

INCOHERENCE AND THE PARAMETRIC TEST FRAMEWORK: MISCONCEIVED RELATIONSHIPS AMONG SAMPLE, SAMPLING DISTRIBUTION, AND POPULATION

Chong Ho Yu, Barbara Ohlund, Samuel A. DiGangi, & Angel Jannasch, Arizona State University
Chong Ho Yu, Instruction and Research Support 0101, Arizona State University, Tempe AZ 85287

Keywords: Parametric test, sample, population, sampling distributions

Parametric tests are frequently applied by researchers, but some researchers may neither understand the theoretical framework behind parametric tests, nor hold beliefs that are consistent with that framework. The parametric test framework is defined by the relationships among sample, sampling distributions, and population. This article points out several omissions in statistics textbooks and common misconceptions concerning these relationships. It is proposed that these relationships should be taught in a coherent fashion. To provide support for this claim, we (a) reviewed 55 statistics textbooks for various majors such as social sciences and engineering, and (b) administered an online survey specific to the concepts of parametric tests to 34 graduate students who have taken 4.9 undergraduate and graduate statistics courses.

Parametric test framework

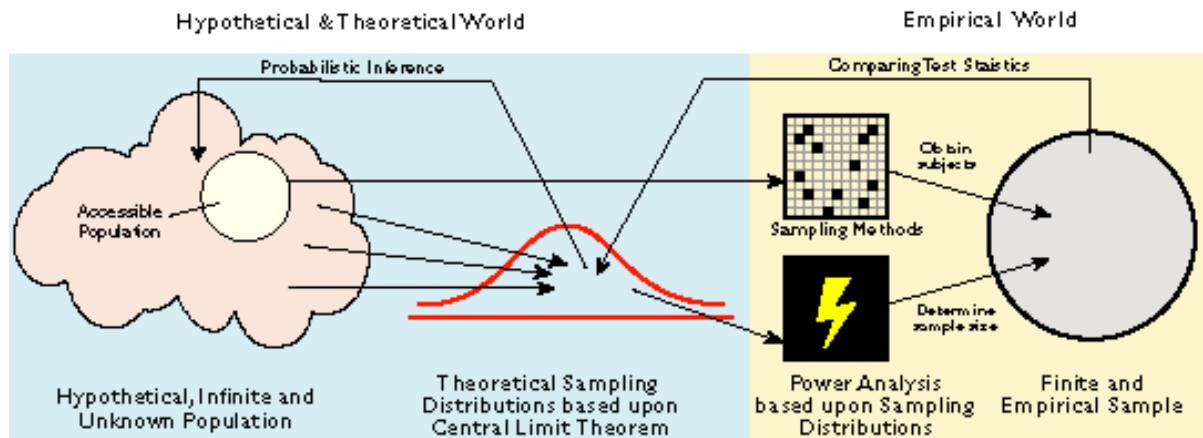
The absence of foundational concepts causes subsequent misconceptions in the interpretation and application of parametric tests. Sixty-two percent of the respondents to our survey did not know what a parametric test was, let alone the assumptions of parametric tests and the criteria of choosing among parametric tests, non-parametric tests, and other data analytical strategies. This lack of awareness of foundational concepts may be traced back to statistics texts. In our textbook review, it was found that only 20 percent of these books explained the term

“parametric tests.” Only one book illustrated a road map of choosing between parametric and non-parametric tests (Sharp, 1979). To address this problem, the following illustration is presented.

Figure 1 illustrates the basic components of the parametric test framework. This framework consists of a theoretical and an empirical world. Although statistical testing appears to be empirical, the foundation is indeed non-empirical. On the theoretical side, there is an infinite population and a sampling distribution, which are the target and the foundation of probabilistic inference, respectively. Probabilistic inference, which leads to a codification of uncertainty by confidence intervals and hypothesis testing, is considered the classical paradigm for parametric tests. This inference rests on the foundation of sampling distributions and the central limit theorem (CLT). In order to generate data for the inference, power analysis and a sampling method are needed on the empirical side. Each component of the framework will be explained in detail.

A parametric test uses the sample statistic to estimate the population parameter. An initial misconception arises from the meaning of “parameter.” According to Webster’s New Word Dictionary (Simon and Schuster, 1991), the term “parameter” denotes a constant with variable values. Another example can be found in computer programming: when a function passes a parameter to another function, the parameter carries an exact value. However, this is not always the case in statistics.

Figure 1. Parametric test framework



Hypothetical and theoretical world

Hypothetical population

Infinite size. Generally speaking, in parametric tests a parameter is introduced as a fixed number that describes the population, but a parameter is viewed by Bayesians as a random variable (Schield, 1997). Indeed, the first statement is correct if the population refers to the accessible population. However, parametric tests start with a hypothetical, infinite population and thus a parameter is hardly a fixed constant. For instance, let us assume that we can measure the height of every American male aged 18 or over. We draw the conclusion that the mean height of these men is 1.51 meters. This mean height is not a fixed constant. Its value will change a second later, since every second thousands of American men die and thousands reach their 18th birthday.

Even if the population size is finite, the population parameter is still not a fixed constant because there are distributions both between people and within people. Since people are different, this between-subject variability forms a distribution. However, the same person also has different task performance levels and attitudes toward an issue at different times. This variability within the same person could also form a distribution. Following this framework, even if the population has a fixed number of members, it could still yield a changing parameter.

Frick (1998) used an example of “the planet of Forty” to illustrate the application of inferences to a finite population. This example could be stretched to illustrate the concept of distribution within. Imagine that in the planet of Forty, there are only 40 residents who can live forever but cannot reproduce offspring. Imagine that their memory can be erased so that a treatment effect will not carry over to the next one. When they are split into two groups and are exposed to two different treatments, are the two mean scores considered fixed parameters? The answer is “no.” A month later when the researcher wipes out what they have learned and asks them to start the experiment over, the scores will vary. This is one of the reasons why statistical tests are still useful even if the researcher has full knowledge of the population. Since there is variability within and between subjects, the researcher needs to know whether the difference is due to chance fluctuations regardless of whether the source of fluctuation is between or within.

A real life example can be found at the debate on university faculty salary equity studies. Haignere, Lin, Eisenberg, and McCarthy (1996) suggested that use of statistical significance is improper while the complete population of faculty members is studied. To counter this argument, Dizinno (1999) stated that the current faculty are only a sample that reflects ongoing, and possible future, salary-setting policies,

and thus they are a sample of the population, not the complete population.

Glass and Hopkins (1996) stated that the population is not actually infinite is of little concern, because unless the ratio between the sample and the population is .05 or greater, the techniques for making inferences to finite populations and those for infinite populations give essentially the same results. Procedurally speaking, the notion that “populations are infinite” is unimportant. Conceptually speaking, the misperception of population as finite and population parameter as a fixed constant has negative consequences, as it leads researchers to seek out an objective, true and final answer that does not exist.

Unfortunately, the difference between infinite and finite populations is not emphasized in most statistics textbooks. Out of fifty-five reviewed books, only sixteen of them (29%) explained the difference.

Unknown distribution. Not only is this hypothetical population infinite in size and fluctuating within, it is also unknown in distribution. Contrary to popular belief, the population distribution is not necessarily normal. This leads to another problem. It is a common belief that a random sample represents the target population and thus random sampling is required for a parametric test. However, when the population is infinite, fluctuating, and unknown, there is no way of knowing whether the sample reflects the population even if the sample is random. (Frick, 1998). A noted physicist Jaynes (1995) identified the perceived equivalence between the random sample and the unknown population as the “mind projection fallacy.”

Regardless of the uncertainty of the population, one must start with an accessible population from which random samples are drawn. Problems arise when one regards the known and accessible population as the target population, to which the inference is ultimately made. For example, if a researcher defines the population as all current college students at a university and an inference regarding the effectiveness of web-based instruction is made to this population, does the inference also apply to future students? If the inference is localized to a particular time and space, then the findings of the experiment cannot be used to construct a theory, since theory by definition is predictive in nature (Kerlinger, 1986).

Summary. In summary, accepting the notion of a finite population will lead to four troublesome consequences: (a) when the entire population is accessible, it is believed that there will be no need to conduct statistics tests, (b) the population parameter will be regarded as a fixed constant and the mission of statistics will be seen as the search for one true answer, (c) a random sample will be believed to be

representative of a population (the “mind projection fallacy”), and (d) the generalizability of the inference will be limited to the accessible population and the construction of a universal theory will be crippled.

Survey results. The survey results confirm our suspicion that the preceding concepts are widely misunderstood. Sixty-one percent of respondents realized that a hypothetical population is infinite in size, but only twenty-three percent were aware that the distribution is unknown. Only thirty-eight percent correctly believe that even when a researcher has full access to the entire population, there is still a need to perform a statistical test.

Theoretical sampling distributions

Central limit theorem. Obtaining a true random sample from an infinite, fluctuating, and unknown population is not possible. Under the CLT, limited cases are used to construct a sampling distribution to approximate the center of the population. This theoretical sampling distribution serves as a bridge between an empirical sample and a hypothetical population. The theorem itself is used to justify making inferences from the sample to the population.

Statistical tests are said to be positivist in nature (Suen, 1992). However, sampling distributions exist in theory only. If some things exist in theory, do they really exist? In theory, a normal distribution is based on infinite cases. One can use a supercomputer to simulate a normal distribution, but of course the simulation cannot run forever. The debate regarding the existence of mathematical reality has a long history and remains inconclusive (Penrose, 1989; Yu, 1998). While mathematics is theoretical in essence and thus sampling distributions seem natural to a mathematician, sampling distributions may not correspond to the practical reality which confronts the practitioners (Good, 1994).

Nonetheless, the founder of statistical testing, Fisher (1956) did not view distributions as outcomes of empirical replications that might actually be conducted. He asserted that theoretical sampling distributions, against which observed effects are tested, have no objective reality “being exclusively products of the statistician's imagination through the hypothesis, which he has decided to test.” (p.81).

Non-normal population. The requirement of data normality in parametric tests is grounded in the CLT. However, some researchers mistakenly believe that non-normal data are undesirable for parametric tests, because the data do not resemble a normal population to which observed data are compared (e.g. Siala, 1999). Burrill (1999) pointed out two problems about the preceding notion: (a) not every statistical test requires normally-distributed variables, and (b) not all statistical tests require the scores to be compared to a normal population. One can see how one

misconception could lead to another, and eventually the entire conceptual model could fall apart.

Questionable statements concerning the CLT and normal distribution could be found in statistics texts. For example, a statistical guide for medical researchers stated, “sample values should be compatible with the population (which they represent) having a normal distribution.” (Airman & Bland, 1995, p.298). In fact, the CLT does not assume the normality of the population distribution. The theorem states that a sampling distribution becomes closer to normality as the sample size increases, regardless of the shape of population distribution. Because the shape of the population distribution is unknown and could be non-normal, in parametric tests data normality resembles the sampling distribution, not the population. In other words, a test statistic from the sample will be compared against the sampling distribution rather than against the population.

Normality is a myth. The belief that most populations are normal is hardly an empirical fact. Physicist Lippmann pointed out the circular logic of proving normality: “Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.” (cited in Thompson, 1959, p.121). In a similar vein to Lippmann, Stigler (1986) criticized the “circular” logic employed by Gauss, who developed the normal distribution. Gauss conceptualized the mean in terms of “least squares”: the mean could be used to summarize a data set, because when more observations are closer to the mean and less observations are farther from the mean, the sum of squares of the deviation is minimal. The mean is only “most probable” if the errors (deviations) are normally distributed; and the supposition that errors are normally distributed leads back to least squares. In response to the lack of proof of universal normal distributions, Geary (1947) stated that normality could be viewed as a special case of many distributions rather than a universal property. However, since the school of Fisher became dominant, universal normality has been favored and interest in non-normality has retreated to the background. Geary suggested that future textbooks should include this warning: “Normality is a myth; there never was, and never will be, a normal distribution.” (p.241). However, none of the reviewed texts carry this warning.

Mathematical efficiency. The belief that observed data are compared to a normal distribution is a serious misunderstanding of the role of normal distribution in hypothesis testing. Normal distributions are used because a statistical test

procedure should be “efficient” and “optimal,” in the sense of a high probability of detecting the falseness of a hypothesis when it is indeed false (Kariya & Sinha, 1989). This probability is known as statistical power, which will be discussed in the next section. An optimal test can maximize its power when the normal distribution is assumed as the underlying distribution. In addition, when normality is satisfied, only the first- and second- order moments (mean and variance) are needed to fully describe the distribution of the variables. The third- and fourth-order moments (skewness and kurtosis) are not necessary (West, Finch, & Curran, 1995). Thus, the requirement of normality is not due to an empirical fact, rather it is driven by mathematical efficiency.

Summary. The lack of the knowledge of sampling distributions and the CLT will result in three problems: (a) sampling distributions serve as a foundation for making the leap from sample to population. Without this knowledge, inferences are believed to be made to the sample or there is no justification for the leap from the empirical world to the hypothetical world, (b) statistical testing is believed to be positivist and empirical. Actually, the foundation of statistical testing, which is sampling distribution, is theoretical and cannot be verified empirically, and (c) the sample normality requirement is not driven by empirical facts, but mathematical efficiency. However, normally distributed data are expected to show that the shape of the sample distribution can match that of the population. In fact, the population distribution is unknown and only the sampling distribution is normal.

Survey results. The survey results are not surprising. Forty-one percent of respondents failed to identify the population as the target of inferences. Fifty-six percent mistakenly believed that the hypothetical population must be normal.

Empirical world

Power analysis

Fusion of null and alternate hypotheses. Sampling distributions provide the basis for power analysis. Power analysis, which is based upon the null sampling distributions and the alternate sampling distributions, is applied to determine the proper sample size for a research project, and thereby determines the efficiency of the test. Sampling methods are used to draw subjects from an accessible population. Hence, a finite sample is obtained and empirical data are computed.

However, in Fisherian statistical testing, the null hypothesis is zero effect. The only conclusion after achieving statistical significance is that “the effect is not nil.” Following this strict Fisherian tradition, researchers would find no room for power analysis since statistical power depends on the unknown

alternate distribution (Lehmann, 1993). To rectify this shortcoming, an effect size, which is the standardized distance between the null and the alternate, must be pre-determined. By sketching a distance from the null, the position of the hypothetical alternate is “pinned down.”

Summary. Failure to recognize that power analysis is based on the alternate sampling distribution introduces two problems: (a) power analysis is perceived as an empirical-based procedure on the population and the sample, and (b) It is disconcerting that one looks for a clear-cut answer (reject/not reject the null hypothesis) while conducting a power analysis based on the alternate hypothesis, which is unknown in nature and is only hypothesized by an estimated effect size.

Survey results. The survey results indicate that only thirty-two percent of participants correctly associated power analysis with sampling distributions.

Sampling methods

Randomness as independence. As previously mentioned, it is impossible to obtain a true random sample from an infinite and unknown population and then empirically verify whether the sample could represent the population. Thus, random sampling emphasizes the properties of the sample derived from the sampling process. i.e. One draws a series of values of independent and identically distributed random variables to form a random sample. The keyword of the preceding statement is “independence.”

Many authors define random sampling as a sampling process in which each element within a set has equal chances to be drawn (e.g. Moore & McCabe, 1993; Aczel, 1995). Equality is associated with fairness. This definition contributes to the myth that if the occurrence of a particular event is very frequent, the outcome is considered “unfair” and thus the sample may not be random. This belief also implies that a random sample should reflect the population when every type of member in the population is “fairly” represented.

In reality, complete fairness does not exist. One should not expect that in an urn of balls, small balls have equal probabilities to be sampled as large balls. Even if we put the same size balls in the urn, we cannot “equalize” all factors relevant to the outcome. Jaynes (1995) fully explained this problem: “The probability of drawing any particular ball now depends on details such as the exact size and shape of the urn, the size of balls, the exact way in which the first one was tossed back in, the elastic properties of balls and urn, the coefficients of friction between balls and between ball and urn, the exact way you reach in to draw the second ball, etc..

(Randomization) is deliberating throwing away relevant information when it becomes too complicated for us to handle...For some, declaring a problem to be 'randomized' is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits." (pp. 319-320)

Phenomena appear to occur according to equal chances, but indeed in those incidents there are many hidden biases and thus observers assume that chance alone would decide. Since authentic equality of opportunities and fairness of outcomes are not properties of randomness, a proper definition of random sampling should be a sampling process in which each member within a set has independent chances to be drawn. In other words, the probability of one being sampled is not related to that of others. Hassad (1999) made a very precise statement about the role of probability in sampling, "The probability in sampling takes care of selection bias only. It does not address representativeness."

At the early stage of the development of the concept "randomness," the essence of randomness was tied to independence rather than fair representation. When Fisher and his coworkers introduced randomization into experiments, their motive was not trying to obtain a representative sample. Instead they contended that the value of an experiment depended upon the valid estimation of error (Cowles, 1989). In other words, the errors must be independent rather than systematic.

Summary. The misconception of random sampling as achieving "fairness" and "representation" is tied to this problem: The target population is finite and known, and therefore one can tell how representative a random sample is. The consequence of misunderstanding random sampling will result in a false sense of security: The sample can represent the population and thus the inference is valid.

Survey results. Although the population, to which the inference is made is hypothetical and unknown, the majority of the participants (32%) believed that a random sample could be more representative of the population, depending on the ratio between the sample size and the population size.

Inferences from empirical to theoretical world Probabilistic inferences

As mentioned before, under the framework of parametric tests, the inference should be made to the population from the sample. In statistical testing, a test statistic is extracted out of a finite sample and used to compare against an infinite sampling distribution. The probability (p-value) indicates how likely the result will surface in the long run. In other words, the interpretation of statistical testing should be a probabilistic inference rather than the pursuit of one true answer. Niels Bohr's "Copenhagen

interpretation" is well applied to statistical inference though he was in a different discipline. Bohr asserted that one can answer questions of the form: "If the experiment is performed, what are the possible results and their probabilities?" One should not answer any question in this form: "What is really happening when ...?" (cited in Jaynes, 1995, p.1012).

The impression that a statistical inference leads to one true answer is due to the subsequent action after the rejection or retention of the null hypothesis. When an experiment indicates that there is a significant difference between the mean scores of the control and treatment groups, the policy maker adopts the treatment although there is no logical connection between the action and the inference. As Schield (1997) said, "Probability itself does not lead to action, rather probability justifies confidence and confidence justifies action" (p.3).

As a matter of fact, it is impossible that every study on the same topic can produce the same result. If there is only one true answer, which one is true? On the other hand, the probabilistic nature of inference is compatible with the philosophy of science that research results are tentative and thereby inquiry is a self-correcting process in the long run (Peirce, 1900/1960). Under this premise, inconsistent results from different research studies do not create any logical dilemma.

Summary

The failure of conceptualizing a statistical inference as a probabilistic inference is tied to other misconceptions in sample, sampling distributions, power, and population. In addition, this failure not only leads researchers to have a false sense of certainty, but also leaves no room for harmonizing inconsistent research results.

Survey results

According to the survey, misconceptions specific to inferences seem to be less serious. Only twenty-nine percent of participants misunderstood the meaning of the p-value, and misperceived that the nature of quantitative research, to some degrees, is truth-seeking by giving a definite answer.

Conclusion

The concept of the relationship among sample, population, and sampling distribution is the foundation of subsequent statistical concepts and procedures. Misconceptions in different components of the framework are inter-related. Without a coherent theoretical framework, one may be able to perform statistical procedures correctly, but fail to interpret the result and conceptualize the nature of the inference properly. The following detrimental beliefs are some examples resulting from such an incoherent framework: "The population is finite and inferences are generalized to this population." "The data

distribution are not normal and thus they cannot represent the population. "...etc. All these misconceptions could be boiled down to a common thread: the failure to identify the difference between the theoretical world and the empirical world, and how the researcher could leap back and forth from one to the other. Teaching concepts in a piecemeal manner tends to increase the risk of forming an incoherent framework. Teaching statistical procedures without introducing a unified framework is even worse. It is recommended that a comprehensive and coherent parametric test framework should be learned by statistics students with each component thoroughly explained to ensure a smooth logical flow from one to the other.

Acknowledgements

Special thanks to Ms. Lynn Tobin for reviewing this paper.

References

- Aczel, A. D. (1995). Statistics: Concepts and applications. Chicago: Richard D. Irwin, Inc.
- Airman, D. G. & Bland, J. M. (1995). The normal distribution. British Medical Journal, *310*, 298.
- Burrill, D. (1999, April 8). Re: Normalization. Educational Statistics Discussion List (EDSTAT-L). [Online]. Available E-mail: edstat-1@jse.stat.ncsu.edu [1999, April 8].
- Cowles, M. (1989). Statistics in psychology: An historical perspective. Hillsdale, New Jersey: LEA.
- Dizinno, G. (1999, March 29). Population vs. sample: Implications for salary equity study. Educational Statistics Discussion List (EDSTAT-L). [Online]. Available E-mail: edstat-1@jse.stat.ncsu.edu [1999, March 29].
- Fisher, R. A. (1956). Statistical methods and scientific inference. Edinburgh: Oliver and Boyd.
- Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. Behavior Research Methods, Instruments, & Computers, *30*, 527-535.
- Geary, R. C. (1947). Testing for normality. Biometrika, *34*, 209-241.
- Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in psychology and education (Third edition). Boston, MA: Allyn and Bacon.
- Good, P. (1994). Permutation tests: A practical guide to resampling methods for testing hypotheses. New York: Springer-Verlag.
- Haignere, L, Lin, Y. J., Eisenberg, B., & McCarthy, J. (1996). Pay checks: A guide to achieving salary equity in higher education. Albany, NY: United University Professors.
- Hassad, R. (1999, March 1). Re: Question about Convenience Sampling. Educational Statistics Discussion List (EDSTAT-L). [Online]. Available E-mail: edstat-1@jse.stat.ncsu.edu [1999, March 1].
- Jaynes, E. T. (1995). Probability theory: The logic of science. [On-line] Available URL: <http://omega.math.albany.edu:8008/JaynesBook.html>
- Kariya, T. & Sinha, B. (1989). Robustness of statistical tests. Boston, MA: Academic Press, Inc.
- Kerlinger, F. N. (1986). Foundations of behavioral science. New York: Holt, Rinehart, and Winston.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, *88*, 1242-1249.
- Moore, D. S. & McCabe, G. P. (1993). Introduction to the practice of statistics. New York: W. H. Freeman and Company.
- Peirce, C. S. (1900/1960). Collected papers of Charles Sanders Peirce. Cambridge: Harvard University Press.
- Penrose, R. (1989). The emperor's new mind: Concerning computers, minds, and the laws of physics. Oxford: Oxford University Press.
- Schild, M. (1997). Intepretating statistical confidence. Proceedings of 1997 American Statistical Association Convention. Alexandria, VA: ASA
- Sharp, V.F. (1979). Statistics for the social sciences. Canada : Little, Brown & Company.
- Siala, H. (1999, April 8). Re: Normalization. Educational Statistics Discussion List (EDSTAT-L). [Online]. Available E-mail: edstat-1@jse.stat.ncsu.edu [1999, April 8].
- Simon and Schuster. (1991). Webster's New Word Dictionary. Cleveland, OH: The Author.
- Stigler, S. M. (1986). The history of statistics: The measurement of uncertainty before 1900. Cambridge, MA: The Belknap Press of Harvard University Press.
- Suen, H. K. (1992). Significance testing: Necessary but insufficient. Topics in Early Childhood Special Education, *12*, 66-81.
- Thompson, D. W. (1959). On growth and form. Cambridge: Cambridge University Press.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. H. Hoyle (Eds.), Structural equation modeling: Concepts, issues, and applications (pp.56-75). Thousand Oaks: Sage Publications.
- Yu, C. H. (1998). Mathematical reality: Do theoretical distributions exist? [On-line] Available: http://seamonkey.ed.asu.edu/~alex/computer/sas/math_reality.htm