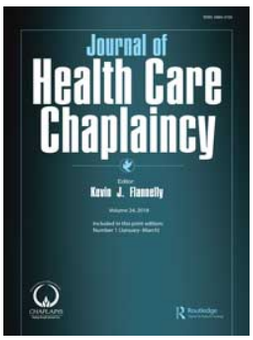

Journal of Health Care Chaplaincy

ISSN: 0885-4726 (Print) 1528-6916 (Online) Journal homepage: https://www.tandfonline.com/loi/whcc20

The *t*-test: An Influential Inferential Tool in Chaplaincy and Other Healthcare Research

Katherine R. B. Jankowski, Kevin J. Flannelly & Laura T. Flannelly

To cite this article: Katherine R. B. Jankowski, Kevin J. Flannelly & Laura T. Flannelly (2018) The *t*-test: An Influential Inferential Tool in Chaplaincy and Other Healthcare Research, Journal of Health Care Chaplaincy, 24:1, 30-39, DOI: 10.1080/08854726.2017.1335050

To link to this article: https://doi.org/10.1080/08854726.2017.1335050

Journal of Health Care Chaplaincy, 24:30–39, 2018

ISSN: 0885-4726 print/1528-6916 online
DOI: 10.1080/08854726.2017.1335050

Research Methodology

The *t*-test: An Influential Inferential Tool in Chaplaincy and Other Healthcare Research

KATHERINE R. B. JANKOWSKI
Iona College, New Rochelle, New York, USA

KEVIN J. FLANNELLY and LAURA T. FLANNELLY
Center for Psychosocial Research, Massapequa, New York, USA

The t-test developed by William S. Gosset (also known as Student's t-test and the two-sample t-test) is commonly used to compare one sample mean on a measure with another sample mean on the same measure. The outcome of the t-test is used to draw inferences about how different the samples are from each other. It is probably one of the most frequently relied upon statistics in inferential research. It is easy to use: a researcher can calculate the statistic with three simple tools: paper, pen, and a calculator. A computer program can quickly calculate the t-test for large samples. The ease of use can result in the misuse of the t-test. This article discusses the development of the original t-test, basic principles of the t-test, two additional types of t-tests (the one-sample t-test and the paired t-test), and recommendations about what to consider when using the t-test to draw inferences in research.

KEYWORDS chaplaincy healthcare, inferential statistic, t-test

INTRODUCTION

"What kind of barley makes a better tasting beer?" This question appears to be unrelated to scientific methodology. Enjoying beer, or not enjoying it, is a

Address correspondence to Katherine R. B. Jankowski, Center for Psychosocial Research, 33 Maple Street, Massapequa, NY 11758, USA. E-mail: krbjankowski@gmail.com

matter of taste! Historically, brewing beer was a tradition-based endeavor. However, in the 1890s, Cecil Guinness and Christopher La Touche of the Guinness Brewing Company changed tradition by hiring chemists to identify what makes the best beer (Box, 1987). William S. Gosset, the man who is known for developing the *t*-test, was one of several chemists hired by Guinness. He, and the other chemists employed by Guinness, began methodically identifying the barley seeds that produced the best barley, and which of those produced the best beer.

It took years for the chemists to identify and measure the qualities of barley seeds associated with the best beer. Even after better seeds were identified, there was much variation in production. One batch of barley grown in one field would produce phenomenal barley while the same seed grown in a different field resulted in unsatisfactory barley. The chemists had to find a way to identify the best barley seed regardless of the variation that is due to things, such as a farmer's field, the weather, the manure used to fertilize it, and the maltster. Many years of planting specific kinds of barley in specific fields provided enough data for Gosset to calculate the average yield and the average amount of variation in the yield. He slowly worked out a mathematical equation to find the best seed and identified which seeds produced better beer than others, on average, even with the variation and the error that was inherent in their experimental research due to sampling and small sample size (Student, 1908). The *t*-test that is used today in scientific research was worked out mainly by Gosset, with some help from mathematicians Karl Pearson and Ronald A. Fisher (Boland, 1984; Box, 1987).

From the Seeds to the Table

The *t*-test that healthcare and other researchers use today has two parts: a mathematical equation that provides a number or value (the *t*-statistic), and a table of all the possible results of that equation (all possible *t*-statistics). The first part, the equation, is a mathematical way to represent the difference between one average (mean) and another number, which can be another average (mean) or a specific value, such as zero, taking into account the variability in the data (see explanation of the mean in Jankowski & Flannelly, 2015, and explanation of variability in K. J. Flannelly, Jankowski, & Flannelly, 2015). The second part, the table, is a standard table that lists all the possible mathematical outcomes of the *t*-test equation (*t*-statistics), and how likely each outcome is (also known as the probability of the outcome, the alpha level, *p* value, or critical value) given the sample size. As a result, the *t*-test is an equation that provides a value and a table that lists how likely that value is in the distribution of such values (the *t*-distribution, which is similar to the normal distribution).

A useful table of the critical values of the *t*-test does not list all of the possible numerical outcomes of the *t*-test equation because the table would

be too large and not very helpful. What is helpful to researchers is knowing when the *t*-test equation produces an unlikely or unusual value of *t* (*t*-statistic). Today, computer programs do the calculations and provide the *t*-test equation result (the value of *t*) and the likelihood of that result (*p*) for a given sample size, which eliminates the need for an actual table of *t* values and their probability (*p*). Researchers, in general, agree that when the *t*-test result (the value of *t*) is unusual, when is expected to occur less than 5% of the time ($p < .05$), 1% of the time ($p < .01$), or less often (e.g., less than 1 time in 1,000; $p < .001$), the *t*-test result is statistically significant. Such unusual *t*-values support the inference that something is very different between the two means (or between a mean and zero) because the *t*-test result is not very likely to occur.

The *t*-test is based, in part, on the logic and mathematical proof of something called the Central Limit Theorem (Pagano, 2012). This theorem states that an entire population is well represented when a large enough sample is randomly drawn from it. In addition, any sample that is large enough and randomly drawn from a population will likely be similar to any other same-sized sample that is randomly drawn from the same population. Mathematically, it can be demonstrated that, if a researcher randomly pulled every possible same-size sample from a population, the means of all these same-size samples would produce a frequency bar graph that resembles a normal distribution or normal curve (the sampling distribution of the mean). Many of the means of the same-size samples would appear more frequently than others in the bar graph, and most would be near the population mean in the center of the curve. Some sample means would occur that are far away from the population mean and they would happen less frequently.

Similarly, the *t*-test equation produces a numerical outcome for a difference between sample means drawn from the same population (Pagano, 2012). A bar graph distribution of all possible *t*-test outcomes from all possible difference tests resembles a normal curve with larger sample sizes. A very unlikely *t*-test value indicates a very large difference between means and occurs away from the center of the bar graph in the tails of the distribution curve (see a description of the normal curve in K. J. Flannelly et al., 2015). In other words, a large difference between the sample means, relative to the variations in the data, yields a high *t*-value, which indicates that the difference between means is unusual and this allows for an inference that perhaps the samples are not from the same population.

TYPES OF *t*-TEST AND THEIR USES

There are three different but related types of *t*-tests, which are: (a) the two-sample or independent-samples *t*-test; (b) the matched- or paired *t*-test; and (c) the one-sample or single-sample *t*-test.

Two-Sample or Independent-Samples *t*-Test

The two-sample or independent samples *t*-test is based on the original *t*-test developed by Gossett. It is often called Student's *t*-test because Gosset published his article about it anonymously under the pseudonym Student (Boland, 1984; Box, 1987; Student, 1908). The two-sample *t*-test is used to compare the means of two samples to see if the difference is unusual and allow for the inference that the samples are not drawn from the same population. Two examples of studies that used two-sample *t*-tests, which we identified from a search of PubMed (see K. J. Flannelly, Jankowski, & Tannenbaum, 2011, about searching PubMed), tested the extent to which different treatment-team approaches improved patient care. One study assessed whether a multidisciplinary intervention improved the quality of life of advanced cancer patients, compared to standard care (Rummans et al., 2006), and the other study assessed the extent to which the implementation of an interdisciplinary care plan improved the end-of-life care of oncology and geriatric patients, compared to usual care (Bookbinder et al., 2005). A more recent study we found on PubMed used a two-sample *t*-test to compare spiritual well-being between samples of patients with generalized anxiety disorder and patients with minor general medical conditions (Ajman & Bokharey, 2015). Examples of articles published in the *Journal of Health Care Chaplaincy* (*JHCC*) include two studies that compared gender differences using two-sample *t*-tests. The first study examined sex differences in pastoral care skills among CPE (clinical pastoral education) students (Jankowski, Vanderwerker, Murphy, Montonye, & Ross, 2008) and the other study examined sex differences in the use of CAM (complementary alternative medical) practices by religious professionals (Jankowski, Silton, Galek, & Montonye, 2010). A third article reported the results of a national survey which found that healthcare chaplains who received workplace support had fewer symptoms of disenfranchised grief than those who did not receive workplace support (Spidell et al., 2011).

Paired or Dependent *t*-Test

The paired or dependent *t*-test is used to examine differences in means of two sets of data that are related to one another (hence, it is also called a correlated *t*-test). Typically, the paired *t*-test is used to compare the mean of a sample of people on a variable at two points in time.

The same type of *t*-test can be used with two samples if individuals in one sample are matched with those in another sample to form pairs based on their similarities on specific characteristics. This kind of matching procedure is mainly performed in epidemiological research (Jewell, 2004; Stewart, 2010; Stroup & Teutsch, 1998), where the paired *t*-test is sometimes referred to as a matched or matched-pairs *t*-test (Stroup & Teutsch, 1998).

The most common use of the paired *t*-test is to assess whether the scores of one sample of people on a scale or other measure have changed over time, as already noted. The paired *t*-test is widely used to examine the effects of an intervention (Katz, 2001) by comparing pretest (i.e., preintervention) and post-test (i.e., postintervention) scores on an *outcome* variable (also called a *dependent* variable, see L. T. Flannelly, Flannelly, & Jankowski, 2014b).

For example, we found two studies on PubMed that used paired *t*-tests to measure the effectiveness of educational programs, including the ability of educational programs to improve the ability of advanced medical students (von Gunten et al., 2012) and respiratory therapists to address end-of-life issues (Brown-Saltzman, Upadhya, Larner, & Wenger, 2010). Another study we found used paired *t*-tests to evaluate the health benefits of a religiously based wellness intervention (Kamieniski, Brown, Mitchell, Perrin, & Dindial, 2000),

There appear to be only two studies published in *JHCC* that used paired *t*-tests, one of which compared the ability of a faith-based chaplaincy intervention to reduce spiritual distress among patients at a U.S. Veteran Affairs Medical Center (Kopacz, Adams, & Searle, 2017). The other is a study that used both two-sample *t*-tests (as previously mentioned) and paired *t*-tests (Jankowski et al., 2008). The paired *t*-tests were used to assess changes between pretest and posttest scores on measures of pastoral skills, emotional intelligence, and self-reflection among CPE students (i.e., their scores before and after taking a unit of CPE).

One-Sample or Single-Sample *t*-Test

The one-sample or single-sample *t*-test is used to test one sample mean against a specific value, such as zero, or a standard or an expected outcome, or against the known population mean. If the *t*-test equation produces a *t*-value that is very unlikely, then the sample mean can be said to be significantly different from the standard mean, such as the normed average of a test. One-sample *t*-tests are often used to compare the mean of a sample of patients on some variable to the mean of the general population on that same variable.

No studies have ever been published in *JHCC* that used a one-sample *t*-test, but we found several interesting studies on PubMed that used it. For example, two recent studies compared the mean of the post-op quality of life of surgical patients in the United States and Italy to the quality of life of the general population in their respective countries (Schiavolin et al., 2015; Steele, Zahr, Kirshbom, Kopf, & Karimi, 2016). Similarly, a Dutch study used a one-sample *t*-test to compare the perceptions of diabetes patients about their state of health to the perceived health of the general population of The Netherlands (Hart, Redekop, Bilo, Berg, & Meyboom-de Jong, 2005).

The reader may think it is odd that this kind of *t*-test is called a single-sample or a one-sample *t*-test, as two means are being compared. However, the term one-sample is based on the fact that the mean of one sample is compared to an accepted standard mean value. In the case of these three studies, the standard to which the sample mean is compared is the mean of the general population.

CONSIDERATIONS LENDING CONFIDENCE TO INFERENCES

Sample Considerations

When using a *t*-test, it is important that the sample of data has 30 data points or more because a sample size of thirty or more will more closely resemble the population and the sampling distribution of the mean of the normal distribution (Triola, 2004). This relates directly to the Central Limit Theorem. A larger sample size helps the researcher have greater confidence that the sample average and variation in the data, closely reflect the population average and variation, unless there is truly something different in the sample due to treatment or other influences. A larger sample size is also needed to detect smaller differences in means (effect sizes). Random sampling from the population is assumed, meaning that the data used for the *t*-test come from a random selection of participants from a larger population. In random sampling, everyone has an equal chance to be selected for the study, and it is also important that everyone has an equal chance to be included in any group in the study (random assignment).

Data Considerations

The data used to calculate the *t*-test must be on at least an interval scale. Ideally, the measurement should provide data that range across at least 11 values (see L. T. Flannelly, Flannelly, & Jankowski, 2014a, for examples of interval and ratio scales). The data must be free from extreme values, also known as outliers. Outliers are found when some participants in a study have scores on a measure that are extremely different from the scores of everyone else. Outliers can be found by arranging all the data in ascending or descending order, or by creating a bar graph. The *t*-test should not be used with outliers in the data because outliers affect the mean and standard deviation, and thereby, affect the accuracy of the *t*-test results. Also, attention should be given to the shape of the distribution of the data by charting the data in a bar graph. The distribution of data in the sample should be balanced around the average of the data in the sample, and there should be only one mode. A distribution of data in this shape, often referred to as a normal distribution or normal curve, can be seen in the distribution of pain scores in K. J. Flannelly et al. (2015; Figure 1). Work by Poncet, Courvoisier, Combescure, and

Perneger (2016) suggests, that in some situations, lack of normality might not negatively affect the interpretation of *t*-test results as dramatically as was once thought.

Reporting and Interpreting the *t*-Test Outcomes

Always report the *t*-test value, the degrees of freedom, and the probability of the observed outcome (*p* value). The degrees of freedom refer to the sample size minus one. It is also important to report the means and standard deviations for the samples tested by the *t*-test. This helps the reader understand and evaluate the results of the *t*-test. It is also becoming accepted practice to report the effect size (Cohen's d; Cohen, 1988), which is the difference between the two means divided by the standard deviation of the combined samples (see the explanation and calculation of the standard deviation in K. J. Flannelly et al., 2015).

Interpretation of the *t*-test result is guided by the probability (*p* value) of the outcome (*t* statistic). If the probability is very small, a researcher can conclude that the difference between sample means is unusual. The *t*-test does not prove anything. It indicates the probability of obtaining the observed difference between the means. When the result is significant, the *t*-test indicates that the outcome happens 5.0%, 2.5%, or 1.0% of the time, or even less often. It is up to the researcher to infer from all of the information regarding the samples if there is truly an important difference between two samples and what that difference means. The *t*-test result is just the beginning to understanding the story of what is being studied.

Interpretation of *t*-test results should also be guided by how many *t*-tests are calculated on the same data. If many *t*-tests are being calculated on the same data it is customary to reduce the critical level for the *t*-test, also known as the alpha level, or *p* level, by dividing the usual .05 by the number of *t*-tests conducted (Gordon, 2012).

Finally, it must always be kept in mind that the results from a single *t*-test in a single study are just the beginning. Science moves forward best by replication of findings, and this means doing more than one study. A similar, second study of the same variables may not necessarily produce the same finding. There is serious difficulty with scientists not being able to replicate the findings of studies, and whole theories built on a single finding might not be supported by further research (see Pashler & Wagenmakers, 2012, for a cautionary tale and Pashler & Harris, 2012, for a fuller explanation of the interpretation of $p < .05$). There are a number of reasons for this recent lack of replication, but going back to the basic understanding of the Central Limit Theorem, some studies might not find similar results because the original results, differences between groups identified by the *t*-test, might just have been due to the fact that the differences observed in means can occur in the same population, just not very often.

CONCLUSIONS

The *t*-test is a go-to statistic that can be used to test a difference between the mean of a sample with the mean of another sample or a standard mean. The *t*-test is easy to use and easy to use incorrectly. The test provides information on whether means are different from one another, provided that certain conditions are met, such as random sampling and a reasonable sample size. The test provides a statistical window to look through to see if the difference observed between two means is notably unusual, occurring 5% of the time or much less. A significant *t*-test result leads to inferences about what the research results might mean, encourages further investigation, and may be the beginning of a new research story.

REFERENCES

Ajman, F., & Bokharey, I. Z. (2015). Comparison of spiritual well-being and coping strategies of patients with generalized anxiety disorder and with minor general medical conditions. *Journal of Religion and Health*, *54*(2), 524–539. doi:10.1007/s10943-014-9834-2

Boland, P. J. (1984). A biographical glimpse of William Sealy Gosset. *The American Statistician*, *38*(3), 179–183. doi:10.1080/00031305.1984.10483195

Bookbinder, M., Blank, A. E., Arney, E., Wollner, D., Lesage, P., McHugh, M., … Portenoy, R. K. (2005). Improving end-of-life care: Development and pilot-test of a clinical pathway. *Journal of Pain and Symptom Management*, *29*(6), 529–543. doi:10.1016/j.jpainsymman.2004.05.011

Box, J. F. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical Science*, *2*(1), 45–52. doi:10.1214/ss/1177013437

Brown-Saltzman, K., Upadhya, D., Larner, L., & Wenger, N. S. (2010). An intervention to improve respiratory therapists' comfort with end-of-life care. *Respiratory Care*, *55*(7), 858–865.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Lawrence Erlbaum.

Flannelly, K. J., Jankowski, K. R. B., & Flannelly, L. T. (2015). Measures of variability in chaplaincy, health care, and related research. *Journal of Health Care Chaplaincy*, *21*(3), 122–130. doi:10.1080/08854726.2015.1054671

Flannelly, K. J., Jankowski, K. R. B., & Tannenbaum, H. P. (2011). Keys to knowledge: Searching and reviewing the literature relevant to chaplaincy. *Chaplaincy Today*, *27*(1), 10–15. doi:10.1080/10999183.2011.10767418

Flannelly, L. T., Flannelly, K. J., & Jankowski, K. R. B. (2014a). Fundamentals of measurement in health care research. *Journal of Health Care Chaplaincy*, *20*(*2*), 75–82. doi:10.1080/08854726.2014.906262

Flannelly, L. T., Flannelly, K. J., & Jankowski, K. R. B. (2014b). Independent, dependent, and other variables in healthcare and chaplaincy research. *Journal of Health Care Chaplaincy*, *20*(4), 161–170. doi:10.1080/08854726.2014.959374

Gordon, R. A. (2012). *Applied statistics for the social and health sciences*. New York, NY: Routledge.

Hart, H. E., Redekop, W. K., Bilo, H. J., Berg, M., & Meyboom-de Jong, B. (2005). Change in perceived health and functioning over time in patients with type I diabetes mellitus. *Quality of Life Research*, *14*(1), 1–10. doi:10.1007/s11136-004-0782-2

Jankowski, K. R. B., & Flannelly, K. J. (2015). Measures of central tendency in chaplaincy, healthcare, and related research. *Journal of Health Care Chaplaincy*, *21*(1), 39–49. doi:10.1080/08854726.2014.989799

Jankowski, K. R. B., Silton, N. R., Galek, K., & Montonye, M. G. (2010). Complementary alternative medicine practices used by religious professionals. *Journal of Health Care Chaplaincy*, *16*(3–4), 172–182. doi:10.1080/08854726.2010.498694

Jankowski, K. R. B., Vanderwerker, L. C., Murphy, K. M., Montonye, M., & Ross, A. M. (2008). Change in pastoral skills, emotional intelligence, self-reflection, and social desirability across a unit of CPE. *Journal of Health Care Chaplaincy*, *15*(2), 132–148. doi:10.1080/08854720903163304

Jewell, N. P. (2004). *Statistics for epidemiology*. New York, NY: Chapman & Hall/CRC.

Kamieniski, R., Brown, C. M., Mitchell, C., Perrin, K. M., & Dindial, K. (2000). Health benefits achieved through the Seventh-Day Adventist wellness challenge program. *Alternative Therapies in Health and Medicine*, *6*(6), 65–69.

Katz, D. L. (2001). *Clinical epidemiology & evidence based medicine*. Thousand Oaks, CA: Sage.

Kopacz, M. S., Adams, M. S., & Searle, R. F. (2017). Lectio Divina: A preliminary evaluation of a chaplaincy program. *Journal of Health Care Chaplaincy*, *23*(3), 87–97. doi:10.1080/08854726.2016.1253263

Pagano, R. P. (2012). *Understanding statistics in the behavioral sciences* (10th ed.). Belmont, CA: Wadsworth.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. doi:10.1177/1745691612463401

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi:10.1177/1745691612465253

Poncet, A., Courvoisier, D. S., Combescure, C., & Perneger, T. V. (2016). Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. *Methodology*, *12*(2), 61–71. doi:10.1027/1614-2241/a000110

Rummans, T. A., Clark, M. M., Sloan, J. A., Frost, M. H., Bostwick, J. M., Atherton, P. J., … Hanson, J. (2006). Impacting quality of life for patients with advanced cancer with a structured multidisciplinary intervention: A randomized controlled trial. *Journal of Clinical Oncology*, *24*(4), 635–642. doi:10.1200/jco.2006.06.209

Schiavolin, S., Broggi, M., Visintini, S., Schiariti, M., Leonardi, M., & Ferroli, P. (2015). Change in quality of life, disability, and well-being after decompressive surgery: Results from a longitudinal study. *International Journal of Rehabilitation Research*, *38*(4), 357–363. doi:10.1097/mrr.0000000000000136

Spidell, S., Wallace, A., Carmack, C. L., Nogueras-González, G. M., Parker, C. L., & Cantor, S. B. (2011). Grief in healthcare chaplains: An investigation of the

presence of disenfranchised grief. *Journal of Health Care Chaplaincy*, *17*(1–2), 75–86. doi:10.1080/08854726.2011.559859

Steele, M. M., Zahr, R. A., Kirshbom, P. M., Kopf, G. S., & Karimi, M. (2016). Quality of life for historic cavopulmonary shunt survivors. *World Journal for Pediatric and Congenital Heart Surgery*, *7*(5), 630–634. doi:10.1177/2150135116658009

Stewart, A. (2010). *Basic statistics and epidemiology: A practical guide* (3rd ed.). New York, NY: Radcliffe.

Stroup, D. F., & Teutsch, S. M. (Eds.) (1998). *Statistics in public health: Quantitative approaches to public health problems*. New York, NY: Oxford University Press.

Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25. doi:10.1093/biomet/6.1.1

Triola, M. F. (2004). *Elementary statistics* (2nd ed.). Boston, MA: Pearson.

von Gunten, C. F., Mullan, P., Nelesen, R. A., Soskins, M., Savoia, M., Buckholz, G., & Weissman, D. E. (2012). Development and evaluation of a palliative medicine curriculum for third-year medical students. *Journal of Palliative Medicine*, *15*(11), 1198–1217. doi:10.1089/jpm.2010.0502