How Would You Analyze Census Data?

WE HAD DATA FROM A CENSUS SURVEY, SO WE TOOK RANDOM SAMPLES FROM CENSUS DATA

We wanted to make inferences about meaningful associations between 30-day ENDS use and other risky behaviors for 10th graders in Wyoming, using data from the 2016 Prevention Needs Assessment (PNA) Survey. But, the 2016 PNA is an attempted census, not a random or representative sample of students. We could have described associations with subjective comparisons without any statistical tests. Instead, we

- Talked to statisticians for ideas,
    - Treated the 2016 PNA as a sampling frame, and
    - Took simple random samples (SRS without replacement) from it.

Sampling steps

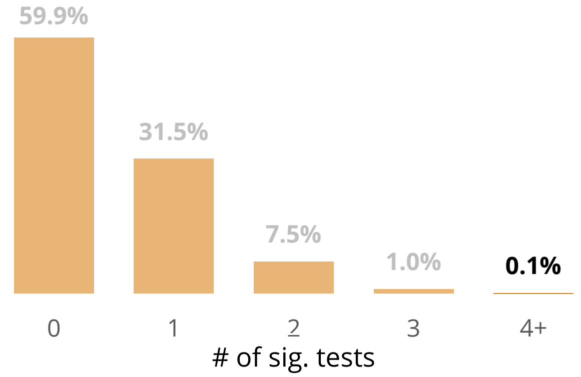
- Clean the data if needed.
- If you want to draw the same samples using syntax every time, follow the steps:
    - Sort the cases in the data in a known, reproducible order. Do not reorder the cases once the order is set.
    - Set a seed value for the random number generator.
        - We used random.org and set Min = 1 and Max = 1000000000 to obtain a random integer.
    - Generate and assign a random number to each case.
- Split the data into 10 samples of roughly equal size. (We determined 10 samples based on our data file and consultation with a statistician. Your needs may differ.)
- We used Stata, and the following blog post is helpful in understanding and implementing our random sampling without replacement: https://blog.stata.com/2012/08/03/using-statas-random-number-generators-part-2-drawing-without-replacement/.

WE REPEATED THE SAME ANALYSIS WITH EACH SAMPLE

Using each sample, we produced a 2x2 contingency table cross-classifying ENDS use and another risky behavior and computed a chi-square value and its p-value for a test of independence using alpha = 0.05. We repeated the same analysis with each of the samples. At this point, we had the result of the test from each of the samples. Some tests were statistically significant while other were not. We also checked the direction of significant relationships.

But, now what? We needed a way to set a threshold for determining how many significant tests from the 10 samples were enough to claim a meaningful relationship. After consulting with a statistician, we used a binomial probability test to set the threshold. Based on a binomial distribution with n = 10 (samples) and the base rate p = .05, we determined 4+ significant tests to claim the significant association. Under these parameters, it is unlikely (<1% probability by chance) to get 4+ significant tests from 10 samples.

Binomial Probability Distribution Given n=10 & p=0.05



For reporting estimates to the client, we re-ran the analyses for the significant relationships on the full dataset.

LESSONS LEARNED & TIPS

- You can sample from you census data.
- Consult with a statistician to determine number of samples and operationalize meaningful.
- With this approach, your census data becomes a sampling frame.
- Use syntax for reproducibility of testing. Don't rely on the point-and-click approach.
- Automate the process of sampling and analysis.
- Other analysis methods such as logistic regression can follow this approach.
- You might not need this approach if you know effect size of practical significance for your topic.

Muneyuki Kato shinze@uwyo.edu

Eric Canen

Laran Despain

This work was completed under contract to the Wyoming Department of Health. This publication was supported by tobacco settlement funds. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Wyoming Department of Health.