Introduction

A construction defect suit typically starts with a number of homeowners in a condominium or single-family home development complaining about imperfections in their dwelling that compromise the value of their dwelling, and in some cases, even their families' health, if toxic mold results from water intrusion. Either plaintiff's attorneys or home owners' associations decide that a critical mass of complaints warrants a class-action law suit, and so a suit is filed in behalf of all (participating) homeowners. To estimate damages – typically, the cost of repair – attorneys retain construction experts, architects or biologists to (1) compute the cost of remediation for defective dwellings, and (2) estimate the extent of the defect among the subject properties. While all three types of experts are competent to perform the first task, the second task – extrapolating damages from tested properties to all properties – is within the parlance of the statistical expert. Without an appropriate application of the principles of random sampling and statistical inference, damage estimates are likely to be much too high. While unscrupulous attorneys and experts may be unfazed by excess compensation, such excess may lead to an inappropriate legal backlash against "frivolous law suits" that may prevent legitimate construction defect claims from just compensation in the future.

The Meaning of Statistical Inference

Statisticians and other professionals, such as economists like me, are plagued by the necessity of inferring the unknown characteristics or *parameters* of a population of interest, from the characteristics or *statistics* from a known sample. For instance, a certain proportion, π, condominium units have a construction defect, and the remaining proportion, $1-\pi$ are free of defect. According to the rules of *statistical inference*, the sample proportion *p*, can be predicted to have the average value π and a standard error equal to $\sigma_p = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$, where *n* is the sample size and *N* is the population size, if and only if the sample is a random one. It follows that, from a random sample, the likely values of the sample proportion follow the normal or Gaussian distribution,

$$\Pr(\pi - 1.96\sigma_p < p < \pi + 1.96\sigma_p) = 95\%$$

Since typically π is unknown, we substitute *p* and use the *t*-distribution in place of the Gaussian distribution, revising the confidence interval to read:

$$\Pr(\pi - 2.0639 s_p < p < \pi + 2.0639 s_p) = 95\%$$

where $s_p = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$. For instance, if a random sample of 25 units from a 280 unit complex showed a 80% defect rate (i.e., 20 units with defects, 5 without defects), we could be 95% confident that the defect rate for the population lies between 64% and 92%. [1]

[1] The sampling error is $s_p = \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{.8(.2)}{25}} \times \sqrt{\frac{280-25}{280-1}} =$ $(.08)(.956) = .0765$. Hence, $\Pr(.8 - 2.0639(.0765) < \pi < .8 + 2.0639(.0765) = 95\%$, or we are 95% confident that .6422 < π < .9578, or 64.22% < π < 95.78%.

Note, that if we wanted to be absolutely certain about the number of defects, we would have only two options: (1) test all 280 units and determine the defect rate (assuming perfectly accurate testing), or (2) incorporate all possibilities into the confidence interval. Since it is possible (though unlikely, unless the sample was biased) that only 20 units are defective, the minimum defect rate is $\pi_{\min} = \frac{20}{280} = 7.14\%..$ At the other extreme, it is possible that only five units in the entire development are defect free, implying the maximum defect rate of $\pi_{\max} = \frac{260}{280} = 92.86\%$. By discarding unlikely events (those with a cumulative probability of only 5%), we obtain a substantially narrower confidence interval, *but only if the sample is a random one.*

The Impact of Sample Bias

However, if the 25 units do not represent a random sample, there is no basis for the inference of the defect rate among untested units. This is because of intentional or unintentional sample *bias*. The most obvious form of sample bias is to include the units owned by the homeowners who brought the suit in the sample. Suppose that 10 of the 20 defective units were identified as such in response to complaints by 10 homeowners. Those 10 units are *not* a random sample, because they were selected specifically because the defect rate was high. Assuming honest testing for defects, those homeowners have legitimate claims and should be included in the pool of properties seeking remediation. However, since the probability of defects for this sub-population is known to be 100%, these units should not be in the pool of units tested to estimate the extent of the damages.[2] Removing these 10 units from the sample would reduce the measured defect rate from 80% to 67%. And, because the sample size is only 15 (out of a remaining population of 270 units), confidence interval becomes:

$$s_p = \sqrt{\frac{.67(.33)}{15} \times \frac{270-15}{270-1}} = \sqrt{.0140} = .1185$$
$$t(.95,14) = \pm 2.145$$
$$\Pr(.67 \pm 2.145(.1185)) = 95\% \rightarrow$$
$$41.25\% < \pi < 92.09\%$$

Whenever the sample of defective units includes the units known to be defective before the suit was filed, the defect proportion will be over-estimated.

[2] In the extreme, suppose the units owned by the 10 complainers were the only defective units in the entire development. Any random sample of 15 units would show a defect rate of 0%, implying that only the original 10 owners should sue. Including the 10 units in any sample of 25 would imply a defect rate of 40%, when the *true* defect rate would only be $\pi = \frac{10}{280} = 3.57\%$!

Another difficulty with obtaining a random sample results from the fact that the fact that plaintiff's experts must secure the permission of homeowners to use "destructive tests" to diagnose the existence of suspected defects. Even if the experts work from a randomized list of units to secure permission, it is much more likely that homeowners will agree to the test if they believe their unit is defective than if they believe that the unit is defect free. If a homeowner suspects that her unit has a defect, there is a much higher probability that it has a defect than if the homeowner does not expect the defect. Suppose that when a homeowner believes their unit has a defect, the probability of that defect is 67%: $\Pr(D \mid B) = 67\%.$, where Pr(*D*|B) is read *"The conditional probability of a defect, given belief*. Alternatively, assume that when the homeowner does not complain, the probability of a defect is only 5%: $\Pr(D \mid NB) = 5\%$, where *Pr*(*D*|*NC*) is read "*The probability of a defect given no belief*. By testing only units where the owner expects a defect (*B*), but treating this as a random sample, one would incorrectly infer that the defect rate was between 20.67% and 59.33%, when the true defect rate is the weighted average of the two sub-populations. Suppose that 50 homeowners believe their units are defective and the remaining 220 homeowners do not.[3] Then the true (but unknown) population proportion would be: $\pi = \frac{50}{270}(.67) + \frac{220}{270}(.05) = .1235 + .1642 = .2877$. If all 270 units had an equal chance of being included in a sample of size 15 (the definition of a true random sample[4]), we would expect to find 95% of random samples to show defect rates between 5% and 52%, while the use of only units whose owners believe they have a problem voluntary would produce sample proportions between 41% and 93%.

A Modest Proposed Solution to Homeowner Selection Bias

If plaintiffs' attorneys are truly interested in an accurate projection of defective properties, there is a relatively straight-forward fix for the "owner selection bias" problem discussed above. After removing the properties known to be defective from the population of all units, a randomized list of all properties should be drawn up. Either an agent for the attorney, or, better, the testing experts, should call owners in random order, and first inquire whether the homeowner believes their property has a defect. The response could be recorded as a simple yes or no, or as a probability.[5] Next, ask each homeowner if they will agree to have their unit tested; since it is critical that experts test some units owners believe to be defect-free, experts must take care to make the tests as convenient as possible for homeowners. Leaving a mess, or holes drilled in walls, merely guarantees that defective units are over sampled. Once a target number of units have been tested,[6] the unbiased proportion of defect units equals the weighted average of the defect rate for homeowners who believe they have a defect, and the defect rate for homeowners who do not believe they have a defect. Then, experts should always include the confidence interval for the number of affected units, since this will communicate to the jury the lack of precision in the test. Also, a confidence interval allows the jury to accept the lower bound of the interval if they believe the defense experts, or the higher bound, if they believe plaintiffs' experts.

---

[3] Remember, we have taken the ten units with known defects out of the population.

[4] Technically, a true random sample means that

[5] Many gamblers may be prepared to give the odds of a defect. To obtain the probability, simply divide the odds by 1 plus the odds; odds of 10 to 1 translate into p_b = 10/11 = 91%, while odds of 1 in 10 translate into p_b = 9.1%. Of course, the probability of non-belief is just 1 minus the probability of belief.

[6] There is no magic number for sample size. If cost were no object, a census of all units would be ideal, since statistical inference would then be unnecessary. However, since cost is an object, the smaller the sample, the wider the confidence interval.