Contents

Preface

The goal of statistical data analysis is to use data to gain and communicate knowledge about processes and phenomena. Comparing means is often part of an analysis, for data arising in both experimental and observational studies. Probably the most common method used to compare the means of several different treatments (or, more loosely, groups arising from stratification) is the analysis of variance (ANOVA). The analysis of means (ANOM) is an alternative procedure for comparing means. While it cannot be used in all the same settings as the ANOVA, when one is specifically interested in comparing means, such as when looking at fixed main effects in a designed experiment, ANOM has the advantages of being much more intuitive and providing an easily understood graphical result, which clearly indicates any means that are different (from the overall mean) and allows for easy assessment of practical as well as statistical significance. The graphical result is easy for nonstatisticians to understand and offers a clear advantage over ANOVA in that it sheds light on the nature of the differences among the populations.

There have been a number of advances in ANOM procedures in the last 20 years, but many of these results have appeared in fairly technical papers. ANOM is actually a multiple comparisons procedure, and the theory behind it is more complicated than that for ANOVA. Rather than dealing with univariate F distributions, one ends up with multivariate negatively correlated singular t distributions. However, the necessary critical values, power curves, and sample sizes for the ANOM procedures have already been obtained, documented and, in some instances, included in statistical software, resulting in methods that are easy to apply and with results that are easy to interpret.

Our intent in writing this book was to present the first modern comprehensive treatment of ANOM containing the information necessary for comparing means using ANOM. The book is intended to be a useful guide for practitioners, not a detailed description of the theory behind the procedures. Only as much theory as was necessary to understand and implement the various ANOM techniques is included. Most of the applications of ANOM that have appeared in the literature are from the physical sciences and engineering. However, ANOM techniques are much more broadly applicable; thus, we have included many examples from other areas, including business, medicine, health care, quality control, and the social sciences. Note that the comparison of means is used in a rather broad sense in that it also includes the comparison of Poisson rates and binomial proportions.

The audience for this book includes quality and process engineers, medical and health care investigators, social scientists, biostatisticians, epidemiologists, and scientists who may work in government, business, or education. The intended uses for this book include as a comprehensive reference for practitioners; as a text in a topics course in biostatistics, engineering statistics, industrial engineering, or business statistics; and as a supplementary text in a design of experiments course or a general course in statistical methods, health statistics, epidemiology, or biostatistics. This book is being used as a supplementary text in an introductory graduate course in statistics for health professionals, and portions of the material in this book are being used in a series of lectures given to researchers at a medical university. In addition, material in this book has been successfully used in an undergraduate course in statistical methods.

Now that ANOM is included as a standard option in some statistical software (including SAS® and MINITAB®), we anticipate the use of ANOM to expand. While not software dependent, this book includes several examples using SAS and an appendix of SAS examples that will make it easy for practitioners to implement ANOM analyses for most settings that arise in practice. While we have included many carefully worked examples that can serve as templates for practitioners who might choose to work solutions by hand, readily available software can be used to do all but the final computations; furthermore, while not explicitly illustrated, spreadsheet programs such as EXCEL™ can readily be used to perform all the calculations as well as to create ANOM decision charts.

We start with the simplest situation, in which one is interested in comparing means associated with changes in the levels of a single factor, and continue with more complicated design situations. Analysis of single-factor experiments based on the usual assumptions of normality and constant variances is covered in Chapters 2 and 3. Chapter 4 describes how to use ANOM-type techniques to test the assumption of constant variances. These chapters, together with Chapter 8, which covers analysis of normal data with nonconstant variances, and Chapter 9, which discusses distribution-free techniques, should be of interest to anyone who has means to compare. Chapter 5 discusses the ANOM for complete factorial designs, and Chapters 6 and 7 cover the ANOM for the more specialized settings of incomplete designs and axial mixture designs.

We would like to thank Robert Rodriguez for his guidance on this project and the reviewers for their insights, and most of all we wish to thank Andi Nelson for allowing us to bring this work to completion.

Peter S. Wludyka
Karen A. F. Copeland

Chapter 1

Introduction

The analysis of means (ANOM) is a graphical procedure for comparing a collection of means, rates, or proportions to see if any of them are significantly different from the overall mean, rate, or proportion. An ANOM decision chart is similar in appearance to a control chart. It has a centerline, located at the overall mean (rate or proportion), and upper and lower decision limits. The group means (rates or proportions) are plotted, and those that fall beyond the decision limits are said to be significantly different from the overall value. These differences are statistical differences, if they exist. The chart also allows one to easily evaluate the practical differences. For example, from the ANOM chart in Figure 1.1, one can conclude that the rate of office visits per member for clinic C (about 0.17 visits per member year) was significantly lower and the rate of visits for clinic A (about 0.218 visits per member year) was significantly higher than the overall rate of office visits for all clinics run by an HMO in a metropolitan area.

In circumstances in which one might use ANOVA to analyze fixed main effects, ANOM is appropriate and generally produces a more useful result. While ANOM can be used to study interactions, its main advantages occur when it is used to study main effects. When studying main effects, ANOM has two advantages over ANOVA: (1) if any of the treatments are statistically different, ANOM indicates exactly which ones are different; and (2) ANOM can be presented in a graphical form, which allows one to easily evaluate both the statistical and the practical significance of the differences.

1.1 ANOM as a Multiple Comparison Procedure

ANOM and ANOVA are only two of many ways to compare a group of means. When one is comparing exactly two means, then often the Student's t-test is used. ANOM is a graphical form of this test. For more than two means, a commonly used technique is the Tukey–Kramer (TK) procedure (Tukey (1953), Kramer (1956)) for comparing all pairwise differences of the means. There are many other multiple comparison procedures that could be used to compare means (see, e.g., Hsu (1996) or Hochberg and Tamhane (1987)). Each procedure approaches the comparison of means differently. That is, there are differences regarding exactly what is being compared. For example, ANOM compares each mean to the

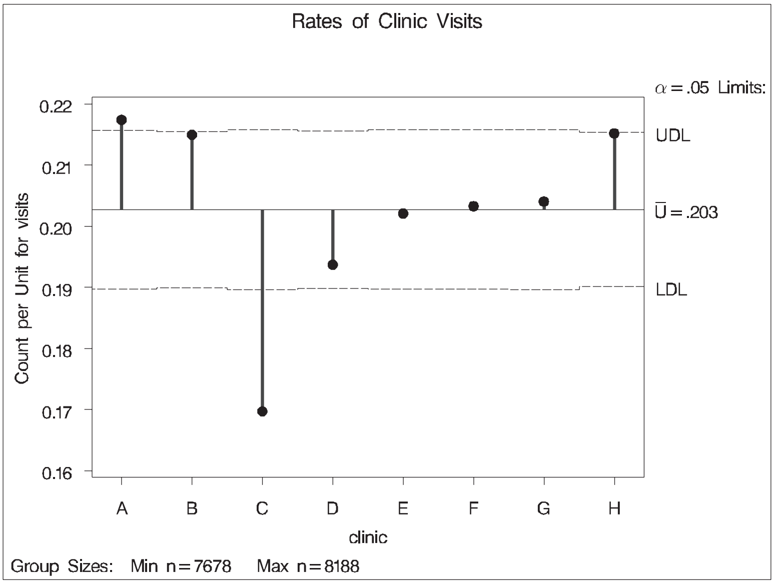

Figure 1.1. *ANOM Chart for Rates of Office Visits.*

Table 1.1. *Sample Means and Variances for Paint Drying Times (Example* 1.1*).*

	Paint Type			
	1	2	3	4
$\overline{y}$	6.88	9.28	9.00	9.90
s^2	1.72	1.85	2.81	1.95

overall mean, while the TK technique considers pairwise differences between the means. We examine the ANOVA, TK, and ANOM procedures in detail in the following example.

Example 1.1 (Paint Drying Data). Consider the following simple example in which one is interested in comparing the drying times (in hours) of four different types of paint used on park benches. Four benches were painted with each of the four types of paint. Summary statistics for the drying times are given in Table 1.1.

ANOVA tests for the equality of means indirectly by comparing estimates of variability that depend on the mean values. Using the ANOVA to test for differences in the drying times, one would compute the mean squares

$$\text{MS}_A = 4\{\text{sample variance of the } \overline{y}\text{s}\} = 4(1.725) = 6.9$$

and

$$\text{MS}_e = \frac{1.72 + 1.85 + 2.81 + 1.95}{4} = 2.08,$$

Table 1.2. *ANOVA Table for the Paint Drying Times (Example* 1.1*).*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20.70250000	6.90083333	3.31	0.0572
Error	12	25.01500000	2.08458333		
Corrected Total	15	45.71750000			

where MS_A is an estimate of the variability, assuming the group means are equal, and MS_e is an estimate of variability regardless of the equality of the group means. The ratio of these two mean squares results in the test statistic

$$F = \frac{6.9}{2.08} = 3.32.$$

If there is not a significant difference in the group means, then the two measures of variability will be similar and F will be close to one. To determine statistical significance one would compare the value of F with the upper α quantile from the appropriate F distribution. Since $3.32 < F(0.05; 3, 12) = 3.49$, no significant differences are found in the drying times at level $\alpha = 0.05$. The ANOVA procedure is generally summarized in an ANOVA table, such as Table 1.2. From this table we obtain a p-value of 0.0572 for the test of equality of means. Since $\alpha = 0.05 < 0.0572$ we do not reject the hypothesis of equal means at the $\alpha = 0.05$ level.

The TK procedure considers all pairwise comparisons between group means. Using the TK procedure, one would compute simultaneous confidence intervals

$$\overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm q(\alpha; I, \nu)\sqrt{\mathrm{MS}_e}/\sqrt{n},$$

where $q(\alpha; I, \nu)$ is the upper α quantile from a Studentized range distribution. The two means $\overline{y}_{i\bullet}$ and $\overline{y}_{j\bullet}$ are declared to be significantly different if the interval does not contain zero or, alternatively, if

$$|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}| > q(\alpha; I, \nu)\sqrt{\mathrm{MS}_e}/\sqrt{n}. \tag{1.1}$$

In our example,

$$q(0.05; 4, 12)\sqrt{\mathrm{MS}_e}/\sqrt{n} = 4.20\sqrt{2.08}/\sqrt{4} = 3.03,$$

and the largest difference in means is $|\overline{y}_{1\bullet} - \overline{y}_{4\bullet}| = |6.88 - 9.90| = 3.02$. Thus, none of the pairs of means are different at level $\alpha = 0.05$. Table 1.3 provides computer output for the TK procedure.

Using the ANOM (details are in Chapter 2), one would compute decision lines

$$\begin{aligned} \overline{y}_{\bullet\bullet} &\pm h(0.05; 4, 12)\sqrt{\mathrm{MS}_e}\sqrt{\frac{3}{16}}, \\ 8.76 &\pm 2.85\sqrt{2.08}\sqrt{\frac{3}{16}} \\ &\pm 1.78 \\ &(6.98, 10.54). \end{aligned}$$

Table 1.3. *TK Output for the Paint Drying Times (Example* 1.1*).*

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	2.084583
Critical Value of Studentized Range	4.19852
Minimum Significant Difference	3.0309

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	type
A	9.900	4	4
A	9.275	4	2
A	9.000	4	3
A	6.875	4	1

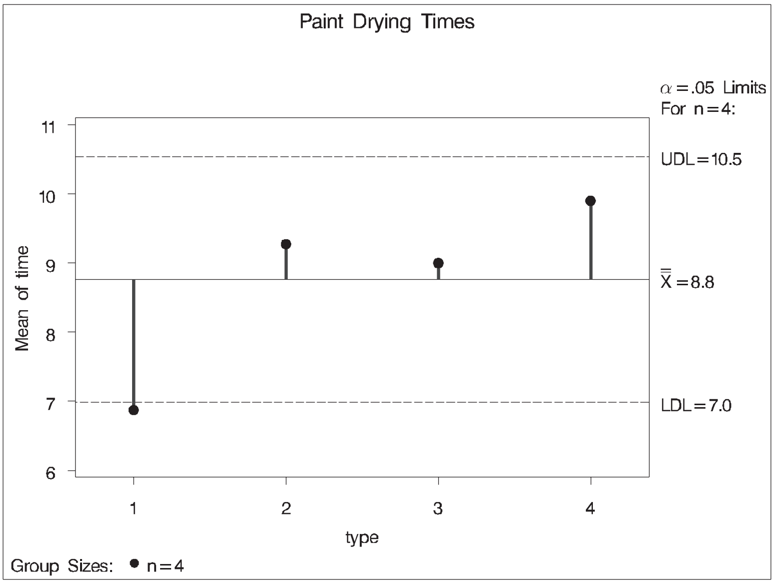

Figure 1.2. *ANOM Chart for Drying Times of Four Types of Paint (Example* 1.1*).*

From the corresponding ANOM chart in Figure 1.2, one concludes that there are significant ($\alpha = 0.05$) differences in the means because paint type 1 has a drying time that is significantly shorter than the overall average. Further, one might conclude that the difference from the overall average drying time of nearly 2 hours ($8.76 - 6.88 = 1.88$ hours) is of practical importance. Of the three procedures in this example, many practitioners will find the ANOM decision chart the easiest of the results to interpret in terms of both statistical and practical significance.

One might wonder what conclusions to draw from the fact that the three approaches to the paint example (ANOVA, TK, and ANOM) produced somewhat different results. What accounts for the different conclusions a researcher might draw? Typically, each of these are preplanned procedures and should be motivated by the purposes of the investigation. That is, what questions in particular are of interest? The three procedures just examined have somewhat different purposes and interpretations in the context of multiple comparisons. The fact that the ANOVA F test had a p-value greater than 0.05 implies that there is no Scheffé-type contrast (or set of multiple comparisons) that is significant. In particular, this implies that the set of comparisons (contrasts) in which the average for each paint is compared to the overall average is not significant at $\alpha = 0.05$ using Scheffé's multiple comparison procedure. This comparison can be made using a pair of decision lines similar to those in Figure 1.2; however, the decision lines based on Scheffé's method are wider than those used in ANOM because the Scheffé decision limits do not take into account the specific correlation structure implied by this particular set of comparisons. (Note that this structure is not a concern for ANOM users since this has been taken into account in the tables and software used in ANOM.)

ANOM uses decision lines associated with this particular set of comparisons and has associated with it exactly the level of significance specified by the test (in this case, exactly four comparisons with simultaneous significance $\alpha = 0.05$). That is, it is specifically designed to compare a group of means to the overall mean.

The TK set of pairwise comparisons is specifically designed to simultaneously test pairwise comparisons and hence is the sharpest test available for this situation. The TK ruler (the critical distance between pairs of means, which is the right-hand side of (1.1)) indicates how far apart the sample means must be to signify significance. This TK ruler typically will be less than the width (difference between the upper decision line and the lower decision line) of the ANOM decision chart. Hence, whenever at least one mean plots below the lower decision line and at least one point plots above the upper decision line, there will be at least one significant pairwise difference using TK. The central point is to use TK when pairwise comparisons will properly answer the research question.

Two Additional ANOM Examples

The following two examples illustrate the flexibility and usefulness of ANOM by showing that binomial count data and Poisson rate data can be analyzed with ANOM. These examples also illustrate how ANOM can be used in observational studies and how ANOM often provides answers to key research questions.

Example 1.2 (Epidemiological Data). A large children's clinic at a teaching hospital conducted a retrospective study of the prevalence (proportion of individuals in the population with the characteristic of interest) of obesity in the population of children they serve (predominately low-income) to determine how to package a nutritional education program. Records for the last 2 years were used to calculate the age- and sex-adjusted body mass index (BMI) percentiles for 535 children. The data were stratified by sex and race/ethnicity into six categories corresponding to sex (male or female) and race/ethnicity (black, white,

Table 1.4. *Epidemiological Study of Obesity Data (Example* 1.2*).*

Sex	Race	At Risk	Sample Size	Prevalence = $\hat{p}_i$
male	black	25	150	0.167
male	white	7	107	0.065
male	other	8	38	0.210
female	black	55	115	0.478
female	white	10	50	0.200
female	other	15	75	0.200

other) combinations. A child at or above the 85th percentile was classified as at risk for obesity. Summary data are given in Table 1.4. The research question of interest is whether the prevalence of those at risk for obesity is the same for the six strata.

One method of analysis is to perform a Pearson chi-squared test for equal prevalence, which would lead to the rejection of the hypothesis at the $\alpha = 0.01$ level. However, this sheds no light on the nature of the differences. Alternatively (or subsequently), one could examine the 15 pairwise differences, adjusting the level of significance to take into account the number of comparisons being made using, for example, the Marascuilo procedure (Marscuilo and Levin (1983)). This may not actually answer the central question since this focuses on comparing between the strata rather than among all strata, and, in addition, the large number of simultaneous comparisons reduces the power of the test.

Using the ANOM one obtains the ANOM decision chart in Figure 1.3. Prevalence for a group is judged to be different from the overall prevalence if the estimated prevalence for that group plots outside the ANOM decision lines. The decisions lines have different widths corresponding to the different sample sizes associated with each strata (wider for small samples and narrower for large samples; see Chapter 3 for details). From the decision chart one can conclude at $\alpha = 0.01$ that the prevalence of those at risk for obesity for black/female children is higher than the prevalence of those at risk for obesity in the overall clinic population. In addition, the prevalence of those at risk for obesity in the white/male population is lower than the prevalence of those at risk for obesity in the overall clinic population. Notice that due to the manner in which data were collected in this example, the overall average, $\bar{p} = (25 + 7 + 8 + 55 + 10 + 15)/535 = 0.224$, has a clear interpretation in this study as an estimate of the prevalence of at-risk children for the clinic population during the period under study. Comparing the strata (sex and race/ethnicity groups) to this has a useful interpretation and is probably more interesting than pairwise comparisons. The study strongly suggests that the nutritional educational piece be constructed to appeal to black females.

Example 1.3 (Tourism/Travel Coupon Data). A charter airline is interested in the manner in and extent to which travelers use coupons for discount opportunities at their destination. A particular traveler may use several coupons during a single travel experience. Data were collected by administering a survey to all passengers for a 2-week period. One research

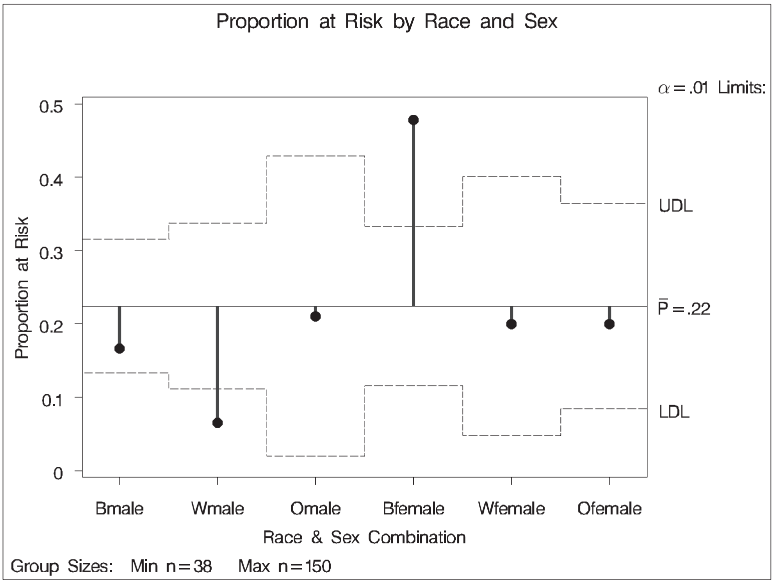

Figure 1.3. *ANOM Chart for the Obesity Data (Example* 1.2*).*

Table 1.5. *Coupon Use Data (Example* 1.3*).*

	Destination			
	Florida	Islands	New Orleans	Asheville
Passengers	525	1100	350	210
Coupons	250	505	50	260
Rate = $\hat{u}_i$	0.479	0.459	0.143	1.238

question asked whether destination affected coupon use. The survey data collected relevant to this question are summarized in Table 1.5, where the rate is the coupon use per passenger. Assuming that the rates are Poisson (see Section 3.3 for details), the ANOM decision chart for this data is shown in Figure 1.4. Since the area of opportunity (number of passengers) is different for the four destinations, the decision lines are different for each destination. In this chart, the rate for each destination is compared to the overall rate (e.g., for Florida, the rate is $\hat{u}_1 = 250/525 = 0.479$ coupons per passenger). The multiple comparison in this case consists of four comparisons, in which each of the destinations is compared to the overall rate $\bar{u} = (250 + 505 + 50 + 260)/(525 + 1100 + 350 + 210) = 0.49$, which is an estimate of the coupon use rate for all passengers. Coupon use by New Orleans passengers is significantly below average, and coupon use by those visiting Asheville is significantly higher than average.

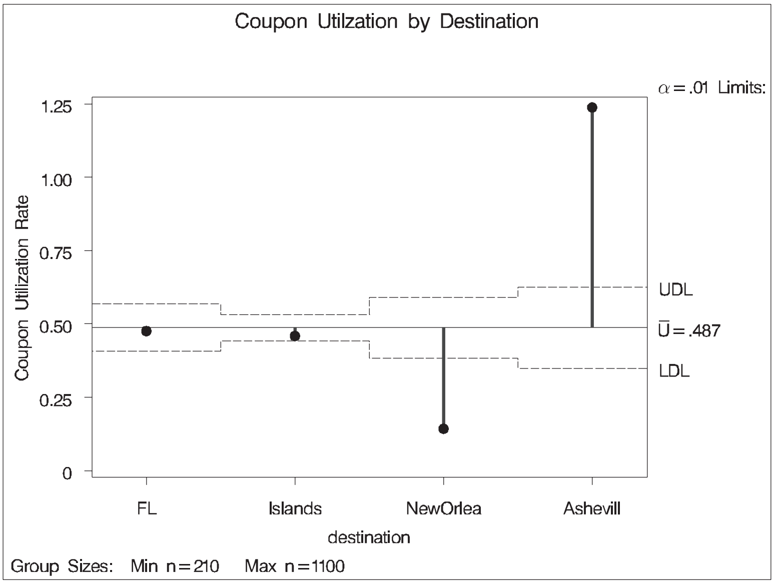

Figure 1.4. *ANOM Chart for the Coupon Use Data (Example 1.3).*

When is ANOM Most Useful?

ANOM is most useful for straightforward studies in which the desired outcome is to identify differences between groups or treatments. Similar to ANOVA, ANOM is a good choice when factor levels are clearly nominal (categorical). When factor levels are continuous, ANOM can still provide useful information; however, regression (response surface) models may be more appropriate. While ANOM has been used primarily in the analysis of experimental data, it can be very useful for analyzing observational data. This is especially true when there is post stratification, since in that circumstance the overall average estimates the population average and hence has a clear interpretation in the context of the problem (see Examples 1.2 and 1.3). Keep in mind that ANOM is a multiple comparison procedure and can be used in conjunction with other procedures. In a particular analysis, ANOM might be used in conjunction with another multiple comparison procedure, such as TK (one might wish to adjust the level of significance to control the experimentwise error rate), or in conjunction with other specific contrasts (e.g., suppose that in the coupon problem, Example 1.3, one wishes to compare tropical to nontropical destinations). A nice feature of ANOM is that the decision chart can be used to convey the conclusions arising from the data analysis. The ANOM decision chart is easy for nonstatisticians to understand. Furthermore, ANOM makes assessing practical significance easy.

1.2 History of ANOM

The basic idea of ANOM was first used by Laplace (1827), almost 100 years before Fisher (1918, 1925, 1935) introduced ANOVA. Laplace was interested in studying the homogeneity over the calendar year of the lunar atmospheric tide in Paris. He had available data on the mean change in barometric pressure from 9:00 a.m. to 3:00 p.m. over a period of 11 years. His analysis for homogeneity consisted of computing the average change for each season and comparing these with the average change over the entire year. He evaluated what we would now call the descriptive level of significance. While Laplace correctly concluded there were significant differences between the four seasons, he made what today would be considered two fundamental errors. First, he didn't account for the dependence among the four differences (seasonal averages minus the overall average), and second, rather than using the pooled seasonal variances as the measure of variability, he used the variance over the entire year.

The next appearance of an ANOM-type procedure was the multiple significance test of Halperin et al. (1955) for several normally distributed populations. They correctly used the pooled sample variances to measure variability, and to account for the dependence among the treatment means minus the overall mean, they used Bonferroni inequalities to obtain upper and lower bounds on the appropriate critical values. They conjectured that the exact critical values were closer to the lower bounds. Ott (1967) suggested applying the test of Halperin et al. (1955) in a graphical form and, based on their conjecture, provided tables of approximate critical values that were the average of the lower and upper Bonferroni bounds. Ott (1967) was the first to use the phrase "analysis of means" to describe this procedure.

Ott (1967) not only used approximate critical values but also advocated using sample ranges rather than sample variances to measure variability. Schilling (1973) extended Ott's work to designs more complicated than factorial designs with fixed effects (e.g., balanced incomplete block designs and mixed effect designs) and discussed using the ANOM with nonnormal data. Rather than using the approximate critical values proposed by Ott (1967), he used the upper bounds from Halperin et al. (1955). Schilling (1973) also advocated using sample ranges to measure variability. Following on Schilling's work, a number of authors discussed various aspects of the ANOM (Nelson (1974), Enrick (1976, 1981), Ohta (1981), Sheesley (1980, 1981)). All of this work continued to be based on conservative critical values obtained using Bonferonni's inequality.

In 1982, exact ANOM critical values for the main effects of ANOM in balanced designs were obtained (see Nelson (1982)). These exact values were based on the variability being estimated using the pooled sample variances. The entire January 1983 issue of the *Journal of Quality Technology* was devoted to ANOM and summarized the state of the art at that time. Since then there have been a number of advances in the ANOM technique. It has been shown that the exact critical values are appropriate not only for main effects from complete balanced designs but also for Latin squares, balanced incomplete block designs, Youden squares, and axial mixture designs (see Nelson (1993)). Exact critical values for main effects when the sample sizes are not equal are now available (see Nelson (1991) and Soong and Hsu (1997)), and for situations in which one doesn't want to go to the trouble of computing the exact critical values for a particular set of unequal sample sizes, conservative critical values, which are less conservative than those obtained using Bonferonni's inequality, are available (see Nelson (1989)).

Power curves for ANOM are now available (see Nelson (1985)), the ANOM technique has been extended to the case of unequal variances (see Nelson and Dudewicz (2002) and Dudewicz and Nelson (2003)), nonparametric ANOM procedures are available (see Bakir (1989) and Nelson (2002)), and the ANOM technique has even been extended to comparison of variances (see Wludyka and Nelson (1997a, 1997b, 1999) and Wludyka et al. (2001)).

Today the ANOM procedure can be found in statistical software such as SAS and MINITAB. The availability of such software has moved ANOM beyond a procedure that relies on specialized look-up tables to a technique that is easily applied in fields from engineering to managed health care.

1.3 This Book

ANOM is actually a multiple comparisons procedure (each treatment mean is compared to the overall mean), and it is only one of many possible multiple comparison procedures that could be used instead of, or as a follow up to, ANOVA. Why choose ANOM? It is easy to understand and apply, its results can be presented in a graphical form (which makes them easy to explain to those not well versed in statistics), and for practical purposes there is no disadvantage in terms of loss of power when compared with the ANOVA (see Nelson (1983a)).

One can find reference to the ANOM in a number of books (e.g., Mason et al. (1989), Farnum (1994), Vardeman (1994), Wheeler (1995), Hsu (1996), Freund and Wilson (1997), Ott et al. (2000), Ryan (2000), Nelson et al. (2003)), but there is no one place to which a practitioner can go to find simple explanations of a wide variety of the possible uses of ANOM and the necessary tables and charts. This book is intended to fill that void and presents the first modern comprehensive treatment of ANOM. While the majority of examples in the literature are of applications to problems in engineering and the physical sciences, ANOM is of much broader use. To illustrate this broad applicability, we include examples from areas such as health care, tourism, and business in addition to engineering and physical science. Further, where applicable, SAS will be used to create the ANOM charts shown in this text, and an appendix will cover the syntax needed to run the ANOM procedure in SAS by providing SAS code for some of the examples used in the book.