Microbial Risk Analysis xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Microbial Risk Analysis

journal homepage: www.elsevier.com/locate/mran

Characterizing the concentration of pathogen occurrence across meat and poultry industries

Q1 Natalia Lukicheva, Eric D. Ebel, Michael S. Williams*, Wayne D. Schlosser

Q2 *USDA-Food Safety Inspective Services, Office of Public Health Science, Risk Assessment and Analytics Staff, 2150 Centre Avenue, Building D, Fort Collins, CO 80526, United States*

ARTICLE INFO

Article history:
Received 6 April 2016
Revised 3 May 2016
Accepted 9 May 2016
Available online xxx

Key Words:
Lorenz curve
Gini coefficient
Food safety

ABSTRACT

Raw meat and poultry contamination frequency demonstrably varies across establishments within an industry. The degree to which more contaminated establishments contribute to the overall production of contaminated products by the entire industry is rarely explored in quantitative microbial risk assessments. We use Lorenz curves and Gini coefficients to describe the degree of concentration of contamination within the industry for several product-pathogen combinations. The analysis is based on beta distributions fit to *Salmonella* sampling data for comminuted chicken, comminuted turkey, chicken parts, and comminuted beef and beef carcasses, as well as for *Campylobacter* in chicken and turkey. We also explore empirically-derived Lorenz curves. Lorenz curves for nine product-pathogen pairs suggest the pattern of contamination ranges from highly dispersed across the industry (e.g., *Salmonella*-comminuted chicken, Gini=0.19) to highly concentrated within a small part of the industry (e.g., *Salmonella*-beef carcass pre-chill, Gini=0.77). Generally, an inverse relationship between an industry's average contamination frequency and its Gini coefficient is observed across these examples. Also, illustrative empirical Lorenz curves are biased relative to the fitted curves because of a substantial number of empirical results with point estimates of zero. Large Gini coefficients may suggest that risk management should focus on just that part of the industry with high contamination frequencies, while small Gini coefficients might suggest that risk management should be more dispersed and holistic across the industry. We discuss monitoring Gini coefficients across time, as well as other applications of these methods, as useful adjuncts to food safety risk assessments.

1. Introduction

Within the United States, the Food Safety and Inspection Service (FSIS) of the United States Department of Agriculture collects pathogen contamination data annually from nearly all establishments producing raw meat and poultry. Typically, such data provide the number of samples positive for a particular pathogen for a specific number of end-product samples collected from each establishment in a given year. Each establishment's set of results suggests the underlying frequency of contamination applicable to that establishment (i.e., its within-establishment proportion contaminated).

Statistical fitting algorithms can be used to estimate how contamination frequency (i.e., within-establishment proportion contaminated) varies across an industry for a particular product-pathogen pair (Williams et al., 2013). Although the resulting distribution may be symmetrical for some pathogens, more commonly the distribution is right-skewed (with a longer right-hand tail) within the 0 to 1 boundaries of proportions. Right-skewed distributions imply a low frequency of establishments that exhibit a high proportion of contaminated product. The degree to which these more contaminated establishments contribute to the overall production of contaminated products by the entire industry is rarely explored in quantitative microbial risk assessments. Nevertheless, risk management strategies might be influenced by a better understanding of just how concentrated or dispersed total pathogen contamination is across an industry.

The Lorenz curve typically is used to analyze inequality in the distribution of wealth across members of a population (Beach and Davidson, 1983; Beach and Richmond, 1985; Bishop et al., 1989; Bishop et al., 1991; Gastwirth and Gail, 1985). The Lorenz curve relates the cumulative proportion of wealth units to the cumulative proportion of the population (Kakwani and Podder, 1976).

The Gini coefficient is a summary metric used to communicate the degree of inequality in the distribution of wealth for a population (Morgan, 1962). The Gini coefficient is scale independent and describes the deviation from a perfectly dispersed distribution of

* Corresponding author. Tel.: 970 492 7189.
E-mail address: mike.williams@fsis.usda.gov (M.S. Williams).

http://dx.doi.org/10.1016/j.mran.2016.05.002
2352-3522/

income (where the Gini equals zero) to a perfectly concentrated distribution of income where the Gini coefficient approaches one. A perfectly dispersed distribution would result if each individual had the exact same income and a perfectly concentrated distribution would result if just one individual controlled all of the income for a population.

Although commonly reported for economic data, researchers also use Lorenz curves and Gini coefficients to measure the concentration of phenomena in other fields. For example, the Gini coefficient, as a social indicator, has been used to examine educational inequality (Holsinger and Jacob, 2009), medical care access inequality (Shortell, 1976; Wagstaff et al., 1989) and labor force inequality (Lindbeck and Snower, 1996). Previously, a food safety application used the Gini coefficient to measure the degree of spatial concentration of human *Campylobacter* cases across a province of Canada (Green et al., 2006). Researchers have also monitored the dynamics of Gini coefficients across time to examine how, for example, income distribution changes as the aggregate wealth of a population changes (Cingano, 2014).

The purpose of this paper is to examine Lorenz curves and Gini coefficients in the context of microbial food safety and risk assessment. In this application, Lorenz curves are used to describe the degree of concentration of contamination for a product-pathogen combination within an industry and can be derived from estimated contamination frequency distributions. Gini coefficients summarize the degree of concentration and are calculable from the Lorenz curves. We compare Lorenz curves and Gini coefficients for different product-pathogen pairs. Additionally, we explore the implications of using an empirically-derived Lorenz curve instead of one based on statistical fitting. Furthermore, monitoring changes in Gini coefficients across time might prove useful for better understanding national food safety programs. It is possible that the magnitudes of Gini coefficients might provide information on the most appropriate risk management strategies for reducing the risk of human illness attributed to foods.

2. Methods

In previous risk assessment work, FSIS has fit beta-binomial distributions to meat and poultry sampling data to describe how the proportion of pathogen-positive samples varies across all slaughter establishments within an industry (Williams et al., 2013). FSIS has estimated these distributions for its *Salmonella* sampling data from comminuted chicken, comminuted turkey, and chicken parts (FSIS, 2015). Similar analyses are also completed for comminuted beef and beef carcasses, as well as for *Campylobacter* in chicken and turkey. The data were summarized (i.e., number of a particular sample type collected and number found positive for a particular pathogen) for each slaughter or processing establishment in which samples were collected. A weight that reflected the share of total industry production of the sampled product was also estimated based on each establishment's daily production volume.

FSIS inspection personnel collected comminuted chicken and turkey, chicken parts, ground beef and beef carcass samples. Comminuted poultry data were collected during a June 2013 through January 2014 exploratory sampling program (FSIS, 2015). Comminuted beef data were collected between July 1, 2014 and June 30, 2015. All of the comminuted or ground samples consisted of 325 g of ground material. Chicken parts samples—comprising rinse samples of 1.8 kg (4 lbs) of legs, breast and wings—were collected during a baseline survey conducted from January through August of 2012 (FSIS, 2013b, 2015). For beef carcasses, inspection personnel surface swabbed an 8000 cm2 area on one side of a carcass immediately after hide removal (post-dehiding) and on the other side just before finished carcasses entered the chiller (pre-chill). FSIS personnel collected the beef carcass samples during a baseline survey conducted from August 2014 through July 2015. All microbiologic detection of pathogens was conducted in FSIS laboratories according to published methods (FSIS, 2011, 2013a, 2014). Results for the following nine product–pathogen pairs are analyzed: comminuted chicken-*Salmonella*, comminuted chicken-*Campylobacter*, comminuted turkey-*Salmonella*, comminuted turkey-*Campylobacter*, comminuted beef-*Salmonella*, chicken parts-*Salmonella*, chicken parts-*Campylobacter*, beef carcasses post-dehiding-*Salmonella*, beef carcasses pre-chill-*Salmonella*.

A weighted beta-binomial fitting algorithm was used to generate maximum likelihood estimates of beta distribution parameters. These parameters determine how the within-establishment proportion positive varies across the industry (Williams et al., 2013). The beta-binomial fitting algorithm accounts for both the binomial sampling variability within each establishment and the variability in the underlying probability of positive samples across establishments. Weighting was necessary because total production volume for an industry is heterogeneously distributed across the industry's establishments.

A Lorenz curve can be derived directly from a distribution describing how the within-establishment proportion of contaminated units varies across the industry. Let $F(x)$ be the beta distribution's cumulative probability for the within-establishment contamination fraction (such that $0 \le x \le 1$). If $p = F(x)$, then the Lorenz curve $L(p)$ for a continuous random variable is defined as

$$L(p) = \int_0^x \frac{t f(t) dt}{\mu},$$

where μ is the expected value of the distribution. The value of $L(p)$ represents the cumulative fraction of all contamination within an industry contained by the p^{th} percentile of the industry. For example, if all establishments had the same positive sample fraction, then $L(0.5) = 0.5$ (i.e., 50% of all contamination is contained within 50% of an industry's production volume).

Practically, for a large sample from the random variable X, points on the Lorenz curve can be approximated by sorting the sample from smallest to largest values, determining the cumulative probability for each value (as the x-axis value of a Lorenz graph) and, for a particular x_k value in the ordered vector of sampled values, calculating

$$L(p_k) = \frac{\sum_{i=1}^{i=k} x_i}{n\mu},$$

where n is the sample size. The latter approach is used here and programmed in R (R Development Core Team, 2015). Sample sizes of 1,00,000 iterations were found to provide sufficiently stable results.

Given a Lorenz curve, the Gini coefficient measures the ratio of the area between a 45° line and the Lorenz curve to the total area of the space (i.e., 0.5). In our practical derivation of the Lorenz curve, the Gini coefficient equals

$$G = \frac{0.5 - \frac{\sum_{i=1}^{i=n} x_i}{n}}{0.5}.$$

In lieu of a fitted parametric distribution, a Lorenz curve can be derived directly from data by binning the ordered results. Nevertheless, because within-establishment sample sizes vary across establishments, binning is based on the positive sample fraction observed for each establishment (i.e., $\tilde{x} = \frac{s}{m}$ where $\tilde{x}$ is the estimated positive sample fraction for an establishment, s is the number of positive samples and m is the number of samples collected in the establishment). Pairing the estimated positive sample fraction with the weighted production volume for each establishment allows calculation of $L(p)$ for each bin of unique $\tilde{x}$ values. Linear interpolation between the discrete values supports a graphical depiction of $L(p)$. The Gini coefficient is calculated by summing the

Please cite this article as: N. Lukicheva et al., Characterizing the concentration of pathogen occurrence across meat and poultry industries, Microbial Risk Analysis (2016), http://dx.doi.org/10.1016/j.mran.2016.05.002