reduce power
  - Increase risk of type 1 errors
  - Attenuated or overestimated correlations, effects

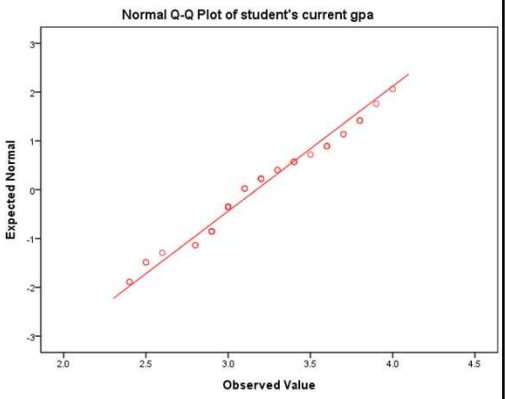
Athabasca University

Testing assumptions

- Specific tests:
  - Reliability, normality, equal variances,…
- Plots
  - Inspect for normality, linearity, …
- Methods
  - Study design, sample selection, measurement, procedures, …

Testing assumptions – example 1

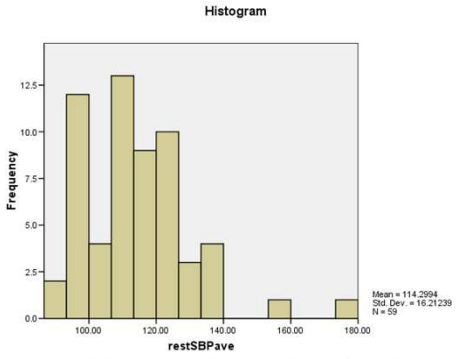


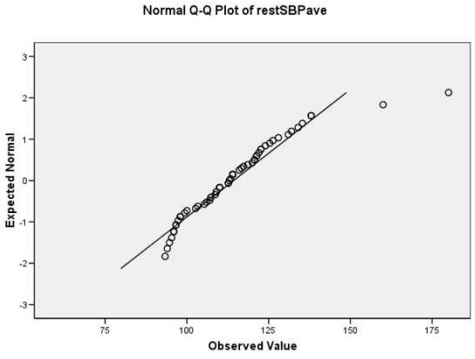


Normality test statistic, Shapiro-Wilk = .96, p = .10.
Do not reject Null, i.e., distribution is 'normal'

Athabasca University

Testing assumptions – example 2

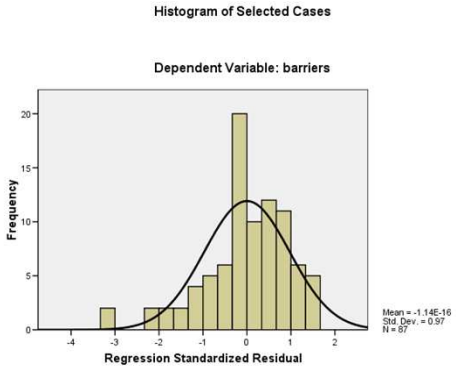


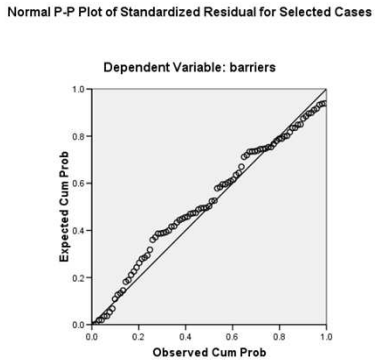


Normality test statistic, Shapiro-Wilk = .89, p < .001
Reject Null, i.e., distribution is 'not normal'
Identify outliers and their impact

Athabasca University

Testing assumptions – example 3

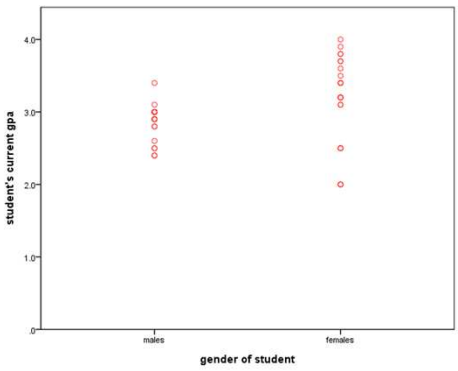


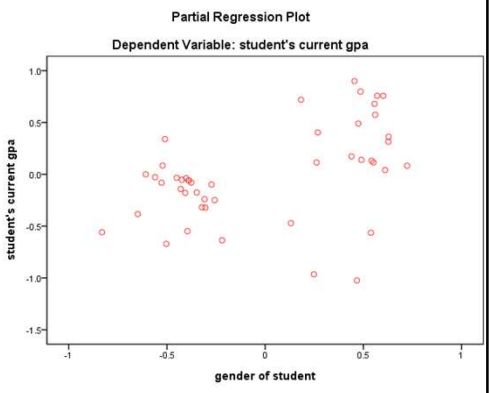


Slightly skewed distribution, minor departure from normality.

Athabasca University

Testing assumptions – example 4





Plot: Variance in GPA differs by gender.
Levene test = 13.77, p = .001. Reject Null (variances are equal)

Athabasca University

Remedial action?

- Clean data first
- Severely non-normal data?
  - Consider transformations
    - Log transformation, square root, …(Stevens, 2009)
  - Non-parametric alternatives to ANOVA
    - Mann-Whitney U, Wilcoxon signed-rank test
  - Alternative formulations for correlation, regression
    - Point biserial correlation, Spearman's rank order correlation, ordinal regression, ….
- Severe unequal variances
    - Not a concern with large and nearly equal sample sizes…generally robust
    - Transformations may help

Athabasca University

Remedial action?

- Non-independence?
  - Correlated observations
    - 'Diagnosed' with the intraclass correlation
    - Even a low ICC does not guarantee data is not clustered or observations are not correlated
  - Not robust, violation is serious
  - Consider hierarchical linear modeling or SEM
  - Paired rather than an independent t-test
- Non-linear
  - Consider transformations
  - higher order coefficients – x, x^2, x^3,…

Athabasca University

Testing assumptions– Summary

- Data will not be perfectly normal
  - Not usually a problem, especially with large data sets
- Some assumptions are robust
  - Linearity, normality, homogeneity of variance
- Some assumptions are not robust
  - Good measurement, independence of observations
- Testing assumptions
  - Is possible and 'easy' enough in popular stats packages

Athabasca University

References

Berry, W.D. (1993). *Understanding regression assumptions*. Newbury Park: Sage.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation*, *9(6)* Online at http://pareonline.net/getvn.asp?v=9&n=6

Osborne, J.W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2). Retrieved April 16, 2011 from http://PAREonline.net/getvn.asp?v=8&n=2 .

Kleinbaum, D.G., Kupper, L.L., & Muller, K.E. (1988). *Applied regression analysis and other multivariable methods*. Boston: PWS-KENT Publishing Company.

Stevens, J. S. (2009). *Applied multivariate statistics for the social sciences*. New York, N.Y. : Routledge.

Athabasca University
Canada's Open University™

Shawn Fraser

shawn.fraser@athabascau.ca