in this issue >>>

A Quarterly Insight on the Services our Division Offers

DEB quarterly

Bulletpr00f Stats

Experimental design and sample size calculations: a closer look at clustering

By: Joseph Billian, MS

One of the most frequent forms of assistance provided to researchers at WMed by biostatisticians is a sample size calculation. That is, the calculation of the smallest sample size necessary for statistical significance, given the hypothesized value being tested against (e.g. μ_0), an estimate of the sample effect size (e.g. x), the variability of the sample estimate (e.g. σ), the significance level of the test (α), and the power of the test ($1-\beta$). These five parameters are required for the most basic sample size calculation.

Real world experiments are hardly ever so simple. Many questions about the experimental design will affect the sample size calculation. Is the endpoint measurement continuous, binary, a count number, or perhaps time-to-event? Is the experiment trying to show a bi-directional difference, a one-directional difference, equivalence, or non-inferiority? Is there one sample or two? If there are two samples, are they dependent? That is, can we pair-up the data, as in a before-and-after study?

One common and complex aspect of experimental design is clustering, which refers to grouping measurements together which tend to be correlated due to their natural relationship. Students may be clustered within classes. Patients may be clustered within hospitals. Lab mice born to the same mother are clustered together. Clustering of data increases the minimum sample size. This increase is often described as a variance inflation factor (VIF). Since sample size varies directly with variance, the VIF describes the necessary "inflation" of the sample size. The VIF depends upon several different parameters.

First one must ask if the clusters are all the same size or if their sizes vary. In the former case, the VIF depends on the fixed cluster size. In the latter case, the VIF depends on both the average cluster size and the variability of the cluster size. Second, if the experiment involves random assignment of treatments, then one must ask whether entire clusters are to be randomly assigned to treatment arms or whether subjects within clusters are to be randomly assigned to treatment arms. In the former case, the VIF depends on the intra-cluster correlation (ICC). In the latter case, there are two types of ICC upon which the VIF depends: ICC reflecting the variability between clusters as a random effect and ICC reflecting the interaction between cluster and treatment effect. (In the former case, the ICC is only reflective of the interaction between cluster and treatment because of the way entire clusters are assigned to treatment arms.)

Other aspects of experimental design also require a VIF, such as stratification of data, non-compliance, or losses-to-follow-up. When researchers consult with a biostatistician from DEB, a great deal of consideration is given to the experimental design in order to make sure that the sample size calculations accurately reflect the experiment to be performed.

A Learning Expedition on Outbreak Investigations

By: Mireya Diaz, PhD

Recently, I had the pleasure to embark in the fascinating endeavor of creating projects for the M1 class so they could have an on-hands learning experience about outbreak investigations. This was motivated by my previous exposure in collaboration with infectious disease investigators in diseases such as HIV-AIDS, malaria, tuberculosis, and West Nile Virus, in which I learned by doing. Little did I know what I would gain in this learning journey!

For example, I have learned that although media display many patients with Ebola bleeding all over their bodies, only about 11%-36% of patients with Ebola actually experience bleeding (Oza et al. 20017; Arranz et al. 2016). I have also learned about the CDC program "Epidemic Intelligence Service." This is a two-year service fellowship in which trainees learn Applied Epidemiology 90% of the time by doing it. Graduates become "disease detectives." This training is not limited to infectious diseases but also include chronic diseases, environmental and occupational health threats, injury investigations. The program holds a free annual conference open to the public usually in late April-early May in which officers of the program present their recent work.

I also learned how much Epidemiology has advanced such that the traditional methods for outbreak investigations are nowadays supplemented with sophisticated molecular tools, social network tracing, and social media tweets. The molecular tools can assist in discerning cases actually belonging to an outbreak from those who are phenotypically identical, and otherwise would be counted as part of the outbreak, creating obstacles to determine the real source(s) and transmission chain(s) (Gardy et al. 2011). Tuberculosis detection has been the champion underneath the developing of the genotyping tool (EID 2002;8(11)). Social network tracing supplements or even substitutes traditional contact tracing. The former is particularly useful when contact tracing fails to find the sources but cases continue to appear, or in populations difficult to reach (Gardy et al. 2011). Social media is used two-fold. It is used as a means to inform the community about outbreaks or as a surveillance tool. For the former we have the examples of outbreak communication about Zika virus in Singapore through Facebook (Lwin et al. 2018), and twitters around the measles outbreak originated in Disneyland (Tang et al. 2018). For the latter there are the examples of Twitter data and detection of influenza (Allen et al. 2016), and Google searches for identifying conjunctivitis epidemics worldwide (Deiner et al. 2019). A study performed with the 2014-2015 flu season identified that Tweeter data obtained correlation coefficients between the Tweeters and official reports from the CDC that ranged between 0.44 and 0.93. Temporal searches in Google can detect up to 80% of outbreaks before the report's issuance date.

There are plenty of sources of data about many ongoing outbreaks which can be followed as they occur. Examples of these sources are: CDC's Nationally Notifiable Infectious Diseases and Conditions, National Outbreak Reporting System and ArboNET; WHO's External Situation Reports; states' Departments of Health.

The projects for the M1 class are geared towards the students being able to perform in teams many of the steps necessary to conduct an outbreak investigation. These steps are summarized in the figure to the right. They also have the task to use the medical knowledge they have gained so far and perform a differential diagnosis trying to figure out what particular condition is afflicting the patients they get to know about. Finding out this gains them access to the data to be able to orient cases in terms of signs and symptoms, place, and time. With the data they can also build an epidemic curve. With this experience, the students will have a glimpse, hopefully in a stimulating fashion, about being *disease detectives*.

STEPS IN OUTBREAK INVESTIGATION

- Verify the diagnosis
- Confirm existence of outbreak
- Provide case definition
- Identify cases
- Use descriptive epidemiology (orient to time, place, person; epi curve)
- Develop hypotheses about likely source, transmission routes
- Test the hypotheses
- Refine the hypotheses
- Implement control (and prevention) measures
- Communicate the findings

Arranz J, Lundeby KM, Hassan S, et al. BMC Infectious Diseases 2016;16:308.
Oza S, Sesay AA, Russell NJ, et al. EID 2017;23:1792-9.
Gardy JL, Johnston JC, Ho Sui SJ, et al. NEJM 2011;364:730-9.
Emerging Infectious Diseases 2002;8(11): Tuberculosis Genotyping Issue.
Lwin Mo, Lu J, Sheldenkar A, Schulz PJ. Int J Environ Res Public Health 2018;15:1974
Tang L, Bie B, Zhi D. Am J Infect Control 2018;46:1375-80.
Allen C, Tsou M-H, Aslam A, et al. PLoS ONE 2016;11:e0157734.
Deiner MS, McLeod SD, Wong J, et al. Ophthalmology 2019;126:1219-29.

Research Project Funding: Where to Look

Writers Block
By: Dr. Laura Bauler, PhD

Finding the right granting agency is essential to successfully funding a research project. Funded projects align with the mission of the granting agency, meaning the grant must be written with the goals of the funder in mind. For any project there are likely a number of grating agencies that could support the research, however identifying those granting agencies can be a challenge. There are two main sources of grants: Federal grants and Foundation grants. Federal grants fund projects that work towards the benefit of the public at large, while foundation grants may support a specific population or have a specific goal in mind.

Federal Grants
Federal grants are one of the largest sources of funding for research in the United States. This includes the U.S. Department of Education, U.S. Department of Defense, the National Science Foundation, and the U.S. Department of Health and Human Services, which funds the National Institutes of Health. Funding from these agencies comes in many shapes and sizes from contracts awarded to carry out a certain project with a deliverable output, to five-year grants utilized for research. All federal grant funding opportunity announcements (FOA) can be found at www.grants.gov.

Foundations
Many non-profit agencies and foundations fund research, including the Bill and Melinda Gates Foundation, the United Way, and the Robert Wood Johnson Foundation. There are also many disease specific foundations that support medical or basic science research, such as the Cystic Fibrosis Foundation, the American Heart Association, or the American Cancer Society. A list of nonprofit organizations can be found at https://www.guidestar.org/, a subscription is required for full access to their information and search engine. For more foundations search https://www.insidephilanthropy.com/ find-a-grant. They have also generated lists of Science Research Funders and Disease specific Funders.

Local Foundations
W.K. Kellogg Foundation
Kalamazoo Community Foundation
United Way Battle Creek Kalamazoo (UWBCK)
The Rhonda E. Stryker and William D. Johnston Foundation
The American Heart Association

Internal Funding
The WMed Office of Research supports pilot research projects. Applications are accepted annually in late summer to support preliminary research projects with a maximum budget of $10,000 for individual projects or $20,000 for collaborative pilot projects. The goal of these proposals is to engage students and residents in research while developing research programs that can be successful in applying for external grant funding.

WMed Resources
For help navigating the granting process contact the Sponsored Programs Administration (SPA), who can assist with budget and protocol development, identification of collaborators, and the administrative submission of the grant. Laura Bauler, PhD, a medical editor and faculty member at WMed is available to help edit grants prior to submission by providing an internal peer review. Early contact with SPA and Dr. Bauler is best for optimal assistance with grant proposals.

Data Bytes

The Importance of Avoiding Free Text in Data Collection

By: Melissa Sherfield
Database Specialist

When creating a survey or data entry form it is common for us to receive requests that we add a text box so that the participant can put whatever answer they want. For example, say you have a study that is looking at the effects of a medication used on a patient with a certain type of infection and how it affects future infections. Many times, we are asked in these studies to include a box for the doctor to enter in the medications the patient is already on, and then to enter in the medication they are testing.

What's the problem? You ask. Let me show you. Recently I was working on a dataset from the State of Michigan regarding STDs. Because 58 of the 99 fields -more than half- were free text boxes, it took 30 days or 240 hours to clean that data to utilize it and upload it to a database. Here's why.

When asked to enter the referring Doctor's FIRST name, the answers looked like this:
(Keep in mind these all correspond to the same Dr.)

Dr., B, Barb, Barbar, Barbara, Dr B, Dr Crunch

When asked to enter the lab where the samples were sent many entered Borgess as the hospital however, the free text looked like this:

Bo, Borg, Brogess, Brog, Birgess, Borges, Borgess, BMH (this could be Borgess or Bronson), and Bogess

When asked to specify the drug used to treat the patient the same drug was entered as:

ZMAX, ZITHRO, ZITHROMAX, AZI, AZITHROMYCIN and AZITHROMAX

This field also included dates, doses, symptoms such as vomiting, and location meds were given etc.

As you can see, what would seem as a simple question is open to a wide variety of answers. Misspellings, abbreviations, even dates, entered quickly and inaccurately (ex// 1/1/2104 or 1/1/0205) can render a whole data collection process totally worthless. A recent article[1] illustrates this problem. Among 20 different medications, mispellings were found at a rate of 28.7 errors/term for a total of 574 mispellings in the set. The answers may not be wrong, but they are difficult to manipulate and analyze. It is best that we simply give the participant a choice to select and not one where they fill in the blank. This allows for a cleaner dataset and a much more accurate and efficient analysis.

[1]A Sarker, G González-Hernández. J Biomed Inf 2018;88:98-107.

final thoughts...

2020 KALAMAZOO COMMUNITY RESEARCH DAY
Prepare your project with time. EPIBIO we will glad to help you. We will be receiving requests until end of December to allow for sufficient time, so you can write your abstract by the deadline.

DEB

1000 Oakland Drive
Kalamazoo, MI 49008

Editor: Mireya Diaz
Contact us at: epibio@med.wmich.edu