

APPLIED RESEARCH DESIGN AND ANALYSIS

Missing Data, Assessing Normality, and Checking for Outliers



1

MISSING DATA



Rare to obtain complete data from every person with applied research



Run **Descriptives** to find number or % of values missing from each variable

If a variable has lots of missing data, need to think why?
Also, is it random or is there a systemic pattern?



Options button in SPSS offers choices to deal with missing data

But be careful – the choice will have dramatic effects

2

Assessing Normality

In SPSS, use the **Explore** function in **Descriptives**

Interpretation from Output:

- Descriptives
- Trimmed Mean
- Extreme values
- Skewness
- Kurtosis
- Tests of Normality – Kolmogorov-Smirnov
- Histogram
- Normal Probability Plots (Normal Q-Q Plot)
- Detrended Normal Q-Q plot
- Boxplots

3

Descriptives

Descriptives		Statistic	Std. Error
Intention to Eat Healthy Score at Time Two	Mean	23.10	.191
	95% Confidence Interval for Mean	Lower Bound 22.72	
		Upper Bound 23.47	
	5% Trimmed Mean	23.06	
	Median	22.00	
	Variance	14.160	
	Std. Deviation	3.763	
	Minimum	2	
	Maximum	30	
	Range	28	
	Interquartile Range	5	
	Skewness	.037	.124
	Kurtosis	1.593	.247

- Provides information on the dataset as a whole unless you specify you want it by a grouping factor
- Compare the original Mean to the 5% trimmed Mean
 - Ideally, should be similar
 - If not, then check the Extreme values box

4

Descriptives

Descriptives		Statistic	Std. Error
Intention to Eat Healthy Score at Time Two	Mean	23.10	.191
	95% Confidence Interval for Mean	Lower Bound 22.72	
		Upper Bound 23.47	
	5% Trimmed Mean	23.06	
	Median	22.00	
	Variance	14.160	
	Std. Deviation	3.763	
	Minimum	2	
	Maximum	30	
	Range	28	
	Interquartile Range	5	
	Skewness	.037	.124
	Kurtosis	1.593	.247

- Check Skewness and Kurtosis
 - Provides information on the distribution of the scores on continuous variables

5

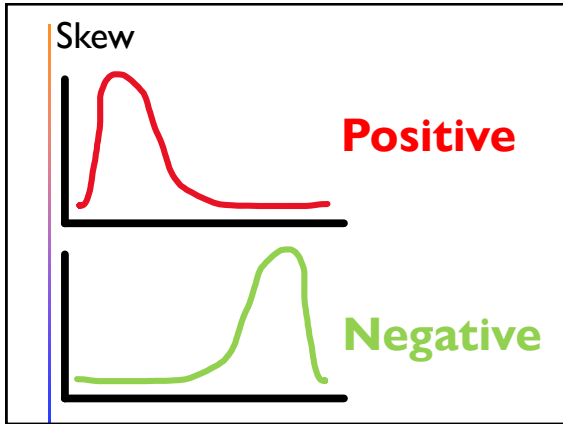
Skewness

Range	28
Interquartile Range	5
Skewness	.037
Kurtosis	1.593

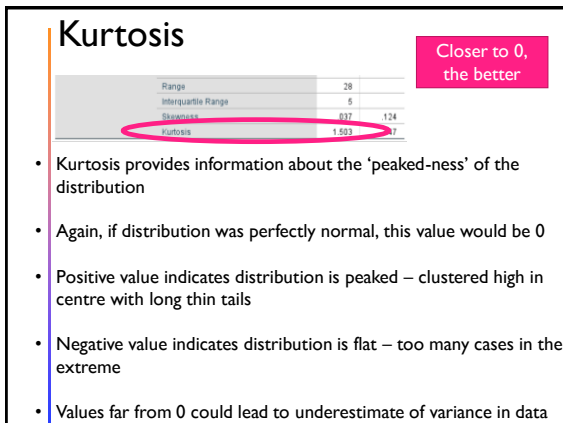
Closer to 0, the better

- Skewness value provides indication of symmetry of the distribution
- If distribution was perfectly normal, this value would be 0
- Positive value indicates positive skew - scores clustered to left on low end of scale
- Negative value indicates negative skew - scores clustered to right on high end of scale
- With large sample size, skew should not make a substantive difference in the analysis (Tabachnick & Fidell, 2007)

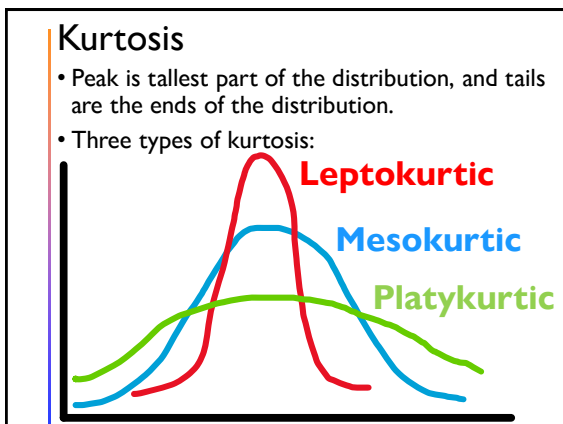
6



7



8



9

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Intention to Eat Healthy Score at Time Two	.136	390	.000	.922	390	.000

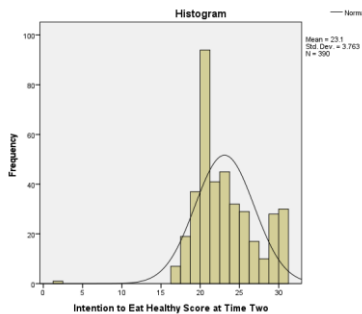
a. Lilliefors Significance Correction

- Two tests of Normality provided:
 - Kolmogorov-Smirnov
 - Shapiro-Wilk
- **Non-sig** result indicates normality
- Here, sig value is .000, suggesting violation of assumption of normality
- Common in applied work and with large samples.

10

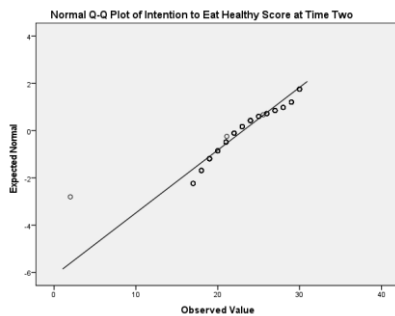
Histogram

- Shows actual shape of distribution



11

Normal Probability Plots (Normal Q-Q Plot)



Ideally, want all data points to fall on or as close as possible to the line

I.e., want a straight line

12

Detrended Normal Q-Q plot

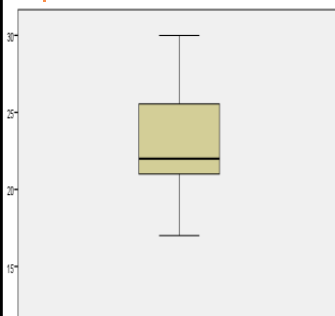


- Obtained by plotting actual deviation scores from straight line

- Should be no clustering of scores – most should be around zero line

13

Boxplots



- Rectangle part represents 50% of cases with line going up to largest value and line going down to lowest value

- Line inside is the median value

14

Outliers



- Number linked to circle outside the range is classified as an outlier – i.e., more than 1.5 box lengths from edge of box

- Number linked to asterisk outside the range is classified as extreme outlier – i.e., more than 3 box lengths from edge of box

- Specific number denotes ID number of the case

15
