SELECTION, SAMPLING AND STATISTICAL ANALYSIS IN MEDICAL AND ETHNOGRAPHIC RESEARCH

A Research Paper

Presented to

Dr. David Sills

The Southern Baptist Theological Seminary

In Partial Fulfillment

of the Requirements for 86110

---

by

Mark David Harris

Goodmedicine1@verizon.net

April 26, 2015

SELECTION, SAMPLING AND STATISTICAL ANALYSIS IN MEDICAL AND ETHNOGRAPHIC RESEARCH

Introduction

The universe is full of data. In every field there are more facts than a person could master in a million lifetimes. The Free Dictionary defines data as "Facts that can be analyzed or used in an effort to gain knowledge or make decisions; information."[1] However individual facts must be analyzed with other facts, often in sets, to provide understanding. For example, if a couple wants to buy a new car, they usually want to know more than its color. The style, the seating capacity, the fuel economy, and the safety information are also important. The couple will analyze these individual facts, perhaps by listing the results for each car they are considering and ranking the score and the importance, to make a decision.

Medical researchers want to understand the impact of various interventions on human disease. A study team may record the blood pressure (BP) of 100 patients taking a new antihypertensive medication, Medication A (the intervention group), and comparing those results with those of 100 patients taking a placebo (the control group) to discover if Medication A reduces BP. Blood pressure readings are numerical data, so the scientists would use quantitative measures for central tendency (such as Mean) and dispersion (such as standard deviation) to describe the data. Then they would use quantitative methods such as the Student's T-Test to compare the intervention and control groups to see if there was a significant difference.

Ethnographers want to understand cultures and subcultures. They generate quantitative data and use quantitative measures and methods, typically in describing the parameters of a people group. However, they will also generate qualitative data and use qualitative measures to

[1] *American Heritage Dictionary of the English Language*, Fifth ed., s.v. "Data," accessed April 22, 2015, http://www.thefreedictionary.com/data.

analyze them, since objects and stories are less easily put into a numerical format and analyzed mathematically. The purpose of this paper is to describe quantitative data, measures and methods, to teach readers how to analyze facts and sets of facts, and in so doing help them provide useful information to their audience.

How to Get Good Quantitative Data

Selection

The first task of any researcher is to identify his research question; "What do I want to discover?" The second task is to answer that question. Medical scientists must select the right set of people that will allow them to answer their research question both for people in the study and for people outside the study. The medical researcher above wanted to know if Medication A was effective in reducing BP. He could pick any group of people whose BP needed to be reduced, split them into groups of those taking the medicine and those not taking it, do they study, and compare the results. If his methods were sound, his results would be true for those people in the study; his research would have internal validity. However, few physicians, drug manufacturers or insurance companies care if Medication A works for 100 people (the group taking the medicine); they want to know if it works for 1,000,000 people. Most stakeholders outside the research team want to know if the study was internally valid; were the findings true for those in the study? But they also want to know if it was externally valid (generalizable); are the study findings true for the general population of people with high BP?[2]

The most common way to ensure that research findings will be generalizable to the larger population is to study a large group of people and to assign study subjects randomly to each group (randomization). Doing so decreases the chance that more people in one group than the other will have a condition that confounds (or confuses) the relationship between the intervention

[2]Stephen B. Hulley et al., *Designing Clinical Research*, 3rd ed. (Philadelphia, PA: LWW, 2007), 17

(medication A) and the condition (high BP). Selection bias occurs when subjects are not assigned randomly to each group.[3]

Suppose an ethnographer wants to study the culture of inmates in a prison in Tennessee, and suppose that she only has one day, a Sunday, to do her field work. She might decide to interview attendees at the prison chapel since she could get many interviews in a short period of time. If this ethnographer's interviewing techniques and other methods were good, her findings would be internally valid; true for the subjects in the study. However her findings may not be externally valid (generalizable to other inmates), since her sample only consisted of people who went to chapel, a group likely to be different from the rest of the prison population. She could avoid this mistake by doing the interviews in the cafeteria (where presumably every inmate has to go) instead of the chapel.

Sampling

The most authoritative way to get information about any group is to obtain it from every member in the group; the entire population. If we checked the BP of every citizen in the United States, we would have an excellent understanding of the average and range of BP of people in our country. Unfortunately it is practically impossible for any researcher or group to check the BP of 310 million people. Therefore researchers take samples of the population in the hopes that the sample data will be about the same as the population data. A physician-researcher team may sample 3000 people, a large sample but still only $1/100{,}000^{th}$ of the entire US population, in the hopes that those 3000 have the same BP average and range as the larger group.

A team of ethnographers and support staff studying culture may be able to sample every member of a people group of 50 located on one isolated island in the Pacific, but it could not do

[3] Joseph LaDou and Robert Harrison, eds., *Current Diagnosis and Treatment: Occupational and Environmental Medicine*, 5th ed. (New York: McGraw-Hill Medical, 2014), 861.

the same for a people group of 5,000. Therefore ethnographers may also take a sample from the overall population.

Convenience samples, getting information from whoever is most accessible to the researcher, are the most common type. Unfortunately they tend to have the worst generalizability. If an investigator interviews people at a mall every Tuesday from 0900 to 1100, she will get many without full time jobs and miss most with them.

Cluster sampling involves getting information from natural groupings in a population. For example, researchers may study students from 10 schools in a school district of 50 schools. This is effective if the groups (in this case, schools) are chosen wisely and the sampling of students within the school is random.

A sample is random when every member of the population from which it is taken has an equal chance of being selected. As noted, this type of sampling has the least chance of leading to biased results since the confounders are more likely to be spread evenly between the intervention and the control groups.[4]

Ethnographic research is often more qualitative than quantitative and usually does not have intervention and control groups, so random sampling is less important. However, internal and external validity still apply and therefore sampling is still an important factor in the success of the study. Investigators should sample key leaders and people from both sexes, all ages, and all walks of life to best answer most ethnographic questions.

Daily logs are a common way to gain information both in medicine and in ethnographic work. Physicians frequently ask patients with insomnia to record how they sleep (bed times, get up times, awakenings, daytime sleepiness), patients with incontinence to record their bathroom

[4]Robert B. Wallace, ed., *Wallace/Maxcy-Rosenau-Last Public Health and Preventive Medicine,* 15th ed. (New York: McGraw-Hill Medical, 2008), 21.

visits and “accidents”, and obese patients to record their diet and exercise. Ethnographers sometimes ask subjects to record key events in a certain time span (day, week, etc.) and compare the results. This can paint a powerful picture of activities across a culture.

Types of Quantitative Data

Once study subjects have been selected, the sampling is done and the subjects have been randomized to at least two groups, the intervention group(s) and the control group, investigators perform the study and get results. In quantitative research there are usually two variables, an independent one and a dependent one. The independent (X axis) variable does not depend on the other variable but the dependent (Y axis) variable does. For example, suppose a research team gives subjects a placebo or one of three different doses of medication A (25, 50 and 75 mg). The independent variable is the dose of the medication and the dependent variable is the patient’s BP.

At first no one knows what the outcome will be and so the inputs to the project are called variables. Once the study is done the variables are known and are called data. For example, investigators do not know the BP of each subject after exposure to Medication A before the study, so the BP is called a variable. Once the researchers have discovered the values, subject #34 might have a BP of 119/64 while subject #78 might have a BP of 176/112. These variables have become data, and there are many different types:

1. Nominal data – categorical data that are problematic to rank such as blood type or skin color. It would be difficult to say that type A blood is better than type B or type O.
2. Binary data – data in only two categories, such as alive/dead, male/female or sick/well.
3. Ordinal data – categorical data that can be reasonably ranked. The academic ranks of instructor, assistant professor, associate professor and professor are hierarchical and could be studied as ordinal data.

4. Continuous data – data that are measured on continuous scales such as BP or income. Each point on the scale is the same interval (size) as each other point on the scale. For example, if a man has an income of $10,000 per year, the interval between $1 and $2 is the same as the interval between $9,999 and $10,000. Continuous data are generally discreet, with clear boundaries for each value. Each type of data is analyzed differently[5]:

First variable	**Second variable**	**Example**	**Appropriate test for significance**
Continuous	Continuous	Age and blood pressure	Correlation coefficient or linear regression
Continuous	Nominal	Income and occupation	ANOVA (F-test)
Continuous	Dichotomous	Weight and sex	Student's T-test
Ordinal	Nominal	Difference in satisfaction (such as a Likert scale) before and after an intervention	Kruskal-Willis test
Nominal	Nominal	Ethnicity and blood type	Chi-square test
Table 1 – Types of Tests			

Continuous data like BP can be analyzed with tests designed for continuous data such as linear regression, or they can be placed into categories and analyzed like nominal or even binary data. Nominal data cannot be analyzed like continuous data, but some try to analyze ordinal data the same as continuous data. A common example is the Likert scale which is used for surveys. The scale usually assigns values as follows:

1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Table 2 – the Likert Scale				

The results from such a survey are clearly ordinal data (capable of being logically ranked), the distance in meaning between any two numbers is not reliably the same as the distance between two other numbers. Therefore they must be analyzed as such.

[5] David L. Katz MD MPH et al., *Jekel's Epidemiology, Biostatistics, Preventive Medicine, and Public Health*, 4th ed. (London: Saunders, 2014), 35.

Statistical Measures

Any set of numerical data can be plotted on a two dimensional (X/Y) graph, in which the value of each piece of data is on the X axis and the number of times that value occurs is on the Y axis.

Measures of Central Tendency

When examining these plots, researchers first identify which values are most prevalent, known as the central tendency. There are three main measures of central tendency:[6]

1. Mean – the total sum of the values divided by the number of data points.
2. Median – the middle value in the set.
3. Mode – the value in the set that occurs most often.

A graph of distributions of income in 100 person samples in two small towns may look like this:

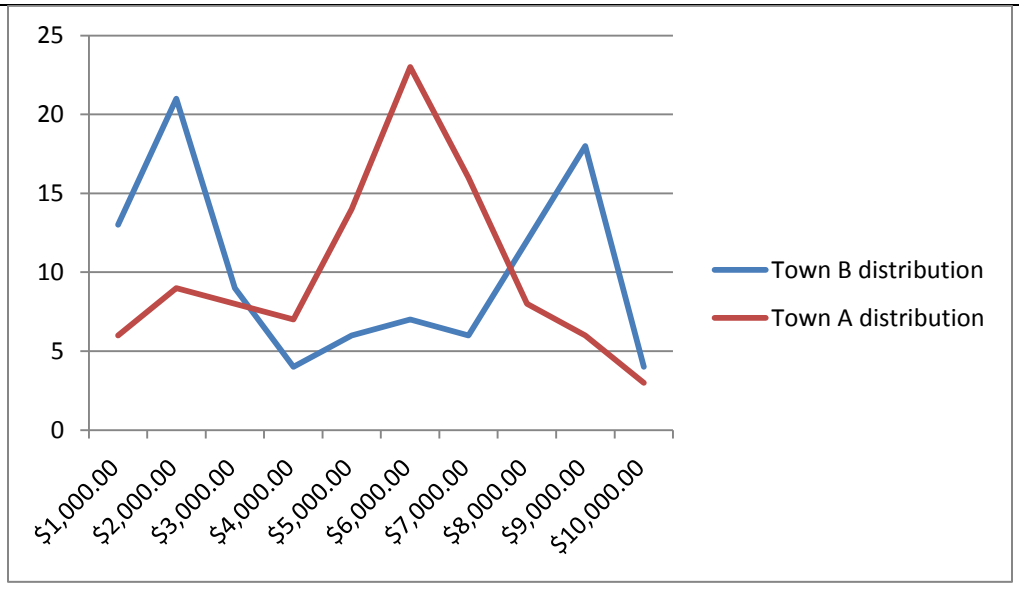

Figure 1 – A sample graph of annual incomes in two towns

[6] Bernard Rosner, ed., *Fundamentals of Biostatistics*, 5th ed. (Pacific Grove, Ca.: Duxbury Thomson Learning, 2000), 9-16.

In this example, the mean income of people in town A is $5440, the median is $4,200 and the mode is $6000. Town A has a normal (bell curve) distribution of a developed nation. The middle class is strong, and there are few people either rich or poor. Normal distributions are the easiest to analyze and are very common in data sets.

Town B has a bimodal income distribution in which the mode is $2,000 and $9,000. Mean and median are not reliable measures in such a distribution. These types of distributions occur in groups in which there are a lot of poorer people (34% make $2,000 or less), a small middle class, and a fair number of rich people (22% make $9,000 or more). This could be a less developed nation.

Measures of Dispersion

It is not enough to describe the central tendency of a data set, researchers must also evaluate how much the values in the set vary.[7]

1. Range – the distance between the highest and lowest values.
2. Quartiles – values falling less than 25%, 25-50%, 50-75%, and greater than 75% of the other values in a data set.
3. Variance – the sum of the squared deviations from the mean
4. Standard deviation (SD) – the variance is usually a large number and so mathematically it is easier to use the standard deviation, the square root of the variance.

In the example above, the range of incomes in both towns is $1,000 to $10,000. The standard deviation of incomes in Town A is $406.53. Since SD depends on the mean, it cannot be calculated for Town B. In a data set with a normal distribution, 67% of all values in a data set

[7] Bernard Rosner, ed., *Fundamentals of Biostatistics*, 5th ed. (Pacific Grove, Ca.: Duxbury Thomson Learning, 2000), 18-24.

fall within one SD of the mean, and 95% of all values in the set fall within two SDs of the mean. In Town A, this means that 95% of the people will have an annual income between $4,600 and $6,300. Variance can be great in different data sets, even if they have the same mean.[8] (Figure 2)

Ethnographers studying cultures will have many reasons to identify central tendency and range in the populations in which they work. Age, income, height, weight, building size, distances, and many other important characteristics can be evaluated with these tools.

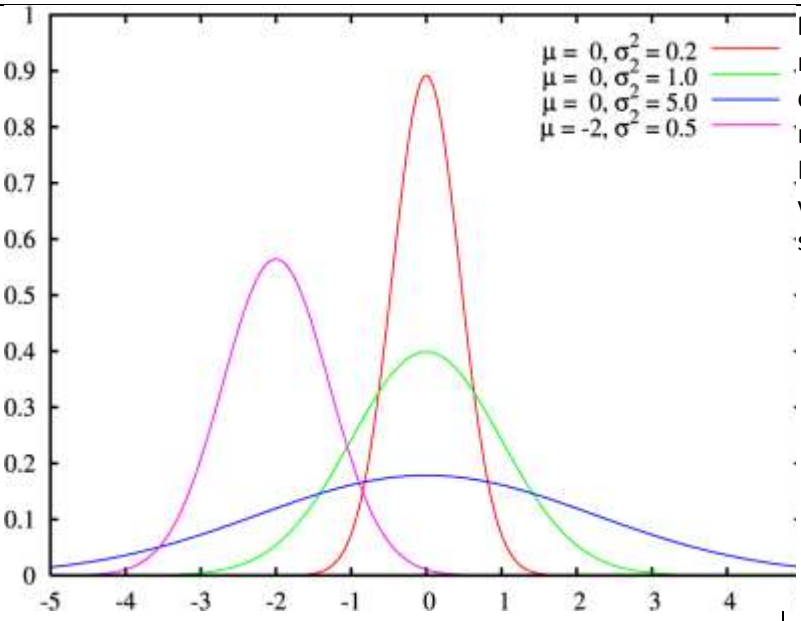

Figure 2[9] – These are all normal distributions (bell curves) with the same number of responses. However the difference in variance between the data sets is great.

Statistical Methods

Most statistical tests compare data sets. Using our blood pressure example, researchers might compare the group that took Medication A with the group that took a placebo. If the distributions were standard (bell shaped), they might use a student's T-test to compare the data and see if the Medication A sample had BP values better, worse or the same as the Placebo

[8] Stanton A. Glantz, *Primer of Biostatistics*, 6th ed. (New York: McGraw-Hill Medical, 2005), 38.

[9] *Archivo: Normal Distribution Pdf.png*, Wikipedia: La encyclopedia libre, accessed April 22, 2015, http://es.wikipedia.org/wiki/Archivo:Normal_distribution_pdf.png.

sample. If the researchers had multiple treatment groups (different doses of Medication A), they would use Analysis of Variance ANOVA.

Studies generally test the null hypothesis. The null hypothesis is usually that there is no difference between the groups being studied. In the Blood Pressure study comparing Medication A and the placebo, the null hypothesis would be that there is no difference in blood pressure between the group of subjects taking Medication A and the group of subjects taking the placebo. If the study did not find a statistically significant difference, researchers would accept the null hypothesis. If it did find a statistically significant difference, they would reject the null hypothesis.

Any study has the risk of making an error. A type 1 error is to find a difference, to reject the null hypothesis, when there is no real difference. Researchers who make this error would believe that Medication A was effective when it actually was not. A type 2 error is to fail to find a difference, to accept the null hypothesis, when a difference actually exists.[10] Researchers who make this error would believe that Medication A was not effective when it actually was. Ethnographic researchers will use the same tests as medical researchers, but their data sets will be different. When using quantitative data, the same types of errors apply.

Software

In bygone days data were analyzed by hand, a laborious task. Now many computer software programs exist to help researchers make sense of the data that they generate. Microsoft Excel can calculate measures of central tendency, dispersion, and basic comparative tests such as the Student's T-Test and ANOVA. Other statistical packages such as SAS, SPSS and Stata can do high level statistics with huge datasets and generate compelling graphics to illustrate results.

[10] Carol V. McKinney, *Globe Trotting in Sandals: a Field Guide to Cultural Research* (Dallas, TX: SIL International, 2000), 119.

Conclusion

Ethnographic research uses a lot of qualitative analysis but quantitative studies are also important. Quantitative information can characterize people and populations and can be used to systematically compare them. It can then be analyzed in mathematically powerful ways. Even more, results can be compared within and even across cultures, and within and across time periods. Quantitative data is not only for medical and social scientists; anthropologists and others involved in this work should be skilled in handling data of all types.

BIBLIOGRAPHY

Katz, David L. MD MPH, Dorothea Wild MD MPH, Joann G. Elmore MD MPH, and Sean C Lucan MD MPH MS. *Jekel's Epidemiology, Biostatistics, Preventive Medicine, and Public Health*. 4th ed. London: Saunders, 2014.

Glantz, Stanton A. *Primer of Biostatistics*. 6th ed. New York: McGraw-Hill Medical, 2005.

Hulley, Stephen B., Steven R. Cummings, Warren S. Browner, Deborah G. Grady, and Thomas B. Newman. *Designing Clinical Research*. 3rd ed. Philadelphia, PA: LWW, 2007.

LaDou, Joseph, and Robert Harrison, eds. *Current Diagnosis and Treatment: Occupational and Environmental Medicine*. 5th ed. New York: McGraw-Hill Medical, 2014.

McKinney, Carol V. *Globe Trotting in Sandals: a Field Guide to Cultural Research*. Dallas, TX: SIL International, 2000.

Rosner, Bernard, ed. *Fundamentals of Biostatistics*. 5th ed. Pacific Grove, Ca.: Duxbury Thomson Learning, 2000.

Wallace, Robert B., ed. *Wallace/maxcy-rosenau-last Public Health and Preventive Medicine*. 15th ed. New York: McGraw-Hill Medical, 2008.