BEFORE THE PUBLIC SERVICE COMMISSION OF UTAH

Explanation of Statistical Methods and Issues

Artie Powell

DPU Exhibit 2.3 DIR-COS

Docket No. 13-035-184

Division of Public Utilities

May 22, 2014

DPU Exhibit 2.2 DIR-COS

Docket No. 13-035-184

Explanation of Statistical Methods and Issues

Several statistics are employed to compare or determine if the average monthly peaks from the Company's Stress Factor Study are statistically different.

F-test: the F-statistic or test is a commonly used statistic when comparing more than one population parameter. For example, comparing the variances for two independent populations or, as employed herein, the means of more than two populations. In the latter example, the Null hypothesis assumes that all of the population means are the same verses the alternative hypothesis that at least one of the means is different:

$$H_0: \mu_1 = \mu_2 \cdots = \mu_n$$
$$H_A: At\ Least\ One\ Mean\ is\ Different \tag{1}$$

In the case where the difference between only two population means are being considered, the F-statistic is the ratio of the two sample variances. When comparing more than two means, as in the present case, the F-statistic is the ratio of the within to the error variance from an Analysis of Variance (or ANOVA) study. (See DPU Workpapers 2.2 DIR COS – 2.5 DIR-COS)

The F-test will reveal whether at least one of the means is different from the others. However, it will not reveal which if any are different. In order to determine which means are different requires pairwise comparisons using tests such as the Student t-Test.

Student t-Test: The Student t-Test is a common statistic to test assumptions about the means of two populations, in particular, whether the means are equal.

$$H_0: \mu_1 = \mu_2 \qquad (2)$$

$$H_A: \mu_1 \neq \mu_2$$

With 12 months, there are 66 pairwise comparisons to consider.[1] However, using the data in this manner can yield misleading or erroneous conclusions: there is an increased probability of rejecting the Null Hypothesis, concluding that two means are statistically different, when in fact they are not. This is commonly known as the multiple or pair wise comparison problem.

Multiple Comparisons: When testing two means in this manner, the researcher usually chooses the test size "α" and a corresponding critical value (cv) as a decision criteria: if the t-statistic is larger than the cv, then the Null Hypothesis is rejected otherwise, we fail to reject the Null Hypothesis. We commit a Type I Error if we reject the Null Hypothesis when in fact the Null Hypothesis is true. If we fail to reject the Null Hypothesis when it is false, we commit a Type II Error. For a single test, the probability of committing a Type I Error is equal to the test size, α. However, for a series of pairwise tests, the probability of a Type I Error can be much larger than α. In particular, for a given test size α and independent tests, the probability of a Type I Error is,

$$\tilde{\alpha} = 1 - (1 - \alpha)^T \qquad (3)$$

where T is the number of tests. If T equals 66, as in the present case, the probability of a Type I Error increases from say α = 5 percent to approximately $\tilde{\alpha} = 97\%$. In other words, it is very likely that we would conclude at least one pair of months were different when in fact they are not, a Type I Error, if we simply proceeded with the Student-t Test.

[1] The 66 comparisons is the result of the combination of 12 months choosing two at a time for comparison: ${}_{12}C_2 = 66$.

As a simple illustration of this problem, consider an experiment of flipping a coin 10 times. In advance, we do not know whether a particular coin is fair, therefore, we define an unfair coin as one that yields nine or more heads. In other words, the Null Hypothesis is that the coin is fair. We will reject that hypothesis if we actually flip the coin 10 times and record 9 or 10 heads. The probability of rejecting the Null Hypothesis (committing a Type I Error) in this case, given the coin is fair, is approximately 0.011 or 1.1 percent:

$$\begin{aligned} P(9\ or\ 10) &= P(9) + P(10) \\ &= \binom{10}{9}(0.5)^9(0.5) + \binom{10}{10}(0.5)^{10}(0.5)^0 \\ &= 0.00980 + 0.00098 \\ &= 0.011 \end{aligned} \qquad (4)$$

Now suppose, we repeat this experiment 100 times. What is the probability that **at least one** trial—flipping the coin ten times—out of the 100 trials yields nine or more heads even if the coin is in fact fair? Let p = 0.011 be the probability of nine or more heads on any one trial. Then the probability of rejecting the Null Hypothesis at least once in the 100 trials is approximately 66%:

$$\begin{aligned} P(At\ Least\ One\ Rejection) &= 1 - (1 - p)^T \\ &= 1 - (1 - 0.011)^{100} \\ &= 0.66 \end{aligned} \qquad (5)$$

Tukey's HSD Test: To maintain a constant probability of a Type I Error requires an adjustment in the initial test size. For example, solving Equation 3 for α yields,

$$\alpha = 1 - (1 - \tilde{\alpha})^{1/T} \qquad (6)$$

For example, if we want the overall probability of a Type I Error to be 5 percent, $\tilde{\alpha}$ = 0.05, then we could choose a test size of α = 0.0008 for each pairwise comparison. With eight degrees of freedom, $8 = n_1 + n_2 - 2$, the critical value would be approximately 5.23 compared to 2.31 for a single comparison.

A somewhat simple but common procedure that accomplishes this adjustment is Tukey's Honestly Significant Difference (HSD) procedure.[2] The HSD procedure combines an adjusted critical value with the information from a standard ANOVA:

$$w(HSD) = q(T, df|\alpha) * \sqrt{\frac{MSE}{n}} \tag{7}$$

where

$q(T, df|\alpha)$ = the adjusted critical value taken from a table of such values;

MSE = mean squared error from the ANOVA; and

n = the sample size.

The Null Hypothesis that two means are the same is rejected if the difference in the means is greater than w(HSD).

---

[2] William Mendenhall, James E, Reinmuth, and Robert J. Beaver, "Statistics for Management and Economics," (1993), Duxbury Press, Belmont, California, pp. 475-478.