Case 6 - Priority Assessment:
ANOVA and Exploratory Data Analysis

Marlene Smith, University of Colorado Denver
Business School

Priority Assessment:
ANOVA and Exploratory Data Analysis

Background

Software development projects typically follow six basic phases: Requirements, design, implementation (and integration), testing (validation), deployment (installation) and maintenance. First, general requirements are gathered, and the scope of the functionality is defined. Then, alternative scenarios for the required functionality are developed and evaluated. Implementation, usually 50% or more of the development time, is the phase in which the design is translated into programs and integrated with other parts of the software – this is when software engineers actually develop the code. During the final phases, programs are tested, software is put into use, and faults or performance issues are addressed.

ApDudes, a developer of applications for tablet computers, was having difficulty meeting project deadlines; only 10% of their projects had been completed within budget and on time last year and that was starting to hurt business. The group's project manager was tasked with studying problems within the implementation phase. He found that software engineers were having difficulty prioritizing their work, and that they often became overwhelmed by the magnitude of the projects.

As a result, two changes were made. Each project was broken down into smaller, distinct tasks, or jobs, and each job was assigned a priority. The project manager believes that this classification and prioritization system would speed the completion of high priority jobs, and thus lower overall project completion time.

The Task

We will focus on the prioritization system. If the system is working, then high priority jobs, on average, should be completed more quickly than medium priority jobs, and medium priority jobs should be completed more quickly than low priority jobs. Use the data provided to determine whether this is, in fact, occurring.

The Data Priority Assessment.jmp

The data set contains a random sample of 642 jobs completed over the last six months. The variables in the data set are:

Days	The number of days it took to complete the job
Priority	The priority level assigned to that job

Analysis

Exhibit 1 shows low priority jobs, as expected, took the longest on average. Surprisingly, high priority jobs actually took longer on average (3.0 days) than medium priority jobs (2.5 days). Variation around the average for medium priority jobs (5.0 days) is also smaller than the other priority levels. For all priority levels, some jobs probably took quite a bit longer than most of the other jobs, since the medians are notably lower than the averages.

Also of note is the number of low, medium and high priority jobs. There were far more high priority jobs than low or medium.

Exhibit 1 Summary Statistics for **Days**

	Days			
Priority	**N**	**Mean**	**Median**	**Std Dev**
Low	67	4.2	1.2	6.3
Medium	249	2.5	0.5	5.0
High	326	3.0	0.6	7.8

(Analyze > Tabulate; drag ***Priority*** *in drop zone for rows, and* ***Days*** *in the drop zone for column as an analysis column. Then, drag Mean, Median and Std Dev from the middle panel to the middle of the table.*

Note that in JMP versions 10 and earlier Tabulate is under the Tables menu.)

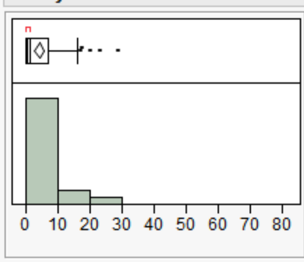
A closer look at the data (Exhibit 2) indicates that the distributions for all priority levels are highly skewed, which is not uncommon for time-based data. We also see three high priority jobs that took extremely long to complete relative to the other jobs.

Exhibit 2 Distribution of **Days** by **Priority Level**

(Analyze > Distribution. Select ***Days*** *as Y, Columns and* ***Priority*** *as By. For a horizontal layout select Stack under the top red triangle. To apply the same x-axis scaling to all graphs, click the control key and select Uniform Scaling from the top red triangle.)*

Distributions Priority=Low

Days



Quantiles

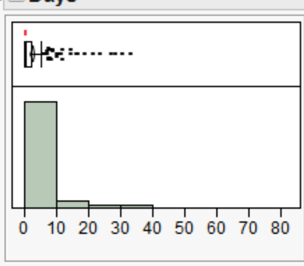
100.0%	maximum	28.5
99.5%		28.5
97.5%		24.79
90.0%		13.48
75.0%	quartile	6.8
50.0%	median	1.2
25.0%	quartile	0.3
10.0%		0.2
2.5%		0.1
0.5%		0.1
0.0%	minimum	0.1

Summary Statistics

Mean	4.2283582
Std Dev	6.2694499
Std Err Mean	0.7659352
Upper 95% Mean	5.7575972
Lower 95% Mean	2.6991192
N	67

Distributions Priority=Medium

Days



Quantiles

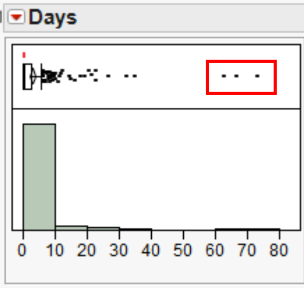
100.0%	maximum	33.2
99.5%		32.5
97.5%		20.775
90.0%		7.4
75.0%	quartile	2.25
50.0%	median	0.5
25.0%	quartile	0.2
10.0%		0.1
2.5%		0.1
0.5%		0.1
0.0%	minimum	0.1

Summary Statistics

Mean	2.5
Std Dev	5.0116638
Std Err Mean	0.3176013
Upper 95% Mean	3.1255398
Lower 95% Mean	1.8744602
N	249

Distributions Priority=High

Days



Quantiles

100.0%	maximum	72.9
99.5%		68.8995
97.5%		22.2075
90.0%		7.7
75.0%	quartile	2.5
50.0%	median	0.6
25.0%	quartile	0.2
10.0%		0.1
2.5%		0.1
0.5%		0.1
0.0%	minimum	0.1

Summary Statistics

Mean	3.0236196
Std Dev	7.7813679
Std Err Mean	0.4309701
Upper 95% Mean	3.8714629
Lower 95% Mean	2.1757764
N	326

Having extreme points in skewed distributions is not unusual. But these jobs all took over 60 days, at least 28 days longer than the next longest job in any priority level. Upon further investigation it was learned that these three jobs were not closed out when they were completed. Instead, a system date was applied at the end of the quarter. Since the actual number of days for these jobs is unknown, we will remove them from the analysis using hide and exclude.

Exhibit 3 shows the summary statistics after removing the three bad data points. Although we have a large sample (323 jobs), both the mean and the standard deviation show a sizeable decrease.

Exhibit 3 Summary Statistics after Removing Bad Data

	Days			
Priority	**N**	**Mean**	**Median**	**Std Dev**
Low	67	4.2	1.2	6.3
Medium	249	2.5	0.5	5.0
High	323	2.4	0.6	4.7

(To exclude and hide, draw a box around the points in the boxplot to select them. Then, select Rows > Hide and Exclude.)

Now that data quality issues have been resolved (no other issues were found), we proceed to focus on the question at hand. Namely, are higher priority jobs being completed more quickly than lower priority jobs?

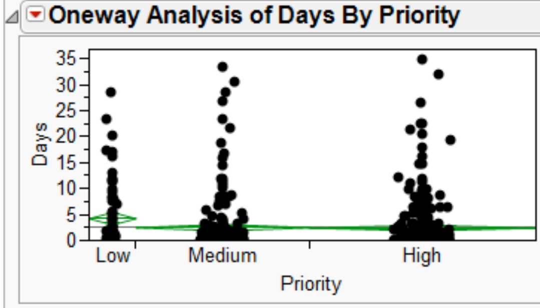
A conclusion regarding the priority system based solely on the descriptive statistics in Exhibit 3 fails to account for sampling error. For example, the average days to complete medium priority jobs (2.5 days) is numerically larger than average days to complete high priority jobs (2.4 days), but is this difference due to random variation (noise)?

This situation involves a continuous response variable (**Days**) and categorical predictor (**Priority**) which has three levels or groups. Since we are interested in comparing the average days of completion for these three groups, the appropriate statistical method is One-Way Analysis of Variance, or, ANOVA.

The p-value (Prob > F =0.0235) in Exhibit 4 indicates that there are significant differences in the average completion time for the three priority levels (at a significance level of 0.05).

Exhibit 4 Oneway ANOVA for **Days** by **Priority**

Oneway Analysis of Days By Priority



Excluded Rows 3

Oneway Anova

Summary of Fit

Analysis of Variance

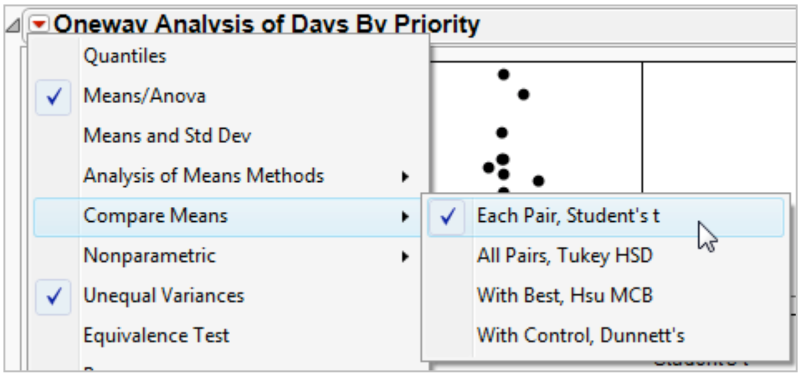
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Priority	2	188.739	94.3695	3.7747	0.0235*
Error	636	15900.393	25.0006		
C. Total	638	16089.132			

*(Analyze > Fit Y by X; use **Priority** as X, Factor and **Days** as Y, Response. Then, select Means/ANOVA under the top red triangle. To jitter the points, select Display options, Points Jittered.)*

ANOVA is an overall test for differences. It tells us that at least one of the three priority levels is different from at least one other – but it doesn't tell us which priority levels differ from one another. Multiple comparison tests allow us to compare individual pairs of means. Each test produces comparison circles and test results for comparing means.

The Each Pair, Student's t test (Exhibit 4) allows us to compare each mean to every other mean.

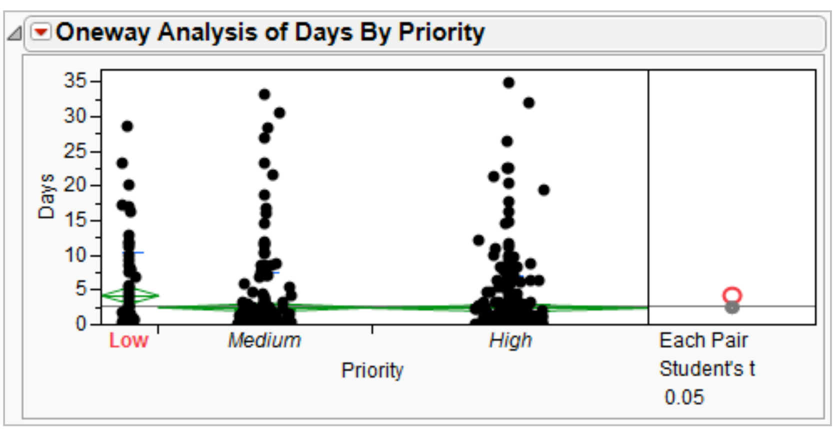
Exhibit 4 Multiple Comparison Procedures



(Four multiple comparison procedures are available. From the Oneway output window, select Compare Means under the red triangle.)

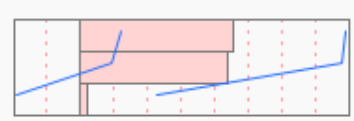
The distribution of days to complete low priority jobs is significantly different (higher) than the days to complete medium or high priority jobs. Interestingly, completion times for high and medium priority jobs aren't statistically different (Exhibit 5).

Exhibit 5 Multiple Comparison Procedure



(Click on a comparison circle to compare a mean to the other means. Means significantly different from the selected mean at significance level of 0.05 will have gray circles.)

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Low	High	1.802662	0.6712264	0.484574	3.120749	0.0074*
Low	Medium	1.728358	0.6881480	0.377041	3.079675	0.0123*
Medium	High	0.074303	0.4216698	-0.753730	0.902337	0.8602



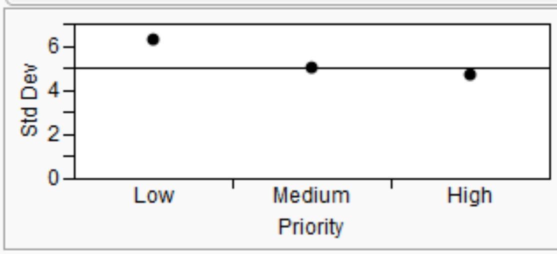
One-Way ANOVA presumes equal variances for the three groups. In Exhibit 3 we saw evidence that the standard deviations may not be equal. Exhibit 6 shows that the equal variance test results are mixed.

Three of the tests indicate that there is a significant difference in the variances (p-values are below 0.05), while the O'Brien test indicates that the differences are not significant.

Since we have some ambiguity about equal variances, we'll refer to the Welch Test. Unlike ANOVA, the Welch Test indicates the difference in average days to completion for the three priority levels is not statistically significant at a 0.05 significance level (Prob > F = 0.0841). According to Welch, differences in the means are marginally significant at best.

Exhibit 6 Unequal Variances Test for **Days** by **Priority**

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Low	67	6.269450	4.705458	3.849254
Medium	249	5.011664	2.840161	2.268273
High	323	4.688176	2.778761	2.198452

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.8606	2	636	0.4234
Brown-Forsythe	3.4093	2	636	0.0337*
Levene	6.9613	2	636	0.0010*
Bartlett	5.2144	2	.	0.0054*

Welch's Test

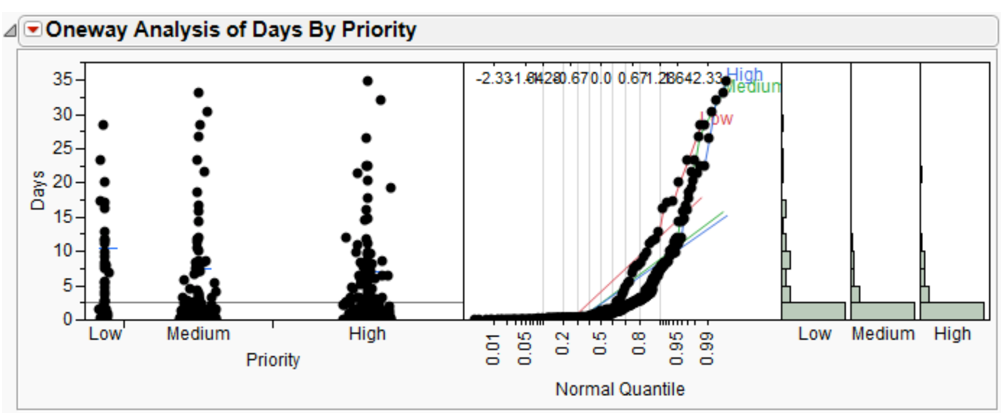
Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
2.5115	2	172.89	0.0841

(From the Oneway output window, select Unequal Variances under the top red triangle.)

One-Way ANOVA also presumes normality. The middle panel in Exhibit 7 suggests a violation of the normality assumption, since the observations do not fall on a straight line in the normal quantile plot. The curved pattern indicates that the distributions are right skewed. We can also see evidence of skewness in the scatter plot (left) and in the histograms (right).

Exhibit 7 Normal Quantile Plots and Histograms for **Days** by **Priority**



(From the previous output, select Normal Quantile Plot from the red triangle. To produce histograms, select Display Options > Histograms.)

Fortunately, ANOVA generally works well even if the underlying distributions are non-normal, so long as the distributions are roughly the same. Because Exhibit 7 shows this to be the case, no subsequent analyses are needed to address non-normality.

Summary

Statistical Insights

Qualitative outcomes of statistical tests (e.g., rejecting or failing to reject null hypotheses) can depend on the test used, even in large samples. Thus, validating that test assumptions are met should become standard practice. In this scenario, ANOVA and the multiple means comparison found that low-priority jobs took statistically longer to complete, yet that difference disappeared when using the Welch test. As sometimes happens, the final conclusion rests on the choice of significance level; the Welch test finds statistically significant differences at the 0.10 significance level but not at 0.05. Regardless, something important has been learned here: there is neither a statistical nor real-world reason to conclude that high-priority jobs are being completed faster than medium-priority jobs on average (2.5 days versus 2.4 days).

The case demonstrates how to test for the equal variances and normality assumptions in an ANOVA setting. The assumption of independence was presumed to hold for these data, since completion times were measured separately for each job. Read on for an after-the-fact assessment of this assumption.

Managerial Implications

Several aspects of the data analysis surprised the project manager. The first was the unexpectedly large number of jobs needing urgent attention—just over half of all jobs completed were identified as high-priority (Exhibit 1). Unless high priority jobs are truly urgent, and rarer than others, the label might have lost meaning over time, leading to a breakdown in the prioritization system. This prompted management to ask further questions about how priorities were assigned, by whom, and by what criteria.

Secondly, simple examination of the histograms revealed unexpected results: about 25% of all jobs took only 90 minutes (0.2 days) or fewer to complete. In retrospect, it appeared as if the data set included both difficult and meaningful development tasks and simple enhancements or minor debugging rework. This prompted the project manager to think harder about what was meant by a “job”, leading to involved discussions with the data warehouse manager about data needs moving forward.

Finally, management learned that high-priority jobs were not completed faster on average than medium-priority jobs, an outcome that prompted the project manager to ask the software engineers about their experiences with the new prioritization system. Several mentioned anecdotally that jobs in the downstream phase of a project seemed to have more bugs than before the prioritization system was put in place, a suspicion that was borne out by further data analysis. As it turned out, developers reacted in part to the prioritization system by conducting fewer diagnostic tests in order to meet a perceived deadline implied by the priority label. The prioritization system seems to have produced an observational bias, also known as the Hawthorne effect, in which behavior is changed in response to a real or perceived sense of being monitored.

JMP Features and Hints

This case uses JMP to examine differences between group means in the Fit Y by X platform: ANOVA to test for differences between means, Unequal Variances to test the equal variance assumption, and the Welch Test as a follow-up to ANOVA. JMP also produces four multiple comparison procedures, comparison circles and test results for comparing individual pairs of means.

JMP's normal quantile plot and graphical displays of data within each group can be used to assess the normality assumption. As an alternative, JMP can be used to produce a family of nonparametric tests to deal specifically with non-normal data. We will visit two of these tests in an exercise to follow.

Exercise 1

Use the Medical Malpractice case and data set to determine whether the average award amount varies by the insurance carrier of the patient.

Exercise 2

Two popular nonparametric tests are the Median Test and the Kruskal-Wallis Test. In both tests, observations are transformed to ranks, and these ranks are used instead of the raw data. JMP provides both tests in the Fit Y by X analysis window under the top red triangle, Nonparametric.

These tests are analogous to ANOVA. However, the Median Test compares the medians of the groups, rather than the means, while the Kruskal-Wallis test compares the distributions of the groups.

1. Use the Median Test to compare the median **Days** for the three priority levels. Compare this to ANOVA results found earlier in the case.

2. Use the Kruskal-Wallis Test (select Nonparametric > Wilcoxon) to compare the distribution of **Days** for the three priority levels. If the test is significant at the 0.05 level, use a nonparametric multiple comparison procedure to determine which priority levels are different (use Nonparametric > Nonparametric Multiple Comparisons > Wilcoxon Each Pair). Compare these results to the multiple comparison results found earlier in the case.