APPENDIX C – STATISTICAL CONFIDENCE LIMITS

STATISTICAL CONFIDENCE LIMITS

A survey such as that used in this project solicits answers from a sample of respondents. If the sample is a random sample (or something reasonably close to being random) one can compute "confidence limits" for the estimates from the responses to the survey questions.

General Idea – In a political poll, the results might be stated in a form such as the following: 46 percent of the population is in favor of candidate Jones, plus or minus five percent. That is, the poll predicts that candidate Jones will get 41-51 percent of the vote. The "plus or minus" statement is the confidence limit. If a larger sample were used, the "plus or minus" limits would get smaller; e.g., ±3% instead of ±5%. If the sample was the whole population, then the result would be known exactly.

The confidence limits are useful to compute so that one knows how confidently the responses from the 716 respondents to the question can be used as an estimate of the response you would get if the whole firefighter population had responded. In most usage, the "confidence limit" is actually a "95 % confidence limit". There is a 95 % probability that the result you would get by sampling the entire population fall within the ± error stated as the confidence limit.

The calculation of confidence limits can be used to state how different the responses to two questions have to be (how many percentage points apart) for us to say that one would be ranked higher than the other if the whole population responded. Likewise, confidence limits can be calculated to show how far apart the responses from two subgroups (such as females and males, or firefighters in the Northwest area vs. firefighters nationwide) have to be for us to be able to say they are significantly different.

Sampling Error – As with any survey, a sampling error is associated with each percentage given in the tables in this report that were based on the random survey of wildland firefighters. In statistical parlance, this error is often referred to as the 95% confidence limit. For example, in Table 4-3, 37% of the approximately 700 respondents felt there was an inadequate exchange of information between shifts on a fire more than half the time. There is a sampling error that comes from having 700 responses to the question instead of response from the whole population of firefighters.

Without getting into a lot of statistical theory, a 95 % confidence limit means you are 95 % sure that the true value of the estimate is within the calculated statistical error. If the error is said to be 4%, then you would be 95% sure (or confident) that if you had data on the whole population of firefighters, the "true" answer to the question would be 37% ± 4% in the example, that is, the true answer would lie in the range 33 - 41%.

The approximate formula for the sampling error is extremely simple, namely $1/\sqrt{n}$, where n is the sample size (number of respondents to the survey). In some statistical tables, n refers to the entire sample (approximately 700 here) while in other tables n refers to a subgroup (such as the number of firefighters in a particular geographic area or agency who responded). The error for n = 700 is $1/\sqrt{700} = 3.78$ %.

Although the value of $1/\sqrt{n}$ is the quantity usually reported as error by the media in public opinion polls, actually it is an approximation of the upper bound. The more precise formula is, with p the observed percentage (e.g. 37% in the above example)[1]:

[1] The formula is based upon the normal distribution approximation to the binomial distribution. This approximation is excellent except when n and p are small and the binomial distribution becomes substantially asymmetric. Also, in the formula some statisticians prefer using n-1 in place of n to allow for the observed p being merely an estimate of the "true" p, but the difference is small for the ns we are dealing with. And of course, to obtain the final error in percent one multiplies by 100.

$$\text{Error (in percent)} = 1.96\ \sqrt{p(1-p)/n}\ \text{x}\ 100\%$$

$$= 1.96\ \sqrt{p(1-p)}\ \text{x}\ 1/\sqrt{n}\ \text{x}\ 100\%$$

The quantity $\sqrt{p(1-p)}$ has a maximum value of 1/2 when p equals 1/2 (i.e. 50%) and the error becomes 1.96 x 1/2x1/$\sqrt{n}$ $= .98\sqrt{1/n} \cong 1/\sqrt{n}$, the same as the approximate maximum error noted above. In general $\sqrt{p(1-p)}$ is symmetric about 1/2 (e.g. it is the same for p=3/4 as it is for 1/4). Near 1/2 the error is only slightly less than at 1/2, while near the extremes, p=0 or 1, there is only a small error (reflecting the smaller statistical variation). The error is reduced by half when p=6.7%. That is, if 6.7% of the survey population felt that a situation was true, you would have half the error than if the 50% of the population thought it were true.

The table below shows the actual error, from the above equation, associated with different values of p and n. For the previous example, in Table 4-3, with p=37 % and the total number of responses n=700, the table shows that the error would be a little less than 3.63% (the error for 40%), so the error of ±3.78% we calculated above was a very good approximation. Note that if there had been 1000 responses (the original survey goal), the error would have been reduced only to about ±2.99%. In general, if one is satisfied with a rough estimate of the error, the $1/\sqrt{n}$ approximation will suffice, at least for p greater than 30%.

The magnitude of the error determines whether there is a statistically significant difference between two responses. For example, the difference in the percentages reported for the two highest reasons in Table 4-3, 37% and 35%, is obviously not significant. It is not as clear, however, whether the difference of 8% between the second and third most important reasons given is significant, since the error for just the 37% estimate alone was ±3.8%, greater than the difference between the two percentages, not even considering the error for the second 35% estimate. Some persons might consider it to be significant if the difference exceeds the error (about 4%), while others might think

that a total difference of 8% (4% for the first and 4% for the second percentage) is needed for significance. Actually the former is too little (since it would apply only if one of the percentages had zero error) while the latter, doubling the error for one estimate, is too stringent (since it presumes worst case sampling errors for each response). The true difference required for significance is intermediate: approximately $\sqrt{2} \cong 1.4$ times the error. That is, the difference - using the $1/\sqrt{n}$ approximation for error - should be, at least approximately, $\sqrt{2/n}$, or $\sqrt{2/700} = 5.35\%$. When the sample sizes differ - for example, in comparing responses by different agencies to the same question - the previous formula for the significant difference, $\sqrt{2/n}$, is replaced by $\sqrt{1/n_1 + 1/n_2}$, where n_1 is the size of the sample from the first agency and n_2 is the size of the sample from the second agency.[2]

A special situation relates to considering the selection of two answers in the same question. Unfortunately, the statistical theory required to determine the appropriate error has not been analyzed. Clearly, however, there will be twice the number of responses than if merely the single most important factor had been requested. It is significant that the same procedure for determining error (and statistical difference) be employed in this case, with of course a doubling of the result.[2]

---

[2] A better approximation for the required difference to obtain significance at the 95% level is obtained by using the average error or the average p, for the two responses from the table. For 35% the error (with n = 700) is +3.53% and p=27% the error is +3.29, giving an average error of 3.41%. Multiplying by $\sqrt{2}$, gives a required difference of 4.82%. A similar computation using an average p of 31% yields a required difference of 4.84%. The exact formula for the required difference for significance (based upon the normal distribution and using possible different sample sizes n_1 and n_2) is $1.96\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$; this formula gives the value 4.83%, not much better than the approximation.

[3] The doubling of percentages can be justified also by the following analysis. If the selected pairs, from among k choices say, were completely at random, there would be a total of k(k-1)/2 different pairs possible. Since any given response could be paired with k-1 different second choices, the ration 2/k, is the average response under random selection of pairs; this is double the response, 1/k, for a single factor selection. It is noted also that the negative correlation in responses - arising from the fact that each selection implies rejection of other choices - is small when k exceeds 2, and hence can be ignored when determining significant differences.

Remarks On Validity – Any survey such as this contains several inherent sources of error. The most important relate to bias due to non-responses, and the fact that many questions require fine distinctions by the respondent that are to a certain extent arbitrary. Unfortunately, the impact of these on the validity of the overall results cannot be established quantitatively. (The quantitative effect of errors due to sampling is treated in a later section of this report.) In this survey the over 50% response (716 out of 1400 surveyed) is exceptionally high, reflecting the high degree of cooperation by the respondents, even with over 300 `questions covering 25 pages. Also very useful were the 300 one-on-one interviews and focus group sessions. Overall the results are undoubtedly valid.

If the number of firefighter responses to the questionnaire had been increased, say to 1000, the effect would have improved the overall results only slightly. The confidence limits on a typical question would only have improved (decreased) to ±3% instead of ±4% if there were 1000 responses instead of 700. There would be a slightly greater impact on confidence limits for subgrouping comparisons. Any sample size, short of 100%, will invariably contain fewer responses than one would like for certain groupings of interest. A tradeoff between cost and accuracy is always required.

The allocation of responses to the different regions, agencies and ranks deviated somewhat from the targeted allocation that reflected the estimates number of personnel in each group. Although one could in principle adjust the response rates by scaling to the targeted allocation, this procedure does not seem appropriate considering that the deviations were quite moderate and the actual numbers of firefighters were not always accurately known.

Summary

The confidence limit for the response to a single question involving the total population of 700 firefighters is in the range of ±3 to ±4%.

The confidence limit for comparing two questions is ± 5 to 6%.

For examining important subgroups of the population, the sample sizes ranged from 32 (agency administrators) to 98 (female firefighters). For comparing responses from the largest subgroup considered (females) to the rest of the population, the error is about ±10.8%. So, for example, if 37% of women agreed to a question, and 26% of men, that would be a significant difference. For the agency administrator sample compared to the rest of the group, the error is 17%; in other words, unless the agency administrator were 17 percentage points different from the full group, it is not a significant difference.

For comparing two small subgroups, (e.g., geographic areas), where the number of respondents averaged 70, the error in comparing two areas is 16.9%. Differences have to be more than 17 percentage points for them to be statistically significant; i.e., one geographic area reports 40% in agreement and another 60% in agreement.

A final note: when the differences between two subgroups or between two questions from the main groups are smaller than the errors quoted above, they are not statistically significant at the 95% level, but they may be statistically significant at the 90% or 80% or some other level. For example, if the response to a question from female firefighters is 8% higher than the response from male firefighters, that means one is not 95% certain that the full population of female firefighters would score higher than the men, but it still is considerably more likely than not.

Table C-1. Sampling Error

SAMPLING ERROR (CONFIDENCE LIMITS							
		Percentage of Survey Respondents					
n/p	Approx. Error	50 %	40%	30%	20%	10%	5%
1000	3.16	3.10	3.04	2.84	2.48	1.86	1.35
700	3.78	3.70	3.63	3.39	2.96	2.22	1.61
625	4.00	3.92	3.84	3.59	3.14	2.35	1.71
400	5.00	4.90	4.80	4.49	3.92	2.94	2.14
200	7.07	6.93	6.79	6.35	6.54	4.16	3.02
100	10.00	9.80	9.60	8.98	7.84	5.88	4.27
50	14.14	13.86	13.58	12.70	11.09	8.32	6.04
25	20.00	19.60	19.20	17.96	15.68	11.76	8.54
16	25.00	24.50	24.00	22.45	19.60	14.70	10.68