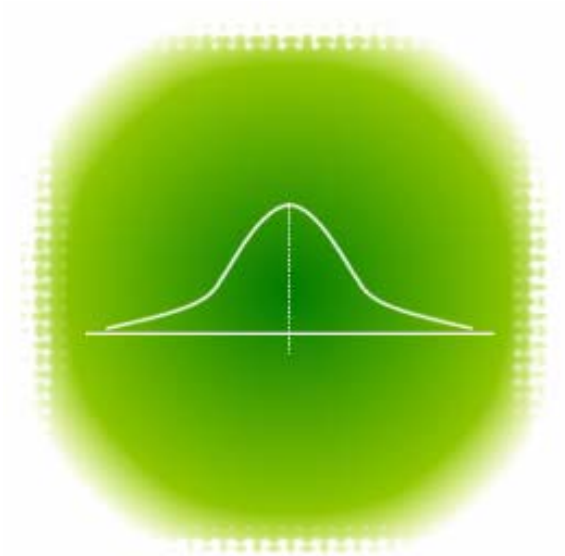
Applied Statistics Handbook

Phil Crewson



Version 1.2

Applied Statistics Handbook

http://www.acastat.com

Table of Contents

Preface

Approach Used in this Handbook

The Applied Statistics Handbook was developed to serve as a quick reference for undergraduate and graduate liberal arts students taking research methods courses. The Handbook augments classroom lecture and commonly available statistical texts by providing an easy to follow outline for conducting and interpreting data analysis and hypothesis tests. It was not designed as a stand-alone statistical text, although some may wish to use it concurrently with a comprehensive lecture series.

The approach of this Handbook is to present commonly used steps and formulas in statistics, provide an example of how to conduct the calculations by hand, and then an example of software output. The software output has annotated interpretations. This approach connects classically taught statistics with statistical software output. The output is very similar to SAS, SPSS, and other common statistical software, so skills learned with the Handbook are transferable to almost any statistical software to include the analysis module available in Microsoft Excel.

AcaStat Software

AcaStat is an easy to use statistical software system that provides modules for analyzing raw (electronic data not aggregated) and summary data (multiple records reduced to counts, means, proportions, etc.). Most of the software output presented in the Handbook was created with AcaStat.

Student Workbook

The Handbook works well with the Student Workbook. The Workbook includes examples and practical exercises designed to teach applied analytical skills. It is designed to work with AcaStat software but can also be used with other statistical software packages.

The Workbook is a free download from http://www.acastat.com.

Software Examples

The following symbols are used to indicate which software is most appropriate for the examples presented.

AcaStat Software: Used for raw data (StatProcs and Data Grid) or summary data (SumStats).

Spreadsheet Template: A spreadsheet for Excel (SumStats.xls) and OpenOffice Calc (SumStats.ods) duplicates the SumStat module available in AcaStat. If you do not use AcaStat, these templates can be used along with SPSS or another statistical software package to produce the examples. OpenOffice is a free download from www.openoffice.org.

Microsoft Excel: Excel provides data analysis tools to conduct statistical analyses to include descriptives, t-tests, ANOVA, correlation, and regression. To access these tools, open Excel and use the Tools pull-down menu to select Add Ins/Data Analysis. Although it is also possible to replicate frequency and crosstabulation procedures using Excel's Pivot Tables, considerable practice is needed to master their use, so the Handbook recommends AcaStat or SPSS instead of Excel for these type of analyses.

Statistical Software: The SPSS icon is used to represent all off-the-shelf statistical software such as SAS, STATA, Minitab, etc. These commercially available packages analyze raw data.

3.1 Univariate Descriptive Statistics

Measures of Central Tendency

Mode: The most frequently occurring score. A distribution of scores can be unimodal (one score occurred most frequently), bimodal (two scores tied for most frequently occurring), or multimodal. In the table below the mode is 32. If there were also two scores with the value of 60, we would have a bimodal distribution (32 and 60).

Median: The point on a rank ordered list of scores below which 50% of the scores fall. It is especially useful as a measure of central tendency when there are very extreme scores in the distribution, such as would be the case if we had someone in the age distribution provided below who was 120. If the number of scores is odd, the median is the score located in the position represented by (n+1)/2. In the table below the median is located in the 4th position (7+1)/2 and would be reported as a median of 42. If the number of scores are even, the median is the average of the two middle scores. As an example, if we dropped the last score (65) in the above table, the median would be represented by the average of the 3rd (6/2) and 4th score, or 37 (32+42)/2. Always remember to order the scores from low to high before determining the median.

Variable➔	Age	← Also known as X		
	24			
	32	← Mode		
	32			
	42	← Median	Xi	←Each score
	55			
	60			
	65			
n=	7	← Number of scores (or cases)		
$\Sigma X_i =$	310	← Sum of scores		
$\overline{X} =$	44.29	← Mean		

Mean: The sum of the scores (ΣX_i) is divided by the number of scores (n) to compute an arithmetic average of the scores in the distribution. The mean is the most often used measure of central tendency. It has two properties: 1) the sum of the deviations of the individual scores (Xi) from the mean is zero, 2) the sum of squared deviations from the mean is smaller than what can be obtained from any other value created to represent the central tendency of the distribution. In the above table the mean age is 44.29 (310/7).

Weighted Mean: When two or more means are combined to develop an aggregate mean, the influence of each mean must be weighted by the number of cases in its subgroup.

$$\overline{X}_w = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3}{n_1 + n_2 + n_3}$$

Example

$\overline{X}_1 = 12, n = 10$

$\overline{X}_2 = 14, n = 15$

$\overline{X}_3 = 18, n = 40$

Wrong Method: $\frac{(12+14+18)}{3} = 14.7$

Correct Method: $\frac{10(12)+15(14)+40(18)}{10+15+40} = 16.2$

Measures of Variation

Range: The difference between the highest and lowest score (high-low). It describes the span of scores but cannot be compared to distributions with a different number of observations. In the table below, the range is 41 (65-24).

Variance: The average of the squared deviations between the individual scores and the mean. The larger the variance the more variability there is among the scores. When comparing two samples with the same unit of measurement (age), the variances are comparable even though the sample sizes may be different. Generally, however, smaller samples have greater variability among the scores than larger samples. The sample variance for the data in the table below is 251.57. The formula is almost the same for estimating population variance. See formula in Appendix.

Standard deviation: The square root of variance. It provides a representation of the variation among scores that is directly comparable to the raw scores. The sample standard deviation in the following table is 15.86 years.

Variable➔	Age	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	← $\overline{X} = 44.29$
	24	-20.29	411.68	
	32	-12.29	151.04	
	32	-12.29	151.04	← squared deviations
	42	-2.29	5.24	
	55	10.71	114.70	
	60	15.71	246.80	
	65	20.71	428.90	
n=	7		1509.43	← $\Sigma\left(X_i - \overline{X}\right)^2$
		$S^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1}$	251.57	← sample variance
		$S = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1}}$	15.86	← sample standard deviation

Software Output

Raw data from previous example

```
Descriptive Statistics
Variable: Age
-------------------------------------------------------------------
Count                 7           Pop Var              215.6327 ②
Sum                310.0000       Sam Var              251.5714
Mean                44.2857       Pop Std               14.6844
Median              42.0000       Sam Std               15.8610
Min                 24.0000       Std Error              6.4752
Max                 65.0000       CV%                   35.8152 ③
Range               41.0000       95% CI (+/-)          15.8449 ④
Skewness             0.0871 ①     t-test (mu=0) |p<|     0.0005 ⑤
-------------------------------------------------------------------

Missing Cases    0
```

Interpretation

① Skewness provides an indication of the how asymmetric the distribution is for a given sample. A negative value indicates a negative skew. Values greater than 1 or less than -1 indicate a non-normal distribution.

② Population variance and standard deviation (Std) will always be less than sample variance and standard deviation, since you are dividing the sum of squares by n instead of n-1.

③ Coefficient of Variation (CV) is the ratio of the sample standard deviation to the sample mean: (sample standard deviation/sample mean)*100 to calcuate CV%. It is used as a measure of relative variability, CV is not affected by the units of a variable.

④ Add and subtract this value to the mean to create a 95% confidence interval.

⑤ This represents the results of a one-sample t-test that compares the sample mean to a hypothesized population value of 0 years of age. In this example, the sample mean age of 44 is statistically significantly different from zero.

Standardized Z-Scores

A standardized z-score represents both the relative position of an individual score in a distribution as compared to the mean and the variation of scores in the distribution. A negative z-score indicates the score is below the distribution mean. A positive z-score indicates the score is above the distribution mean. Z-scores will form a distribution identical to the distribution of raw scores; the mean of z-scores will equal zero and the variance of a z-distribution will always be one, as will the standard deviation.

To obtain a standardized score you must subtract the mean from the individual score and divide by the standard deviation. Standardized scores provide you with a score that is directly comparable within and between different groups of cases.

$$Z = \frac{X_i - \overline{X}}{S}$$

Variable➔	Age	$X_i - \overline{X}$	$Z = \frac{X_i - \overline{X}}{S}$	Z
Doug	24	-20.29	-20.29/15.86	-1.28
Mary	32	-12.29	-12.29/15.86	-0.77
Jenny	32	-12.29	-12.29/15.86	-0.77
Frank	42	-2.29	-2.29/15.86	-0.14
John	55	10.71	10.71/15.86	0.68
Beth	60	15.71	15.71/15.86	0.99
Ed	65	20.71	20.71/15.86	① 1.31

Interpretation

① As an example of how to interpret z-scores, Ed is 1.31 standard deviations above the mean age for those represented in the sample. Another simple example is exam scores from two history classes with the same content but difference instructors and different test formats. To adequately compare student A's score from class A with Student B's score from class B you need to adjust the scores by the variation (standard deviation) of scores in each class and the distance of each student's score from the average (mean) for the class.

Software Output

Raw data from previous example

The following displays z-scores automatically created by software when conducting a descriptive analysis (an option on most software). Software calculates the mean and standard deviation for the sample data and then calculates the z-score for each observation and outputs the result to the data file thereby creating a new variable (in this case Z-Age).

Obs	Age	Z-Age
1	24	-1.27897
2	32	-0.77459
3	32	-0.77459
4	42	-0.14411
5	55	0.67551
6	60	0.99075
7	65	1.30599

The following displays the descriptive statistics for the Z-Age variable. Notice that the mean is zero and variance and standard deviation are one.

```
Descriptive Statistics
Variable: Z-Age
-------------------------------------------------------------------
Count                    7         Pop Var                   0.8571
Sum                      0.0000    Sam Var                   1.0000
Mean                     0.0000    Pop Std                   0.9258
Median                  -0.1441    Sam Std                   1.0000
Min                     -1.2790    Std Error                 0.4082
Max                      1.3060    CV%              -70000119.5370
Range                    2.5850    95% CI (+/-)              0.9990
Skewness                 0.0871    t-test (mu=0)  |p<|       1.0000
-------------------------------------------------------------------

Missing Cases    0
```