Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data

Seven: Extreme and Influential Data Points: Debunking the Myth of Equality

Contributors: Jason W. Osborne
Book Title: Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data
Chapter Title: "Seven: Extreme and Influential Data Points: Debunking the Myth of Equality"
Pub. Date: 2013

Access Date: November 11, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412988018
Online ISBN: 9781452269948
DOI: http://dx.doi.org/10.4135/9781452269948.n7
Print pages: 139-169

http://dx.doi.org/10.4135/9781452269948.n7*Old Dominion University*

[p. 139 ↓]

Seven: Extreme and Influential Data Points: Debunking the Myth of Equality

Next time you read an article from a top journal in your field, look for any mention of looking for influential data points (or extreme scores or outliers). Odds are you will find none (as my students and I have found in several surveys across several disciplines). Authors spend a great deal of time describing the importance of the study, the research methods, the sample, the statistical analyses used, results, and conclusions based on those results, but rarely mention having screened their data for outliers or extreme scores (sometimes referred to as influential data points). Many conscientious researchers do check their data for these things, perhaps neglecting to report having done so, but more often than not, this step is skipped in the excitement of moving directly to hypothesis testing. After all, researchers often spend months or years waiting for results from their studies, so it is not surprising they are excited to see the results of their labors. Yet jumping directly from data collection to data analysis without examining data for extreme scores or inappropriately influential scores can, ironically, decrease the likelihood that the researcher will find the results they so eagerly anticipate.

Researchers from the dawn of the age of statistics have been trained in the effects of extreme scores, but more recently, this seems to have waned. In fact, a recent article of mine examining publications in respected educational psychology journals (Osborne, 2008) found that only 8% of these articles reported testing any sort of assumption, and almost none specifically discussed having examined data for extreme scores. There is no reason to believe that the situation is different in other disciplines. Given what we know of the importance **[p. 140 ↓]** of assumptions to accuracy of estimates and error rates (Micceri, 1989; Yuan, Bentler, & Zhang, 2005), this is troubling, and it leads to the conclusion that research in the social sciences is probably at increased risk for errors of inference, problems with generalizability, and suboptimal outcomes. One can only conclude that most researchers assume that extreme scores do not exist, or that if they exist, they have little appreciable influence on their analyses. Hence, the goal of this chapter is to debunk the myth of equality, the myth that all data points are equal. As this chapter shows, extreme scores have disproportionate, usually detrimental, effects on analyses.

Some techniques, such as “robust” procedures and nonparametric tests (which do not require an assumption of normally distributed data) are often considered to be immune from these sorts of issues. However, parametric tests are rarely robust to violations of distributional assumptions (Micceri, 1989) and nonparametric tests benefit from clean data, so there is no drawback to cleaning your data and looking for outliers and fringeliers, whether you are using parametric or nonparametric tests (e.g., Zimmerman, 1994, 1995, 1998).

The goal of this step is to decrease the probability you will make a significant error of inference, as well as to improve generalizability, replicability, and accuracy of your results by making sure your data includes only those data points that belong there.

What are Extreme Scores?

Figure 7.1, on page 142, visually shows the concept of the extreme score, including the outlier and the fringelier. Although definitions vary, an outlier is generally considered to be a data point that is far outside the norm for a variable or population (e.g., Jarrell, 1994; Rasmussen, 1988; Stevens, 1984). It is an observation that “deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980, p. 1). Hawkins's description of outliers reinforces the notion that if a value is very different because it reflects different processes (or populations), then it does not belong in the analysis at hand. Outliers also have been defined as values that are “dubious in the eyes of the researcher” (Dixon, 1950, p. 488) and contaminants (Wainer, 1976), all of which lead to the same conclusion: extreme scores probably do not belong in your analyses. That is not to say that these extreme scores are not of value. As discussed next, they most likely should be examined more closely and in depth.

[p. 141 ↓]

The scholarly literature on extreme scores reveals two broad categories: scores typically referred to as *outliers*, which are clearly problematic in that they are far from the rest of the distribution, and *fringeliers*, which are scores hovering around the fringes of a normal distribution that are unlikely to be part of the population of interest but less clearly so (Wainer, 1976, p. 286). We can operationalize fringeliers as those scores around ± 3.0 standard deviations (*SD*) from the mean, which represents a good (but by no means the only possible) rule of thumb for identifying scores that merit further examination.

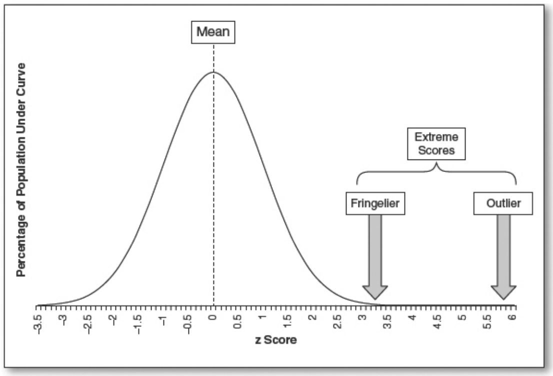
Why are we concerned with scores around ± 3.0 standard deviations from the mean? Recall from Chapter 5 that the standard normal distribution is symmetrical distribution with known mathematical properties. Most relevant to our discussion, we know what percentage of a population falls at any given point of the normal distribution, which also gives us the probability that an individual with a given score on the variable of interest would be drawn at random from a normally distributed population. So, for example, we know that in a perfectly normal distribution, 68.2% of the population will fall within 1 standard deviation of the mean, about 95% of the population should fall within 2 standard deviations from the mean, and 99.74% of the population will fall within 3 standard deviations. In other words, the probability of randomly sampling an individual more than 3 standard deviations from the mean in a normally distributed population is 0.26%, which gives me good justification for considering scores outside this range as suspect.

Because of this, I tend to be suspicious that data points outside ± 3.0 standard deviations from the mean are *not part of the population of interest*, and furthermore, despite being plausible (though unlikely) members of the population of interest, these scores can have a disproportionately strong influence on parameter estimates and thus need to be treated with caution. In general, since both outliers and fringeliers represent different magnitudes of the same problem (single data points with disproportionately high influence on statistics) I refer to them here collectively as *extreme values*.

How Extreme Values Affect Statistical Analyses

Extreme values can cause serious problems for statistical analyses. First, they generally increase error variance and reduce the power of statistical tests by altering the skew or kurtosis of a variable (which can become very problematic **[p. 142 ↓]** in multivariate analyses). As many statistical tests compare variance accounted for to error (unexplained) variance, the more error variance in the analyses, the less likely you are to find a statistically significant result when you should (increasing the probability of making a Type II error).

Figure 7.1 Extreme Scores: Outliers and Fringeliers



Second, they can seriously bias or influence estimates that may be of substantive interest, such as means, standard deviation, and the like (for more information on these points, see Rasmussen, 1988; Schwager & Margolin, 1982; Zimmerman, 1994). Since extreme scores can substantially bias your results, you may be more likely to draw erroneous conclusions, and any conclusions you do draw will be less replicable and generalizable, two important goals of scientific quantitative research.

I explore each of these effects and outcomes in this chapter.

What Causes Extreme Scores?

Extreme scores can arise from several different mechanisms or causes. Anscombe (1960) sorts extreme scores into two major categories: those arising from errors **[p. 143 ↓]** in the data and those arising from the inherent variability of the data. I elaborate on this idea to summarize six possible reasons for data points that may be suspect.

Let me first be careful to note that not all extreme scores are illegitimate contaminants, and not all illegitimate scores show up as extreme scores (Barnett & Lewis, 1994). Although the average American male stands about 5′ 10″ tall, there are 7-foot-tall males and 4-foot-tall males. These are legitimate scores, even though they are relatively extreme and do not describe the majority of the American male population. Likewise, it is possible that a score of 5′5″ (what seems to be a very legitimate score) could be an error, if the male was in reality 6′5″ but the data was recorded incorrectly.

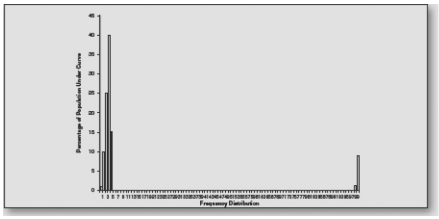
It is therefore important to consider the range of causes that may be responsible for extreme scores in a given data set. What should be done about an outlying data point is very much a function of the inferred cause.

The Case of the Mysterious 99s

Early in my career I was working with data from the U. S. National Center for Educational Statistics (NCES), analyzing student psychological variables such as self-esteem. With many thousands of subjects and previous research showing strong correlations between the variables we were researching, I was baffled to discover correlations that were substantially lower than what we expected.

Exasperated, I informed a professor I was working with of the problem, who merely smiled and suggested checking the data for outliers.

Figure 7.2 The Case of the Mysterious 99s



[p. 144 ↓]

Immediately the problem became apparent. Items, such as the one in Figure 7.2, had responses from 1–5, a typical Likert-type scale item, and then a small but significant number of 98s and 99s. I learned that many researchers and government data sets use numeric codes for missing data, rather than just leaving the data field blank. There are good reasons for this.

First, in earlier days, computers had difficulty handling blanks in data, so entering numeric codes for missing data was important. Second, there are sometimes different reasons for missing data, and the NCES had different codes so they could analyze the missingness in the data (as we discussed in Chapter 6). Identifying 99 and 98 as missing data immediately solved the problem, but I never forgot the lesson: always check for extreme scores!

1. *Extreme Scores From Data Errors*. Extreme scores, particularly outliers, are often caused by human error, such as errors in data collection, recording, or entry. Data gathered during research can be recorded incorrectly and mistakes can happen during data entry. One survey I was involved with gathered data on nurses' hourly wages, which at that time averaged about $12.00 per hour with a standard deviation of about $2.00 per hour. In our data set one nurse had reported an hourly wage of $42,000.00, clearly not a legitimate hourly wage in nursing. This figure represented a data collection error (specifically, a failure of the respondent to read the question carefully—she reported *yearly* wage rather than *hourly* wage). The good news about these types of errors is that they can often be corrected by returning to the original documents or even possibly contacting the research participant, thus potentially eliminating the problem. In cases such as this, another option is available—estimation of the correct answer. We used anonymous surveys, so we could not contact the nurse in question, but because

the nature of the error was obvious, we could convert this nurse's salary to an estimated hourly wage because we knew how many hours per week and how many weeks per year she worked.

Data entry is a significant source of extreme scores, particularly when humans are hand-entering data from printed surveys (the rise of web-based surveys are helpful in this regard). Recently, I was analyzing some hand-entered data from a Likert scale where values should range from 1 to 7, yet I found some *0* and *57* values in the data. This obviously arose from human entry error, and returning to the original surveys allowed for entry of correct values.

[p. 145 ↓]

If extreme scores of this nature cannot be corrected they should be eliminated as they do not represent valid population data points, and while it is tempting to assume the 0 was supposed to be a 1 or 2 (which is right above the 0 on a numeric keypad) and the 57 was supposed to probably be a 5 (and the data entry person hit both keys accidentally) researchers *cannot make those assumptions* without reasonable rationale to do so. If you do such a thing, be sure to be transparent and report having done so when you present your results.

A final, special case of this source of extreme score is when researchers (such as government agencies) use numeric codes for missing data, but researchers fail to identify those codes to the statistical software as missing. This is a simple process that all modern statistical software does easily, but can be disastrous to analyses if these codes are in the data but researchers fail to realize this (see sidebar, The Case of the Mysterious 99s).

2. *Extreme Scores From Intentional or Motivated Misreporting*. Sometimes participants purposefully report incorrect data to experimenters or surveyors. In Chapter 10, I explore various motivations for doing so, such as impression management or malingering. Yet these types of motives might not always result in extreme scores (social desirability pressures often push people toward average rather than toward unrealistic extremes).

This also can happen if a participant makes a conscious effort to sabotage the research, is fatigued, or may be acting from other motives. Motivated misreporting also can happen for obvious reasons when data are sensitive (e.g., teenagers misreporting drug or alcohol use, misreporting of sexual behavior, particularly if viewed as shameful or deviant). If all but a few teens underreport a behavior (for example, cheating on a test or driving under the influence of alcohol), the few honest responses might appear to be extreme scores when in fact they are legitimate and valid scores. Motivated overreporting can occur when the variable in question is socially desirable (e.g., income, educational attainment, grades, study time, church attendance, sexual experience) and can work in the same manner.

Environmental conditions can motivate misreporting, such as if an attractive female researcher is interviewing male undergraduates about attitudes on gender equality in marriage. Depending on the details of the research, one of two things can happen: inflation of all estimates, or production of extreme scores. If all subjects respond the same way, the distribution will shift upward, not generally causing extreme scores. However, if only a small subsample of the group responds this way to the experimenter, or if some of the male undergraduates are interviewed by male researchers, extreme scores can be created.

[p. 146 ↓]

Identifying and reducing this issue is difficult unless researchers take care to triangulate or validate data in some manner.

3. *Extreme Scores From Sampling Error or Bias*. As I discuss in Chapter 3, sampling can help create biased samples that do not reflect the actual nature of the population. Imagine you are surveying university undergraduates about the extent of their alcohol usage, but due to your schedule, the only time you could perform interviews was 8:00 to 10:00 in the mornings Friday, Saturday, and Sunday. One might imagine that heavy alcohol users might not be willing or able to get up that early on the weekend, so your sample may be biased toward low usage. If most of your sample is biased toward nondrinkers, but a few average—drinking college students by chance slip into the sample, you may well see those as extreme scores when in fact they are part of the

normal diversity in the population. Ideally, upon realizing this, you would correct your sampling plan to gather a representative sample of the population of interest.

Another cause of extreme scores is sampling error. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample. For example, in the previously described survey of nurse salaries, nurses who had moved into hospital administration were included in the database we sampled from, as they had maintained their nursing license, despite our being primarily interested in nurses currently involved in routine patient care. In education, inadvertently sampling academically gifted or mentally retarded students is a possibility, and (depending on the goal of the study) might provide undesirable extreme scores. These cases should be removed if they do not reflect the target population.

4. *Extreme Scores From Standardization Failure*. Extreme scores can be caused by research methodology, particularly if something anomalous happened during a particular subject's experience. One might argue that a study of stress levels in schoolchildren around the country might have found some significant extreme scores if the sample had included schoolchildren in New York City schools during the fall of 2001 or in New Orleans following Hurricane Katrina in 2005. Researchers commonly experience such challenges—construction noise outside a research lab or an experimenter feeling particularly grouchy, or even events outside the context of the research lab, such as a student protest, a rape or murder on campus, observations in a classroom the day before a big holiday recess, and so on can produce extreme scores. Faulty or noncalibrated equipment is another common cause of extreme scores.

[p. 147 ↓]

Let us consider two possible cases in relation to this source of extreme scores. In the first case, we might have a piece of equipment in our lab that was miscalibrated, yielding measurements that were extremely different from other days’ measurements. If the miscalibration results in a fixed change to the score that is consistent or predictable across all measurements (for example, all measurements are off by 100) then adjustment of the scores is possible and appropriate. If there is no clear way to defensibly adjust the measurements, they must be discarded.

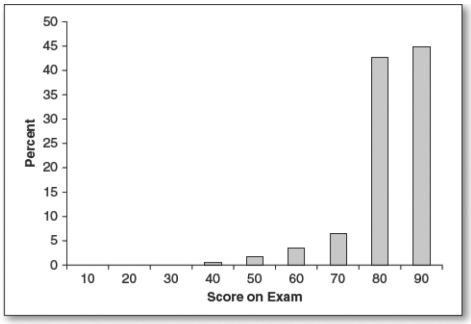
Other possible causes of extreme scores can cause unpredictable effects. Substantial changes in the social, psychological, or physical environment (e.g., a widely known crime, substantial noise outside the research lab, a natural disaster) can substantially alter the results of research in unpredictable ways, and these extreme scores should be discarded as they do not represent the normal processes you wish to study (e.g., if one were not interested in studying subjects' reactions to construction noise outside the lab, which I experienced one summer while trying to measure anxiety in a stereotype threat study).

5. *Extreme Scores From Faulty Distributional Assumptions*. Incorrect assumptions about the distribution of the data also can lead to the presence of suspected extreme scores (Iglewicz & Hoaglin, 1993). Blood sugar levels, disciplinary referrals, scores on classroom tests where students are well-prepared, and self-reports of low-frequency behaviors (e.g., number of times a student has been suspended or held back a grade) may give rise to bimodal, skewed, asymptotic, or flat distributions, depending on the sampling design and variable of interest, as Figure 7.3 shows.

The data in Figure 7.3, taken from an exam in one of the large undergraduate classes I teach, shows a highly skewed distribution with a mean of 87.50 and a standard deviation of 8.78. While one could argue the lowest scores on this test are extreme scores by virtue of distance from the mean, a better interpretation might be that the data should not be expected to be normally distributed. Thus, scores on the lower end of this distribution are in reality valid cases. In this case, a transformation could be used to normalize the data before analysis of extreme scores should occur (see Chapter 8 for details of how to perform transformations effectively) or analyses appropriate for nonnormal distributions could be used. Some authors argue that splitting variables such as this into groups (i.e., dichotomization) is an effective strategy for dealing with data such as this. I disagree, and demonstrate why in Chapter 11.

[p. 148 ↓]

Figure 7.3 Performance on Class Unit Exam, Undergraduate Educational Psychology Course



Similarly, the data may have a different structure than the researcher originally assumed, and long- or short-term trends may affect the data in unanticipated ways. For example, a study of college library usage rates during the month of August in the United States may find outlying values at the beginning and end of the month—exceptionally low rates at the beginning of the month when students are still on summer break and exceptionally high rates at the end of the month when students are just back in classes and beginning research projects. Depending on the goal of the research, these extreme values may or may not represent an aspect of the inherent variability of the data, and they may or may not have a legitimate place in the data set.

6. *Extreme Scores as Legitimate Cases Sampled From the Correct Population*. Finally, it is possible that an extreme score can come from the population being sampled legitimately through random chance. It is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails (Evans, 1999; Sachs, 1982). As a researcher casts a wider net and the data set becomes larger, the more the sample **[p. 149 ↓]** resembles the population from which it was drawn, and thus the likelihood of legitimate extreme values, becomes greater.

Specifically, if you sample in a truly random fashion from a population that is distributed in an exact standard normal distribution, there is about a 0.25% chance you will get a data point at or beyond 3 standard deviations from the mean. This means that, on average, *about 0.25% of your subjects should be 3 standard deviations from the mean*. There is also a nontrivial probability of getting individuals far beyond the 3 standard deviation threshold. For example, in the United States, assume the average height for a woman is 5′ 4″ (64 inches), with a standard deviation of 2.5 inches.[1] While the odds are highest that a sample of women will be between 4′ 11″ and 5′ 9″, if one of our female volleyball players from North Carolina State University randomly happens to participate in your study, you could easily get a legitimate data point from a woman that is 6′0″. Or if you had happened to ever meet my great-aunt Winifred Mauer, you could have included a woman about 4′ 6″ in your data set.

When legitimate extreme scores occur as a function of the inherent variability of the data, opinions differ widely on what to do. Due to the deleterious effects on power, accuracy, and error rates that extreme scores can have, I believe it is important to deal with the extreme score in some way, such as through transformation or a recoding/truncation strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference (for more on this point see Chapter 8). The alternative is removal.

Extreme Scores as a Potential Focus of Inquiry

We all know that interesting research is often as much a matter of serendipity as planning and inspiration. Extreme scores can represent a nuisance, error, or legitimate data. They can be inspiration for inquiry as well. When researchers in Africa discovered that some women were living with HIV for many years longer than expected despite being untreated (Rowland-Jones et al., 1995), those rare cases constitute extreme scores compared to most untreated women infected with HIV, who die relatively rapidly. They could have been discarded as noise or error, but instead they served as inspiration for inquiry: what makes these women different or unique, and what can we learn from them? Legitimate exceptionality (rather than motivated misinformation or exaggeration motivated by social motives) can be the source of important and useful insight into **[p. 150 ↓]** processes and phenomena heretofore unexplored. Before discarding extreme scores, researchers should consider whether those data contain valuable information that may not necessarily relate to the intended study, but has importance in a more global sense.

I have a small sample and can't afford to lose data. Can I keep my extreme scores and still not violate my assumptions?

Yes, at least in some cases. But first, let's talk about why you want to keep your data, because keeping extreme scores can cause substantial problems. Are you dealing with a specialized population that precludes you from getting a large enough sample to have sufficient power? You should be aware that you might be better off without that data point anyway. Extreme scores that add substantial error variance to the analysis may be doing more harm than good.

If your extreme case is a *legitimate* member of the sample then it is acceptable to keep that case in the data set, provided you take steps to minimize the impact of that one case on the analysis.

Assuming you conclude that keeping the case is important, one means of accommodating extreme scores is the use of transformations or truncation. By using transformations, extreme scores can be kept in the data set with less impact on the analysis (Hamilton, 1992).

Transformations may not be appropriate for the model being tested, or may affect its interpretation in undesirable ways (see Chapter 8). One alternative to transformation is truncation, wherein extreme scores are recoded to the highest (or lowest) reasonable score. For example, a researcher might decide that in reality, it is impossible for a teenager to have more than 20 close friends. Thus, all teens reporting more than this value (even 100) would be recoded to 20. Through truncation the relative ordering of the data is maintained and the highest or lowest scores remain the highest or lowest scores, yet the distributional problems are reduced. However, this may not be ideal if those cases really represent bad data or sampling error.

To be clear on this point, even when the extreme score is either a legitimate part of the data or the cause is unclear, and even if you will study the case in more depth, that is a separate study. If you want the most replicable, honest estimate of the population parameters possible, Judd and McClelland (1989) suggest removal of the extreme data points, and I concur. However, not all researchers feel that way (Orr, Sackett, & DuBois, 1991). This is a case where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions.

[p. 151 ↓]

Interestingly, analysis of extreme scores is now becoming a growth industry in data forensics, where companies attempt to catch students cheating on high-stakes tests by looking at statistical anomalies like unusual patterns of answers, agreement across test-takers that indicates copying, and unusually large gain scores (Impara, Kingsbury, Maynes, & Fitzgerald, 2005).

Advanced Techniques for Dealing with Extreme Scores: Robust Methods

Instead of transformations or truncation, researchers sometimes use various “robust” procedures to protect their data from being distorted by the presence of extreme scores. These techniques can help accommodate extreme scores while minimizing their effects. Certain parameter estimates, especially the mean and least squares estimations, are particularly vulnerable to extreme scores, or have low breakdown values. For this reason, researchers turn to robust, or high breakdown, methods to provide alternative estimates for these important aspects of the data.

A common robust estimation method for univariate distributions involves the use of a trimmed mean, which is calculated by temporarily eliminating extreme observations at both ends of the sample (Anscombe, 1960). Alternatively, researchers may choose to compute a Windsorized mean, for which the highest and lowest observations are temporarily censored, and replaced with adjacent values from the remaining data (Barnett & Lewis, 1994).

A *Univariate Extreme* Score: is one that is relatively extreme when considering only that variable. An example would be a height of $36''$ in a sample of adults.

A *Bivariate Extreme* Score: is one that is extreme when considered in combination with other data. An example would be a height of 5′2″ in a sample of adults. This score would not necessarily stand out from the overall distribution. However, in considering gender and height, if that height belonged to a male, that male would be considered an outlier within his group.

A *Multivariate Extreme* Score: is one that is extreme when considering more than two variables simultaneously. My nephew is 5′8″, which is not extreme for a male, but considering he is only 10 years old, he is extreme when age is considered.

Assuming that the distribution of prediction errors is close to normal, several common robust regression techniques can help reduce the influence of outlying data points. The least trimmed squares (LTS) and the least median of squares **[p. 152 ↓]** (LMS) estimators are conceptually similar to the trimmed mean, helping to minimize the scatter of the prediction errors by eliminating a specific percentage of the largest positive and negative extreme scores (Rousseeuw & Leroy, 1987), while Windsorized regression smooths the Y-data by replacing extreme residuals with the next closest value in the dataset (Lane, 2002). Rand Wilcox (e.g., Wilcox, 2008) is a noted scholar in the development and dissemination of these types of methods, and I would encourage readers interested in learning more about these techniques to read some of his work.

In addition to the above-mentioned robust analyses, researchers can choose from a variety of nonparametric analyses, which make few if any distributional assumptions. Unfortunately, nonparametric tests are sometimes less powerful than parametric analyses and can still suffer when extreme scores are present (e.g., Zimmerman, 1995).

Identification of Extreme Scores

The controversy over what constitutes an extreme score has lasted many decades. I tend to do an initial screening of data by examining data points three or more standard deviations from the mean, in combination with visual inspection of the data in most cases.[2] Depending on the results of that screening, I may examine the data more closely and modify the extreme score detection strategy accordingly.

However, examining data for univariate extreme scores is merely a starting point, not an end point. It is not uncommon to find bivariate and multivariate extreme scores once you start performing data analyses. Bivariate and multivariate extreme scores are easily identified in modern statistical analyses through examination of things such as standardized residuals (where I also use the ±3.0 rule for identifying multivariate extreme scores) or diagnostics commonly provided in statistical packages, such as Mahalanobis distance and Cook's distance. The latter two indices attempt to capture how far individual data points are from the center of the data, and thus larger scores are considered more problematic than smaller scores. However, there is no good rule of thumb as to how large is too large, and researchers must use their professional judgment in deciding what data points to examine more closely.

For ANOVA-type analyses, most modern statistical software will produce a range of statistics, including standardized residuals. ANOVA analyses suffer from a special type of multivariate extreme score called a within-cell extreme score. In this case, within-cell extreme scores are data points that may not be extreme in the **[p. 153 ↓]** univariate analysis, but are extreme compared to the other data points within a particular cell or group (as in the example of my nephew's height in the earlier example above). Fortunately, most modern statistical packages will allow researchers to save standardized residuals in ANOVA, regression, and many other types of analyses, allowing for straightforward examination of data for extreme scores.

Why Remove Extreme Scores?

Extreme scores have several specific effects on variables that otherwise are normally distributed. To illustrate this, I will use some examples from the National Education Longitudinal Study (NELS 88) data set from the National Center for Educational Statistics (http://nces.ed.gov/surveys/NELS88/). First, *socioeconomic status* (SES) represents a composite of family income and social status based on parent occupation (see Figure 7.4). In this data set, SES scores were reported as z scores (a distribution with a mean of 0.00 and a standard deviation of 1.0). This variable shows good (though not perfect) normality, with a mean of -0.038 and a standard deviation of 0.80. Skew is calculated to be -0.001 (where 0.00 is perfectly symmetrical).

Samples from this distribution should share these distributional traits as well. As with any sample, larger samples tend to better mirror the overall distribution than smaller samples. To show the effects of extreme scores on univariate distributions and analyses, my colleague Amy Overbay and I (Osborne & Overbay, 2004) drew repeated samples of *N* = 416, that included 4% extreme scores on one side of the distribution (very wealthy or very poor students) to demonstrate the effects of extreme scores even in large samples (I use a similar methodology to discuss the effects of extreme scores on correlation and regression and on t-tests and ANOVAs later in this chapter).

With only 4% of the sample (16 of 416) classified as extreme scores, you can see in Figure 7.5 the distribution for the variable changes substantially, along with the statistics for the variable. The mean is now -0.22, the standard deviation is 1.25, and the skew is -2.18. Substantial error has been added to the variable, and it is clear that those 16 students at the very bottom of the distribution do not belong to the normal population of interest. To confirm this sample was strongly representative of the larger population as a whole, removal of these extreme scores returned the distribution to a mean of -0.02, standard deviation = 0.78, skew = 0.01, not markedly different from the sample of more than 24,000.

[p. 154 ↓]

Figure 7.4 Distribution of Socioeconomic Status

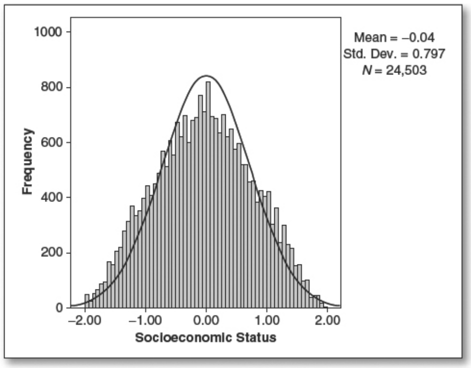
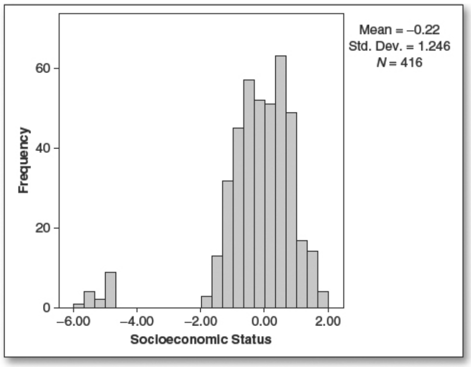
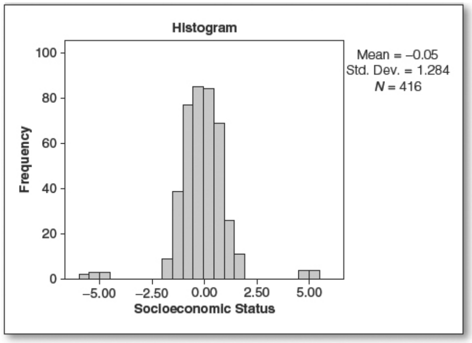


Figure 7.5 Distribution of Socioeconomic Status With 4% Extreme Scores



[p. 155 ↓]

Figure 7.6 Distribution of Socioeconomic Status With 4% Extreme Scores, Both Tails



To emphasize the point that you need to examine your data visually, we repeated the process for drawing a sample of 416 from the same data, however this time half the extreme scores were in each tail of the distribution. As you can see in Figure 7.6, the distribution is still symmetrical and the mean is not significantly different from the original population mean (mean = -0.05, standard deviation = 1.28, skew = -0.03). In this case only the standard deviation is inflated because of added error variance caused by the extreme scores. This increase in error variance would have deleterious effects on any analyses you would want to perform if these extreme scores were not dealt with in some way.

Removing Univariate Extreme Scores

A simple way to handle this problem is to do a z transformation, converting all scores in a distribution to a z (standard normal distribution by subtracting the mean from each score and dividing by the standard deviation) distribution, which has a mean of 0.00 and standard deviation of 1.0, something most modern statistical packages can do automatically. You can then select cases with scores greater than -3.0 and less than 3.0 (or another cutoff point of your choosing) and continue analyses.

[p. 156 ↓]

Effect of Extreme Scores on Inferential Statistics

Dr. Overbay and I also demonstrated the effects of extreme scores on the accuracy of parameter estimates and Type I and Type II error rates in analyses involving continuous variables such as correlation and regression, as well as discrete variable analyses such as t-tests and ANOVA.[3]

In order to simulate a real study in which a researcher samples from a particular population, we defined our population as the 23,396 subjects with complete data on all variables of interest in the NELS 88 data file (already introduced earlier in the book).[4] For the purposes of the analyses reported below, this population was sorted into two groups: “normal” individuals whose scores on relevant variables were between $z = -3.0$ and $z = 3.0$, and “extreme scores,” who scored at least $z = \pm 3.0$ on one of the relevant variables.

To simulate the normal process of sampling from a population, but standardize the proportion of extreme scores in each sample, one hundred samples of $N = 50$, $N = 100$, and $N = 400$ each were randomly sampled (with replacement between each samples but not during the creation of a single sample) from the population of normal subjects.

Then an additional 4% were randomly selected from the separate pool of extreme scores, bringing samples to $N = 52$, $N = 104$, and $N = 416$, respectively. This procedure produced samples that simulate samples that could easily have been drawn at random from the full population, but that ensure some small number of extreme scores in each sample for the purposes of our demonstration.

The following variables were calculated for each of the analyses below.

- *Accuracy* was assessed by checking whether the original statistics or cleaned statistics were closer to the population correlation. In these calculations the absolute difference was examined.
- *Error rates* were calculated by comparing the outcome from a sample to the outcome from the population. An error of inference was considered to have occurred if a particular sample yielded a different conclusion than was warranted by the population.

Effect of Extreme Scores on Correlations and Regression

The first example looks at simple zero-order correlations. The goal was to demonstrate the effect of extreme scores on two different types of correlations: **[p. 157 ↓]** correlations close to zero (to demonstrate the effects of extreme scores on Type I error rates) and correlations that were moderately strong (to demonstrate the effects of extreme scores on Type II error rates). Toward this end, two different correlations were identified for study in the NELS 88 data set: the correlation between locus of control and family size (“population” $\rho = -.06$), and the correlation between composite achievement test scores and socioeconomic status (“population” $\rho = .46$). Variable distributions were examined and found to be reasonably normal.

After all samples were drawn, correlations were calculated in each sample, both before removal of extreme scores and after. For our purposes, $r = -.06$ was not significant at $p < .05$ for any of the sample sizes, and $r = .46$ was significant at $p < .05$ for all sample sizes. Thus, if a sample correlation led to a decision that deviated from the “correct” state of affairs, it was considered an error or inference.

As Table 7.1 demonstrates, extreme scores had adverse effects upon correlations. In all cases, removal of extreme scores had significant effects on the magnitude of the correlations, and the cleaned correlations were more accurate (i.e., closer to the known “population” correlation) 70% to 100% of the time. Further, in most cases, errors of inference were significantly less common with cleaned than uncleaned data.

As Figure 7.7 shows, a few randomly chosen extreme scores in a sample of 100 can cause substantial misestimation of the population correlation. In the sample of almost 24,000 students, these two variables were correlated very strongly, $r = .46$. In this particular sample, the correlation with four extreme scores in the analysis was $r = .16$ and was not significant. If this was your study, and you failed to deal with extreme scores, you would have committed a Type II error asserting no evidence of an existing relationship when in fact there is a reasonably strong one in the population.

Removing Extreme Scores in Correlation and Regression

Merely performing univariate data cleaning is not always sufficient when performing statistical analyses, as bivariate and multivariate extreme scores are often in the normal range of one or both of the variables, so merely converting variables to z scores and selecting the range $-3.0 < z > 3.0$ may not work (as I mention above). In this type of analysis, a two-stage screening process is recommended. First, checking all univariate distributions for extreme scores before calculating a correlation or regression analysis should be done automatically. As you can see in Figure 7.8, after all extreme univariate

[p. 158 ↓]

Table 7.1 The Effects of Extreme Scores on Correlations

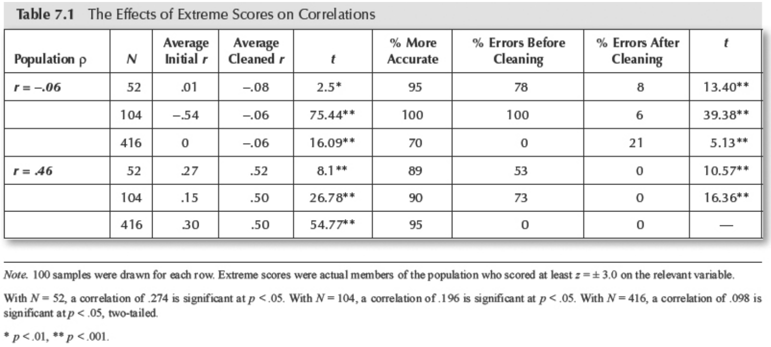


Table 7.1 The Effects of Extreme Scores on Correlations

Population ρ	N	Average Initial r	Average Cleaned r	t	% More Accurate	% Errors Before Cleaning	% Errors After Cleaning	t
$r = -.06$	52	.01	–.08	2.5*	95	78	8	13.40**
	104	–.54	–.06	75.44**	100	100	6	39.38**
	416	0	–.06	16.09**	70	0	21	5.13**
$r = .46$	52	.27	.52	8.1**	89	53	0	10.57**
	104	.15	.50	26.78**	90	73	0	16.36**
	416	.30	.50	54.77**	95	0	0	—

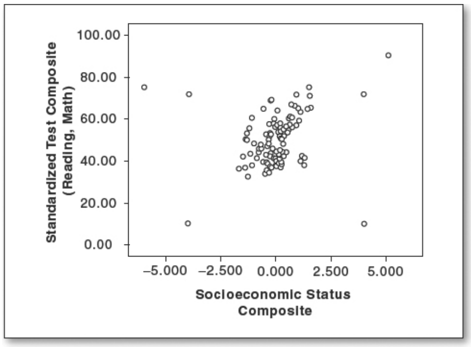
Note. 100 samples were drawn for each row. Extreme scores were actual members of the population who scored at least $z = \pm 3.0$ on the relevant variable.

With $N = 52$, a correlation of .274 is significant at $p < .05$. With $N = 104$, a correlation of .196 is significant at $p < .05$. With $N = 416$, a correlation of .098 is significant at $p < .05$, two-tailed.

* $p < .01$, ** $p < .001$.

[p. 159 ↓]

Figure 7.7 Correlation of SES and Achievement, 4% Extreme Scores

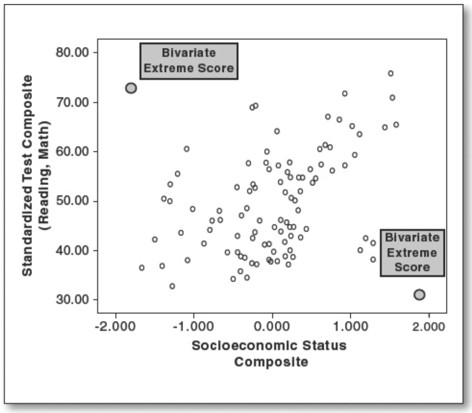


extreme scores were removed, some bivariate extreme scores remain. Most statistical programs allow you to save various statistics when you perform an analysis such as a regression. You will see several different types of residuals and many types of statistics. For simplicity, let us talk about two particular types: standardized residuals and distance indexes.

If you know what a residual is (the difference between the actual value of Y and the predicted value of Y from the analysis; also it can be conceptually defined as the vertical distance a data point is from the regression line), then a standardized residual is easy. It is essentially the z score of the residual and can be interpreted the same way as a univariate z score (e.g., higher numbers mean you are farther from the regression line, and standardized residuals outside the ± 3.0 range should be viewed suspiciously).[5]

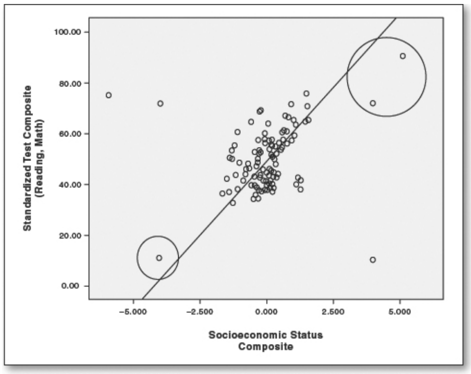
Additionally, bivariate and multivariate extreme scores can exist in multiple directions (not just vertically from the regression line), but standardized residuals only identify scores that fall far from the regression line in a vertical direction. Thus, going back to our example in Figure 7.7, in Figure 7.9 the data points that are clearly extreme scores but are not vertically separated from the regression line (circled) would *not* be detected by examining standardized **[p. 160 ↓]** residuals as they are very near the regression line. Thus, while visual inspection is helpful, particularly with simple analyses containing only two variables, once we get past two variables we need other indices, especially as we get beyond two-dimensional space into multiple regression.

Figure 7.8 Correlation of SES and Achievement, Bivariate Extreme Scores Remain After Univariate Outliers Removed



Indices of distance, such as Mahalanobis distance and Cook's distance, attempt to capture distance in more than one direction. While the details of their computations are beyond the scope of this chapter, imagine there is a center to the large group of data points in the middle of the scatterplot in Figure 7.9. As discussed above, the Mahalanobis distance and Cook's distance attempt to quantify distance from the center of the multivariate distribution and would likely pick up these extreme scores as being very far from the center of where most data points are, even though they are not vertically separated **[p. 161 ↓]** from the regression line. Using these indices to help with extreme score identification is relatively simple. Since statistical packages save these values as a separate variable, you can easily select or remove cases based on these scores.

Figure 7.9 Extreme Scores Not Detected by Standardized Residuals



Note that the same process of decision making we covered in the previous discussion of extreme scores should apply here as well—a case might be a multivariate extreme score for many reasons, some of which are legitimate and interesting and some not. You need to decide on an individual basis for each analysis and data point how to handle them, with the same options (removal, separate study, truncation or recoding, transformation, correction, and so on) available.

Effect of Extreme Scores on T-Tests and Anovas

The second example deals with analyses that look at group mean differences, such as t-tests and ANOVA. For the purpose of simplicity, I used t-tests for this example, but these results easily generalize to more complex ANOVA-type **[p. 162 ↓]** analyses. For these analyses, two different conditions were examined: when there were no significant differences between the groups in the population (sex differences in socioeconomic status produced a mean group difference of 0.0007 with a standard deviation of 0.80 and with 24,501 *df* produced a *t* of 0.29, which is not significant at *p* < .05) and when there were significant group differences in the population (sex differences in mathematics achievement test scores produced a mean difference of 4.06 and standard deviation of 9.75 and 24,501 *df* produced a *t* of 10.69, *p* < .0001). For both analyses, the effects of having extreme scores in only one cell as compared to both cells were examined. Distributions for both dependent variables were examined and found to be reasonably normal.

Similar to the previous set of analyses, in this example, t-tests were calculated in each sample, both before removal of extreme scores and after. For this purpose, t-tests looking at SES should not produce significant group differences, whereas t-tests looking at mathematics achievement test scores should. Two different issues were examined: mean group differences and the magnitude of the *t*. If an analysis from a sample led to a different conclusion than expected from the population analyses, it was considered an error of inference.

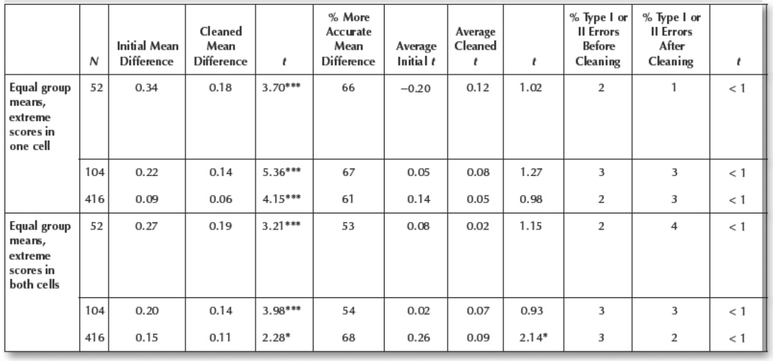
The results in Table 7.2 illustrate the unfortunate effects of extreme scores on ANOVA-type analyses, and they again highlight the importance of including this step in your routine data cleaning regimen. Removal of extreme scores produced a significant change in the mean differences between the two groups when there were no significant group differences expected, but tended not to when there were strong group differences (as these group differences were very strong to begin with). Removal of extreme scores produced significant change in the *t* statistics primarily when there were strong group differences. In both cases the tendency was for both group differences and *t* statistics to become more accurate in a majority of the samples. Interestingly, there was little evidence that extreme scores produced Type I errors when group means were equal, and thus removal had little discernable effect. But when strong group differences were revealed, extreme score removal tended to have a significant beneficial effect on error rates, although not as substantial an effect as seen in the correlation analyses.

The presence of extreme scores appears to produce similar effects regardless of whether they are concentrated only in one cell or are present in both.

[p. 163 ↓]

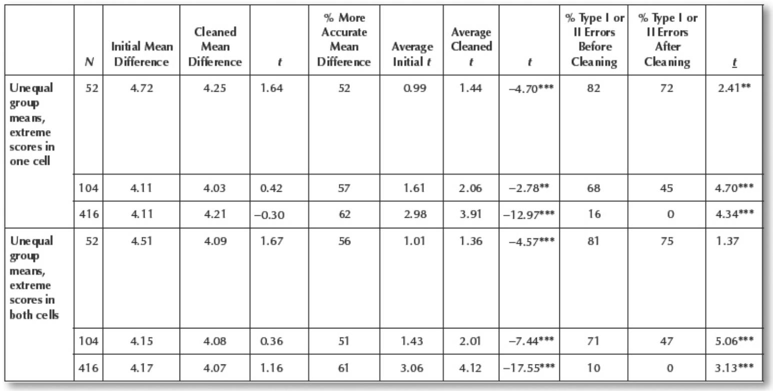
Table 7.2 The Effects of Extreme Scores on ANOVA-Type Analyses

Table 7.2 The Effects of Extreme Scores on ANOVA-Type Analyses



	N	Initial Mean Difference	Cleaned Mean Difference	*t*	% More Accurate Mean Difference	Average Initial *t*	Average Cleaned *t*	*t*	% Type I or II Errors Before Cleaning	% Type I or II Errors After Cleaning	*t*
Equal group means, extreme scores in one cell	52	0.34	0.18	3.70***	66	−0.20	0.12	1.02	2	1	< 1
	104	0.22	0.14	5.36***	67	0.05	0.08	1.27	3	3	< 1
	416	0.09	0.06	4.15***	61	0.14	0.05	0.98	2	3	< 1
Equal group means, extreme scores in both cells	52	0.27	0.19	3.21***	53	0.08	0.02	1.15	2	4	< 1
	104	0.20	0.14	3.98***	54	0.02	0.07	0.93	3	3	< 1
	416	0.15	0.11	2.28*	68	0.26	0.09	2.14*	3	2	< 1

[p. 164 ↓]



	N	Initial Mean Difference	Cleaned Mean Difference	t	% More Accurate Mean Difference	Average Initial t	Average Cleaned t	t	% Type I or II Errors Before Cleaning	% Type I or II Errors After Cleaning	t
Unequal group means, extreme scores in one cell	52	4.72	4.25	1.64	52	0.99	1.44	−4.70***	82	72	2.41**
	104	4.11	4.03	0.42	57	1.61	2.06	−2.78**	68	45	4.70***
	416	4.11	4.21	−0.30	62	2.98	3.91	−12.97***	16	0	4.34***
Unequal group means, extreme scores in both cells	52	4.51	4.09	1.67	56	1.01	1.36	−4.57***	81	75	1.37
	104	4.15	4.08	0.36	51	1.43	2.01	−7.44***	71	47	5.06***
	416	4.17	4.07	1.16	61	3.06	4.12	−17.55***	10	0	3.13***

Note. 100 samples were drawn for each row. Extreme scores were actual members of the population who scored at least $z = \pm 3.0$ on the relevant variable.

*$p < .05$, ** $p < .01$, *** $p < .001$

[p. 165 ↓]

Detecting Extreme Scores in ANOVA-Type Analyses

Similar to regression type analyses, ANOVA-type analyses can contain bivariate or multivariate extreme scores not removed by simple univariate data cleaning. As mentioned previously, most statistical packages will save standardized residuals, which allows for identification of these types of extreme scores. In the case of ANOVA-type analyses, the residual is the difference between an individual score and the group mean, and by standardizing it, the same ± 3.0 standard deviation rule can apply.

To Remove or Not to Remove?

Some authors have made the argument that removal of extreme scores produces undesirable outcomes, such as making analyses less generalizable or representative of the population. I hope that this chapter persuades you that the opposite is in fact true: that your results probably will be more generalizable and less likely to represent an error of inference if you do conscientious data cleaning, including dealing with extreme scores where warranted (remember, there are many possible reasons for extreme scores, and the reason for them should inform the action you take). In univariate analyses, the cleaned data are closer to our example population than any sample with extreme scores—often by a substantial margin. In correlation and regression and in ANOVA-type analyses, my colleague and I demonstrated several different ways in which statistics and population parameter estimates are likely to be *more* representative of the population after having addressed extreme scores than before.

Though these were two fairly simple statistical procedures, it is straightforward to argue that the benefits of data cleaning extend to more complex analyses. More sophisticated analyses, such as structural equation modeling, multivariate analyses, and multilevel modeling, tend to have more restrictive and severe assumptions, not fewer, because they tend to be complex systems. Thus, it is good policy to make sure the data are as clean as possible when using more complex analyses. Ironically, even analyses designed to be robust to violations of distributional assumptions, such as nonparametric procedures, seem to benefit from solid, more normally distributed data.

For Further Enrichment

- Data sets from the examples given in this chapter are available online on this book's website. Download some of the examples yourself and see how removal of outliers generally makes results more generalizable and closer to the population values.
- Examine a data set from a study you (or your advisor) have previously published for extreme scores that may have distorted the results. If you find any relatively extreme scores, explore them to determine if it would have

been legitimate to remove them, and then examine how the results of the analyses might change as a result of removing those extreme scores. And if you find something interesting, be sure to share it with me. I enjoy hearing stories relating to real data.

- Explore well-respected journals in your field. Note how many report having checked for extreme scores, and if they found any, how they dealt with them and what the results of dealing with them were (if reported). In many of the fields I explored, few authors explicitly discussed having looked for these types of issues.

Notes

1. Data comes from the Health and Nutrition Examination Survey (HANES), performed by the U.S. Centers for Disease Control and Prevention (CDC).

2. Researchers (Miller, 1991; Van Selst & Jolicoeur, 1994) demonstrated that simply removing scores outside the ± 3.0 standard deviations can produce problems with certain distributions, such as highly skewed distributions characteristic of response latency variables, particularly when the sample is relatively small. If you are a researcher dealing with this relatively rare situation, Van Selst and Jolicoeur (1994) present a table of suggested cutoff scores for researchers to use with varying sample sizes that will minimize these issues with extremely nonnormal distributions. Another alternative would be to use a transformation to normalize the distribution prior to examining data for extreme scores.

3. Some readers will recognize that both regression and ANOVA are examples of general linear models. However, as many researchers treat these as different paradigms and there are slightly different procedural and conceptual issues in extreme scores, we treat them separately for the purpose of this chapter.

4. This is a different number from the univariate examples as there are different numbers of missing data in each variable, and for these analyses we removed all cases with missing data on *any* variable of interest. For more information on more appropriate ways of handling missing data, be sure to refer to Chapter 6.

5. However, standardized residuals are not perfect. In some cases a *studentized* residual is more helpful (studentized residuals are standardized residuals that account for the fact that extreme scores can inflate standard errors, thus potentially masking extreme scores, particularly in small data sets).

[p. 167 ↓]

References

Anscombe, F. J. Rejection of outliers. Technometrics, vol. 2 (1960). (2), pp. 123–147.

Barnett, V. Lewis, T. (1994). Outliers in statistical data . New York: Wiley.

Dixon, W. J. Analysis of extreme values. Annals of Mathematical Statistics, vol. 21 (1950). (4), pp. 488–506.

Evans, V. P. (1999). Strategies for detecting outliers in regression analysis: An introductory primer . In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 213–233). Stamford, CT: JAI Press.

Hamilton, L. (1992). Regression with graphics: A second course in applied statistics : Belmont, CA: Duxbury Press.

Hawkins, D. M. (1980). Identification of outliers . New York: Chapman & Hall.

Iglewicz, B. Hoaglin, D. C. (1993). How to detect and handle outliers . Milwaukee, WI: ASQC Quality Press.

Impara, J. Kingsbury, G. Maynes, D. Fitzgerald, C. (2005, April). Detecting cheating in computer adaptive tests using data forensics . Paper presented at the annual meeting of the National Council on Measurement in Education and National Association of Test Directors, Montreal, Canada.

Jarrell, M. G. A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. Research in the Schools, vol. 1, (1994). pp. 49–58.

Judd, C. M. McClelland, G. H. (1989). Data analysis: A model comparison approach . San Diego, CA: Harcourt Brace Jovanovich.

Lane, K. (2002, February). What is robust regression and how do you do it? Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Micceri, T. The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, vol. 105 (1989). (1), pp. 156–166.

Miller, J. Reaction time analysis with outlier exclusion: Bias varies with sample size. The Quarterly Journal of Experimental Psychology, vol. 43 (1991). (4), pp. 907–912.

Orr, J. M., Sackett, P. R., and DuBois, C. L. Z. Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. Personnel Psychology, vol. 44, (1991). pp. 473–486.

Osborne, J. W. Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. Educational Psychology, vol. 28 (2008). (2), pp. 1–10.

Osborne, J. W. and Overbay, A. The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation, vol. 9 (2004). (6), pp. 1–12.

Rasmussen, J. L. Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. Multivariate Behavioral Research, vol. 23 (1988). (2), pp. 189–202.

Rousseeuw, P. Leroy, A. (1987). Robust regression and outlier detection . New York: Wiley.

Rowland-Jones, S., Sutton, J., Ariyoshi, K., Dong, T., Gotch, F., McAdam, S., and Corrah, T. HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women. Nature Medicine, vol. 1 (1995). (1), pp. 59–64.

Sachs, L. (1982). Applied statistics: A handbook of techniques (2nd ed.). New York: Springer-Verlag.

Schwager, S. J. and Margolin, B. H. Detection of multivariate normal outliers. The Annals of Statistics, vol. 10 (1982). (3), pp. 943–954.

Stevens, J. P. Outliers and influential data points in regression analysis. Psychological Bulletin, vol. 95 (1984). (2), pp. 334–344.

Van Selst, M. and Jolicoeur, P. A solution to the effect of sample size on outlier elimination. The Quarterly Journal of Experimental Psychology, vol. 47 (1994). (3), pp. 631–650.

Wainer, H. Robust statistics: A survey and some prescriptions. Journal of Educational Statistics, vol. 1 (1976). (4), pp. 285–312.

Wilcox, R. (2008). Robust methods for detecting and describing associations . In J. W. Osborne (Ed.), Best practices in quantitative methods (pp. 263–279). Thousand Oaks, CA: Sage.

Yuan, K.-H., Bentler, P. M., and Zhang, W. The effect of skewness and kurtosis on mean and covariance structure analysis. Sociological Methods & Research, vol. 34 (2005). (2), pp. 240–258.

Zimmerman, D. W. A note on the influence of outliers on parametric and non-parametric tests. Journal of General Psychology, vol. 121 (1994). (4), pp. 391–401.

Zimmerman, D. W. Increasing the power of nonparametric tests by detecting and downweighting outliers. Journal of Experimental Education, vol. 64 (1995). (1), pp. 71–78.

Zimmerman, D. W. Invalidation of parametric and nonparamteric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education, vol. 67 (1998). (1), pp. 55–68.

http://dx.doi.org/10.4135/9781452269948.n7