# GUIDANCE ON STATISTICAL INFERENCE *Christopher Adolph*

## Credentials as a statistical expert

I am a tenured associate professor of political science and an adjunct associate professor of statistics at the University of Washington, Seattle, and hold a Ph.D. in political science from Harvard University. I have taught at the University of Washington as either an assistant professor or associate professor since 2004. Additionally, I am a core faculty member of the University of Washington's Center for Statistics and the Social Sciences, where I am also the director of statistical consulting. My primary areas of teaching and research pertain to statistical research methodology and political economy. I have published in a variety of peer-reviewed journals in political science and public policy. I generally teach three or four graduate statistical methodology courses each year at the University of Washington as well as at the Essex Summer School in Social Science Data and Analysis at the University of Colchester, United Kingdom. My opinions below reflect my personal expertise as a statistical methodologist and are not intended to represent the opinion of the University of Washington or any of its departments or centers.

From time to time, I have served as an expert witness on statistical questions in Washington state and elsewhere. In 2005, I testified as an expert witness in *Borders v. King County*, which contested the gubernatorial election of Christine Gregoire. In 2010, I was an expert witness in *Sarantou v. Lucas County*, which contested a county commissioner election in Ohio. In 2014–2015, I served as an expert witness in a consolidated rate setting case before the Washington Utilities and Transportation Commission (*WUTC v. Puget Sound Energy*).

## Summary of findings

Surveys that sample only a small percentage of the population are usually sufficient for most statistical purposes. For example, experts conducting national voting surveys routinely contact approximately 1,000 individuals out a potential pool of more than 200 million registered voters, or about 0.0005 percent of the survey population. In most survey applications, appropriately constructed samples of 1,000 to 2,000 individuals are sufficiently large to produce reliable estimates of population statistics. Conservative methods explained in detail below allow researchers to establish the confidence of those estimates based just on the size of the sample, without any reference to the size of the population or the fraction of the total population surveyed.

Accordingly, the State should expect the apple harvest wage data it collected from agricultural employers who employed between 1,610–3,632 apple harvesters (depending on the variety) in 2017 to provide useful statistical estimates of prevailing wage conditions in the relevant populations of harvesters. One option is simply to accept these estimates as the best available evidence of the prevailing wage: that would be preferable to rejecting them altogether. If the state is unwilling to simply certify these samples, then the State should follow standard scientific procedure and provide conventional scientific summaries of the confidence intervals around those sample means as described below. The State should then accept those confidence intervals as a valid summary of the likely range containing the prevailing wage. The alternative the State has chosen – to not certify or report large samples of data – cannot be justified on scientific grounds and relies on flawed reasoning about the size of samples needed for confidence in statistical inference.

## Appropriate methods for inferring prevailing wages from a sample

The Employment Security Department (ESD) and Department of Labor (DOL) seek to estimate the prevailing wage for workers in specific sectors of apple production in Washington state. In particular, there is interest in estimating from survey data the prevailing wage for harvesters of Gala, Honey Crisp, Red Delicious, and Golden Delicious apples, respectively. These surveys report prevailing payments to workers for each bin of apples collected. There is additionally interest in determining whether, when converted to an implicit hourly rate, the current prevailing wages are significantly higher than specific alternative hourly wage rates. In this letter, I will take "prevailing wage" to refer to the average wage earned by a worker in the population of all workers in a relevant sector (e.g., all workers harvesting Gala apples in Washington state). Thus, as an example, we may want to determine whether the current average wages of Gala harvesters are statistically significantly higher than $14 per hour, or any other specific amount.

In statistics, the concept of "statistical significance" relates to whether we can be confident that a statistical summary of a random sample drawn from a population of interest accurately reflects what we would have found if we had surveyed the entire population.[1] For example, if we want to know the mean hourly wage across every worker harvesting a specific apple crop in Washington state this year, we could gather a random sample of workers, compute their average hourly wage, and use simple formulas to determine whether we can be confident the average wage in the full population of workers harvesting that crop is above a specific level.

In 2017, ESD conducted a wages and working conditions survey to, in part, determine the prevailing wages for apple harvesters in Washington state for several different varieties of apples. ESD sampled employers who employ apple harvest workers. In a June 1, 2018 memo, ESD updated its findings in regard to four varieties of apples listed in this table:

| Crop | (1) Total workers employed by survey respondents | (2) Estimated total harvest workers | Column (1) as a percentage of column (2) |
|---|---|---|---|
| Gala | 3632 | 27129 | 13.39% |
| Golden Delicious | 1610 | 14394 | 11.18% |
| Honey Crisp | 2991 | 24593 | 12.16% |
| Red Delicious | 2831 | 21336 | 13.26% |

In the June 1 memo, ESD retracted their estimates for wages in these areas on the grounds of insufficient data. The basis of this decision appears to be a handbook (ET Handbook Number 385: Wage Finding Process) published in 1981. The handbook seems to imply that a certain minimum proportion of the population must be sampled in order to make statistically valid conclusions about this population. This is a classic error: in statistical methodology, the confidence we place in estimates from a sample does not in any way depend on the size of the population, but instead on the size of the sample taken from that population, and whether that sample was drawn randomly from the population, or can otherwise be assumed to be representative of the population.

The methods for estimating the sample mean of a continuous variable from a survey are so straightforward and commonly used that they are routinely taught in the earliest statistics classes offered to undergraduates. They are simple enough to be recounted

1 Statistical significance is therefore a property of an estimate from a sample and not a property of the sample itself. There is no such thing as a "statistically significant sample." Instead, it is a specific estimate from a sample that may or may not be significantly different from some comparison group or amount.

here. Let us call the wage of the $i$th surveyed worker $w_i$. To determine whether we can say with confidence that the average wage in the entire population of workers is above a specific value, we need to compute the following quantities: (1) the total number of workers in the sample (which we call $n$); (2) the sample mean wage ($\bar{w}$), which is simply the average wage in the sample, or:

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i;$$

and (3) the standard deviation of the sampled wages ($\sigma$), which is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(w_i - \bar{w})^2}{n-1}}.$$

Notice that we have not referred to the total number of workers in the population in any way. The only count of workers we need is the number we have surveyed.

Using the above quantities, we can compute the lower bound of the 95% confidence interval around the sample mean: from a statistical point of view, we can be reasonably confident the prevailing wage in whole population is at least this amount. For samples as large as those collected by ESD, we can be 95% confident that the prevailing wage in the population is at least as large as the following lower bound:

$$\text{lower bound on population wage} = \bar{w} - 1.96 \times \frac{\sigma}{\sqrt{n}}.$$

There is a complementary upper bound to the estimate calculated in a similar fashion:

$$\text{upper bound on population wage} = \bar{w} + 1.96 \times \frac{\sigma}{\sqrt{n}}.$$

Together, these bounds constitute the 95% confidence interval around the estimate of the prevailing wage. Again, these quantities – and therefore our confidence in this estimate – does not depend on the size of the population or the percentage of the population surveyed in any way.[2]

2 If we make additional assumptions, specifically, that we can identify the size of the population and that this population is essentially fixed, then we can get even more precise results if we adjust our lower bound for the proportion of the population surveyed. However, this ad-

If the lower bound of the estimate prevailing wage from the sample is above a given alternative wage, such as $14 per hour, then we could conclude with 95% confidence that the current mean wage in the overall population of workers is higher than $14 per hour. This is what is mean by a "statistically significant difference."

These formulas reveal that larger samples do produce more precise results. In layman's terms, a bigger sample produces an estimate of the prevailing wage that is more likely to be spot on. But there is a catch: each additional sampled worker makes less and less improvement to the estimate. Because adding more surveys to an already large sample doesn't offer much "bang for the buck," scientists tend not to administer more than 1,000 or 2,000 surveys in most applications.

None of the formulas above depend in any way on the size of the overall population (to be technical, these formulas already assume the population is very large). For this reason, even in surveys of quite large populations, such as the population of voters in American presidential elections, it is rare to see samples larger than 1,000 to 2,000 randomly sampled voters. This principle is a matter of simple mathematics and does not vary across different areas of study unless extraordinary precision is required. Regardless of the circumstance, there is no reason to think that large surveys of workers will depend on the percentage of workers surveyed; instead, what matters is the size of the sample and whether the sample was generated following appropriate procedures to ensure it is representative of the population.

While it is common to provide guidelines for appropriate sample size before a study is conducted, the goal of such guidelines is merely to mitigate *ex ante* the chance of either having too small a sample to draw conclusions or so large a sample that resources were wasted on unnecessary data gathering. Once a survey has been conducted, the confidence interval around any quantity of interest can be computed and the statistical significance of any result can be found using the formulas above without further reference to recommended sample sizes. Our confidence in these results need not depend on the size of the population or the fraction of the population surveyed, but on whether the sample was drawn in a random or representative fashion from that population.

Likewise, one should be cautious in drawing broad inferences about how to conduct significance tests from very specific examples. The procedure outlined above can be

justment will never make our results less certain than the bound presented here: it will give a bonus, albeit typically a small one, when we manage to survey an appreciable fraction of the worker population. For the four agricultural workforces discussed in this note, adjusting for this so-called "finite population correction" would bring each bound 10–15 percent closer to the estimated prevailing wage.

found in almost any first-year statistics text. It is easy to apply and broadly applicable to any case where populations and samples are reasonably large. In contrast, when a statistician is asked for guidance on how much data is "enough" in a specific case – for example, a case where a population is so small that even a complete census of the labor force would only produce a few hundred surveyed workers – he or she is likely to give advice that does not usefully apply to cases where large populations allow for large samples. A sample with more than a thousand surveyed workers is a large sample and should not typically provoke concerns about sample size.

If ESD were to provide the underlying survey data, it would be straightforward to compute whether current wages are significantly higher than any specific proposed alternative hourly wage using the formula above and an appropriate conversion factor between piece rates and hourly rates. For this reason, rather than simply asserting that a survey of hundreds or even thousands of workers is "insufficient" to determine with statistical precision the prevailing wage, ESD should provide the underlying survey data to allow interested parties the opportunity to assess whether or not sound statistical evidence exists regarding the prevailing wage in these agricultural categories. The procedures to do so are easy to implement and are widely used and accepted across the social and natural sciences.

CHRISTOPHER ADOLPH
20 JULY 2018