Unadjusted Bivariate Two-Group Comparisons: When Simpler is Better

Thomas R. Vetter, MD, MPH,* and Edward J. Mascha, PhD†

Hypothesis testing involves posing both a null hypothesis and an alternative hypothesis. This basic statistical tutorial discusses the appropriate use, including their so-called assumptions, of the common unadjusted bivariate tests for hypothesis testing and thus comparing study sample data for a difference or association. The appropriate choice of a statistical test is predicated on the type of data being analyzed and compared. The unpaired or independent samples *t* test is used to test the null hypothesis that the 2 population means are equal, thereby accepting the alternative hypothesis that the 2 population means are not equal. The unpaired *t* test is intended for comparing dependent continuous (interval or ratio) data from 2 study groups. A common mistake is to apply several unpaired *t* tests when comparing data from 3 or more study groups. In this situation, an analysis of variance with post hoc (posttest) intragroup comparisons should instead be applied. Another common mistake is to apply a series of unpaired *t* tests when comparing sequentially collected data from 2 study groups. In this situation, a repeated-measures analysis of variance, with tests for group-by-time interaction, and post hoc comparisons, as appropriate, should instead be applied in analyzing data from sequential collection points. The paired *t* test is used to assess the difference in the means of 2 study groups when the sample observations have been obtained in pairs, often before and after an intervention in each study subject. The Pearson chi-square test is widely used to test the null hypothesis that 2 unpaired categorical variables, each with 2 or more nominal levels (values), are independent of each other. When the null hypothesis is rejected, 1 concludes that there is a probable association between the 2 unpaired categorical variables. When comparing 2 groups on an ordinal or nonnormally distributed continuous outcome variable, the 2-sample *t* test is usually not appropriate. The Wilcoxon-Mann-Whitney test is instead preferred. When making paired comparisons on data that are ordinal, or continuous but nonnormally distributed, the Wilcoxon signed-rank test can be used. In analyzing their data, researchers should consider the continued merits of these simple yet equally valid unadjusted bivariate statistical tests. However, the appropriate use of an unadjusted bivariate test still requires a solid understanding of its utility, assumptions (requirements), and limitations. This understanding will mitigate the risk of misleading findings, interpretations, and conclusions. (Anesth Analg 2018;126:338–42)

Just tea for two and two for tea,
Just me for you
And you for me alone.

Vincent Youmans and Irving Caeser,
from *Tea for Two* (1925)

Research in anesthesia, perioperative medicine, critical care, and pain medicine commonly relies on inferential statistics. Inferential statistics essentially allows investigators to make a valid inference about an association of interest for a specific population that is based on data collected from a random sample of that population. An unknown population parameter representing the clinical association or treatment effect of interest can thus be estimated from the study sample.[1]

From the *Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Austin, Texas; and †Departments of Quantitative Health Sciences and Outcomes Research, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication October 4, 2017.

Funding: None.

The authors declare no conflicts of interest.

Address correspondence to Thomas R. Vetter, MD, MPH, Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Health Discovery Bldg, Room 6.812, 1701 Trinity St, Austin, TX 78712. Address e-mail to thomas.vetter@austin.utexas.edu.

DOI: 10.1213/ANE.0000000000002636

As previously discussed in this series of statistical tutorials, hypothesis testing involves posing both a null hypothesis and an alternative hypothesis.[1] In this basic statistical tutorial, we discuss the appropriate use—including their so-called assumptions—of the following, common unadjusted bivariate tests for hypothesis testing[2] and thus comparing study sample data for a difference or association:

- Unpaired *t* test
- Paired *t* test
- Chi-square test for association
- Chi-square single sample goodness of fit test
- Wilcoxon-Mann-Whitney test
- Wilcoxon signed-rank test

We also highlight how the appropriate choice of a statistical test is predicated on the type of data being analyzed and compared. Hence, the reader is referred to a previous tutorial in this series on the types of data.[3] These unadjusted bivariate tests will serve as the basis for the next tutorial in this series on type I error (α), type II error (β), sample size, power analysis, and effect size.

UNPAIRED *t* TEST

The unpaired or independent samples *t* test is used to test and possibly to reject the null hypothesis that the 2 population means are equal (Ho: $\mu_1 = \mu_2$), thereby accepting the alternative hypothesis that the 2 population means are not equal (Ha: $\mu_1 \neq \mu_2$).[1,4–6]

Table. Common Bivariate Statistical Tests by Type of Observation (Independence), Number of Groups, and Type of Data

Type of Observations	Number of Groups	Type of Data: Continuous (Interval or Ratio)[a]	Type of Data: Categorical	Type of Data: Ordinal[b]
Independent	2	Unpaired *t* test	Chi-square	Wilcoxon-Mann-Whitney test
(Different groups)	>2	ANOVA	Chi-square	Kruskal-Wallis test
Dependent	2	Paired *t* test	McNemar test	Wilcoxon signed-rank test
(Same or paired subjects)	>2	GEE test	Cochran Q test	Jonckheere-Terpstra test

Abbreviations: ANOVA, analysis of variance; GEE, generalized estimating equation model adjusting for within-unit correlation.
[a]Drawn from a normally distributed population.
[b]Or continuous (interval or ratio) data that are not normally distributed.

Operationally, the 2-sample unpaired *t* test compares the random sample means from 2 independent groups ($\bar{x}_1$ and $\bar{x}_2$) to assess whether the underlying population means (μ_1 and μ_2) are significantly different (Table).[4]

The unpaired or independent samples *t* test is a parametric test.[7] The unpaired *t* test has several primary assumptions or requirements[4,8] that should be met:

1. Independent variable is categorical (ie, 2 study groups).
2. Dependent variable is continuous (ie, interval or ratio data).
3. Values are independent within each of the 2 samples.
4. Two samples are independent of each other.
5. There is approximately normal distribution of the dependent variable data for each group.
6. There are approximately equal (homogeneous) variances of the data across the 2 groups.

Levene test can be used to assess for the equality (homogeneity) of variances of data for a continuous variable calculated for 2 (or more) groups.[9] A violation of this *t* test assumption of equality (homogeneity) of variance is a greater concern with small sample sizes of <30.[4]

The unpaired *t* test is intended for comparing dependent continuous (interval or ratio) from 2 study groups. A common mistake is to apply several unpaired *t* tests when comparing data from 3 or more study groups (eg, A, B, and C): Group A versus group B, group B versus group C, and group A versus group C.[10] As will be discussed in a future tutorial in this series, in this situation, an analysis of variance (ANOVA) with post hoc (posttest) intragroup comparisons should instead be applied.[4,10,11]

Another common mistake is to apply a series of unpaired *t* tests when comparing sequentially collected data from 2 study groups (eg, blood pressure measurements at 5, 10, 15, 20, 25, and 30 minutes after anesthetic induction or pain intensity scores at 1, 4, 8, 12, 24, 48, and 72 hours after surgery). As will be discussed in a future tutorial in this series, in this situation, a repeated-measures ANOVA, with tests for group-by-time interaction and post hoc comparisons, as appropriate, should instead be applied in analyzing data from sequential collection points.[10]

Mayell et al[12] applied an unpaired *t* test in assessing the effect of a single preoperative dose of gabapentin on the postoperative mean opioid consumption and mean pain scores in adolescents undergoing idiopathic scoliosis surgery. However, Mayell et al[12] undertook a series of unpaired *t* tests across 7 sequential data collection points rather than an ANOVA with repeated measures. In their similar study of the postoperative effect of a single preoperative dose of gabapentin in pediatric idiopathic scoliosis surgery patients, Rusy et al[13] also applied a series of unpaired *t* tests across 5 sequential data collection time points for morphine consumption, but more appropriately used a repeated-measures ANOVA, with Bonferroni post hoc comparisons, for the sequentially obtained pain intensity scores.

PAIRED *t* TEST

As its name implies, the paired *t* test is used to assess the difference in the means of 2 study groups ($\bar{x}_1$ and $\bar{x}_2$) when the sample observations have been obtained in pairs.[5,14,15] The null hypothesis with a paired *t* test is that the expected difference in the population means is zero (Ho: $\mu_1 - \mu_2 = 0$), or equivalently, that the mean of the within-pair differences is zero (Ho: $\mu_D = 0$).[14]

Such a paired analysis is appropriate, for example: (1) when the dependent outcome variable is measured in each study subject before and after a treatment or other intervention; or (2) when study subjects are recruited as pairs who are matched for important demographic and/or clinical characteristics.[16]

The paired or matched *t* test is a parametric test.[7] The paired *t* test has a few primary assumptions or requirements[14,16] that should be met:

1. The paired values are randomly sampled from or are at least representative of the underlying population of paired samples.
2. Each pair of values is selected independently of the other pairs.
3. The differences between the paired or matched values are normally distributed (Gaussian distribution).

CHI-SQUARE TEST FOR ASSOCIATION

Chi-square for R × C Tables

Perhaps the most common statistical test is the Pearson chi-square test for association, named after famous statistician and philosopher Karl Pearson.[17–19]

The Pearson chi-square test is widely used to test the null hypothesis that 2 unpaired categorical variables, each with 2 or more nominal levels (values), are independent of each other (Table). If they are found to not be statistically independent, the null hypothesis is rejected, and 1 concludes that there is a probable association between the 2 unpaired categorical variables.[18–21]

The term "R × C tables" indicates that there may be multiple (>2) levels or categories for each of the 2 variables being compared and thus the cross-tabulation of the 2 variables can contain multiple number of rows (R) and columns (C).[18,22]

For example, in a clinical trial, investigators might test whether the intervention (eg, 2 anesthetic regimens) is associated with the incidence of postoperative nausea and vomiting (yes/no)—generating 2 rows and 2 columns, or investigators might assess whether race (eg, 4 categories) is associated with disposition immediately after hospital discharge (eg, home, skilled nursing facility, or inpatient rehabilitation unit)—generating 4 rows and 3 columns in the cross-tabulation.

The main requirements for the Pearson chi-square test[19,23,24] are as follows:

1. Two independent categorical variables, each measured on a set of subjects.
2. Each variable can have 2 or more categories.
3. Each variable should be nominal, such that there is no natural ordering of its levels or values.
4. There is a sufficient number of observations in each cell of the R × C table.

The categorical variables being compared would not be independent, for example, if the same variable was measured on the same patients before and after surgery. In such a setting, tests such as McNemar tests for paired proportions or a generalized estimating equation model to account for within-subject correlation would be used instead of Pearson chi-square test (Table).[20]

If the observed frequency is <5 in more than 20% of the cells, a Fisher exact test may be more appropriate than a Pearson chi-square test.[20–22] However, Fisher exact test is quite conservative and should only be used if truly needed.[23] In many cases, investigators can combine levels or categories of one or both of the variables of interest to remove the problem of very small cell sizes in an R × C table before conducting the chi-square test.

Abd-Elsayed et al[25] applied a Pearson chi-square test in their randomized controlled study that assessed patients' understanding of and consenting for clinical trials using 2 different consent form presentations. These authors concluded that consent forms presented in an enhanced format (ie, printed on fine paper and presented in a folio) did not improve patients' understanding or willingness to consent to participate in clinical trials.[25]

An article by McHugh[26] provides a more detailed yet user-friendly explanation of the Pearson chi-square test, including its usages, assumptions, and how to actually calculate the test. The so-inclined reader can use the data reported by Abd-Elsayed et al[25] for this latter exercise.

Ordinal Categories

The chi-square test assumes both variables have nominal levels or categories. If 1 or both of the variables of interest have >2 ordinal categories, the Pearson chi-square test is not appropriate because it ignores any natural ordering of the data, will be underpowered, and may also give invalid results. In such situations, the Mantel-Haenszel chi-square test for ordinal-by-ordinal data can instead be used.[23]

An example would be assessing the association between the number of years of postresidency clinical experience (<1, 1–3, 4–5, and >5) and the Cormack-Lehane airway classification system in which the grades range from 1 (full view of glottis) to 4 (neither glottis nor epiglottis seen).

In their study of the association between preoperative frailty and postoperative delirium after cardiac surgery, Brown et al[27] initially applied a chi-square test to compare the unadjusted incidence of delirium in frail versus nonfrail patients. In their sample of older patients, the overall observed incidence of postoperative delirium was significantly higher in the frail as compared with the nonfrail patients. These investigators then controlled for the potential confounding effect[28] of age, previous stroke, depression, and quintile of the Charlson comorbidity score on the relative risk of postoperative delirium. After this statistical adjustment, the risk of delirium remained significantly increased in the frail compared with the nonfrail patients.[27]

z Test for 2-Group Proportions

For a 2 × 2 table, an equivalent test for association is derived using a *z* test comparing the proportion having the outcome (usually, the "column" variable) between the 2 groups represented by the 2 rows.[29]

The null hypothesis for this test is Ho: $\pi_1 - \pi_2 = 0$, where π_1 and π_2 are the population proportions of patients having the outcome of interest in rows 1 and 2. The alternative hypothesis is Ha: $\pi_1 - \pi_2 \neq 0$. The disadvantage of this *z* test for 2 proportions is that it is only useful when there are exactly 2 groups to compare and exactly 2 levels or categories of the other variable.

An analogous 1-sample *z* test can be conducted if the question of interest is whether a certain proportion was equal to a hypothesized or historical constant, testing the null hypothesis that Ho: $\pi = c$, where c was the hypothesized constant.

CHI-SQUARE SINGLE SAMPLE GOODNESS OF FIT TEST

Sometimes researchers want to test the null hypothesis that a single variable of interest follows a hypothesized distribution or not. A chi-square goodness of fit test can be used in this setting.[20,23,24,30]

Similar to the Pearson chi-square test, the goodness of fit chi-square test compares the observed frequency in each category to the frequency that would be observed, conditional on the observed total sample size, if the variable did follow the hypothesized distribution. If the aggregate amount of discrepancies across the different categories is large, the null hypothesis will be rejected.[20,30]

In practical terms, goodness of fit chi-square test can also be used to compare the frequency observed in a sample with a predetermined value. In their study of arriving at a consensus regarding optimal care, Vetter et al[31] performed a 1-sample chi-square test for goodness of fit on the observed frequency of agreement among clinicians about the optimal management of high-risk coronary artery stent patients versus an ideal 95% agreement.

WILCOXON-MANN-WHITNEY TEST (MANN-WHITNEY *U* TEST, WILCOXON RANK SUM TEST)

When comparing 2 groups on an ordinal or nonnormally distributed continuous outcome variable, the 2-sample *t* test is usually not appropriate. The Wilcoxon-Mann-Whitney test is instead preferred.[20,32–35] This test is akin to a *t* test on the ranks of the data.[33] Data are ranked from smallest to largest while ignoring group assignment, then the groups are compared on the mean rank (accounting for ties).[34] If we refer to the outcomes in group 1 as *X* and the outcomes for group 2 as *Y*, the null hypothesis for the Mann-Whitney test is Ho: $P(X > Y) = .5$. In other words, the probability that a randomly sampled patient from group 1 has a higher value than a randomly sampled patient from group 2 is equal to .50, or a coin flip. The alternative hypothesis is Ha: $P(X > Y) \neq .5$. Assumptions or requirements of the Wilcoxon-Mann-Whitney test[35] are as follows:

1. Random samples are for 2 populations.
2. Outcomes are independent within samples.
3. Measurement scale is at least ordinal.
4. Variance of outcome is the same for each group.

The Wilcoxon-Mann-Whitney test assesses whether there is a location or pattern shift between the 2 populations of interest.[35] If the null is rejected, one concludes that values of one group tend to be higher than the other group. Exemplary uses of this test might be to compare randomized or propensity-matched groups on an ordinal outcome such a numerical rating scale pain score, or a skewed (nonnormally distributed) variable like opioid consumption. An extension of the Wilcoxon-Mann-Whitney test for comparing more than 2 groups is the Kruskal-Wallis test.[20,33,34] When reporting Wilcoxon-Mann-Whitney test results, it is important to avoid stating that a significant result implies that there is evidence that the medians differ or that they do not differ. The test does not explicitly compare medians, but rather a general location or pattern shift.[35] In fact, it is quite possible to have a valid statistically significant Wilcoxon-Mann-Whitney result when the observed medians are exactly the same. Results for the Wilcoxon-Mann-Whitney can be sufficiently reported as the median (interquartile range) for each group, along with the median difference between groups and its confidence interval, calculated using the Hodges-Lehman estimator.[36]

It is also appropriate to report the observed Mann-Whitney probability, $P(X > Y)$, simply called "*p*." It can be even more informative to report the odds for this probability, or $p/(1-p)$, which is called the Wilcoxon-Mann-Whitney odds. A confidence interval for the Wilcoxon-Mann-Whitney odds should also be reported. Further details, including sample size calculations and full examples, are provided in the excellent article by Divine et al.[37]

In their study of the effect of intraoperative dexmedetomidine on postoperative opioid consumption or pain scores after multilevel deformity correction spine surgery, Naik et al[38] analyzed their skewed (nonnormally distributed) continuous data using the Wilcoxon rank sum test (equivalent to the Wilcoxon-Mann-Whitney test and Mann-Whitney *U* test). The authors reported that compared to placebo, the median intraoperative opioid use was reduced in the dexmedetomidine group but not at 24 hours postoperatively. As compared to placebo, no difference in median pain scores, as measured by an 11-point discrete scale, was observed at 24 and at 48 hours postoperatively in the dexmedetomidine group.[38]

As mentioned above, while this was an appropriate use of the Wilcoxon-Mann-Whitney test, the correct interpretation would be that values of intraoperative opioid use were reduced (not that the median per se was reduced) in 1 group versus the other. Naik et al[38] could also have reported the median difference and its confidence interval, as well as the Wilcoxon-Mann-Whitney probability and odds.

WILCOXON SIGNED-RANK TEST

When making paired comparisons on data that are ordinal, or continuous but nonnormally distributed, the Wilcoxon signed-rank test can be used.[20,33,34,39] It is an alternative to the paired *t* test when the paired differences do not appear to be normally distributed.[34] Typical uses of this test would be comparing patients from before to after an intervention or procedure, such as surgery. Results could best be reported as the median difference and confidence interval for the difference. Technically, it is used to assess whether 2 dependent samples were selected from the populations having the same distribution.

Assumptions for the Wilcoxon signed-rank test[39] are as follows:

1. Data are paired and come from the same population.
2. Each pair is chosen randomly and independently.
3. The data are at least ordinal.

In their study of the influence of nociceptive stimulation on the analgesia nociception index (ANI) during propofol–remifentanil anesthesia, Gruenewald et al[40] applied a Wilcoxon signed-rank test. Specifically, the primary outcome variables of the ANI and the surgical pleth index were obtained in 25 patients before and after nociceptive stimulation. Each study subject thus provided paired data for the Wilcoxon signed-rank test. The ANI and surgical pleth index were significantly changed by the insertion of a laryngeal mask airway and the delivery of neuromuscular tetanic stimulation.[40]

CONCLUSIONS

In analyzing their data, researchers should consider the continued merits of the simple yet equally valid unadjusted bivariate statistical tests presented here. However, the appropriate use of any unadjusted bivariate test still requires a solid understanding of its utility, assumptions (requirements), and limitations. This understanding will mitigate the risk of misleading findings, interpretations, and conclusions. ■

DISCLOSURES

Name: Thomas R. Vetter, MD, MPH.
Contribution: This author helped write and revise the manuscript.
Name: Edward J. Mascha, PhD.
Contribution: This author helped write and revise the manuscript.
This manuscript was handled by: Jean-Francois Pittet, MD.

REFERENCES

1. Vetter TR, Mascha EJ. In the beginning-there is the introduction-and your study hypothesis. *Anesth Analg*. 2017;124:1709–1711.
2. Hazra A, Gogtay N. Biostatistics series module 2: overview of hypothesis testing. *Indian J Dermatol*. 2016;61:137–145.
3. Vetter TR. Fundamentals of research data and variables: the devil is in the details. *Anesth Analg*. 2017;125:1375–1380.
4. Salkind NJ. t(ea) for two: tests between the means of different groups. In: *Statistics for People Who (Think They) Hate Statistics*. 6th ed. Thousand Oaks, CA: Sage Publications; 2016:211–227.
5. Hazra A, Gogtay N. Biostatistics series module 3: comparing groups: numerical variables. *Indian J Dermatol*. 2016;61:251–260.
6. Kalpić D, Hlupić N, Lovrić M. Student's *t* tests. In: Lovric M (ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011:1559–1563.
7. Field A. Comparing two means. In: *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll*. Los Angeles, CA: Sage, 2013:357–391.
8. Motulsky H. Comparing two means: unpaired *t* test. In: *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. New York, NY: Oxford University Press, 2014:261–271.
9. Field A. The beast of bias. In: *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll*. Los Angeles, CA: Sage, 2013:163–212.
10. Motulsky H. Analysis of variance. In: *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. New York, NY: Oxford University Press, 2014:369–376.
11. Salkind NJ. Two groups too many: try analysis of variance. In: *Statistics for People Who (Think They) Hate Statistics*. 6th ed. Thousand Oaks, CA: Sage Publications, 2016:243–260.
12. Mayell A, Srinivasan I, Campbell F, Peliowski A. Analgesic effects of gabapentin after scoliosis surgery in children: a randomized controlled trial. *Paediatr Anaesth*. 2014;24:1239–1244.
13. Rusy LM, Hainsworth KR, Nelson TJ, et al. Gabapentin use in pediatric spinal fusion patients: a randomized, double-blind, controlled trial. *Anesth Analg*. 2010;110:1393–1398.
14. Hsu H, Lachenbruch PA. Paired *t* test. In: *Wiley Encyclopedia of Clinical Trials*. Hoboken, NJ: John Wiley & Sons, Inc, 2007:1–3.
15. Urdan TC. *T* tests. In: *Statistics in Plain English*. 4th ed. New York, NY: Routledge, Taylor & Francis Group, 2017:93–111.
16. Motulsky H. Comparing two paired groups. In: *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. New York, NY: Oxford University Press, 2014:272–283.
17. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag*. 1900;50:157–175.
18. Urdan TC. The chi-square test of independence. In: *Statistics in Plain English*. 4th ed. New York, NY: Routledge, Taylor & Francis Group, 2017:205–212.
19. Mendenhall W, Beaver RJ, Beaver BM. Analysis of categorical data. In: *Introduction to Probability and Statistics*. Boston, MA: CL-Wadsworth, 2013:574–605.
20. Salkind NJ. What to do when you're not normal: chi-square and some other nonparametric tests. In: *Statistics for People Who (Think They) Hate Statistics*. 6th ed. Thousand Oaks, CA: Sage Publications, 2016:315–331.
21. Field A. Categorical data. In: *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll*. Los Angeles, CA: Sage, 2013:720–759.
22. Howell DC. Chi-square test: analysis of contingency tables. In: Lovric M (ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011:250–252.
23. Hazra A, Gogtay N. Biostatistics series module 4: comparing groups - categorical variables. *Indian J Dermatol*. 2016;61:385–392.
24. Motulsky H. Comparing proportions. In: *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. New York, NY: Oxford University Press, 2014:233–241.
25. Abd-Elsayed AA, Sessler DI, Mendoza-Cuartas M, et al. A randomized controlled study to assess patients' understanding of and consenting for clinical trials using two different consent form presentations. *Minerva Anestesiol*. 2012;78:564–573.
26. McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb)*. 2013;23:143–149.
27. Brown CH IV, Max L, LaFlam A, et al. The association between preoperative frailty and postoperative delirium after cardiac surgery. *Anesth Analg*. 2016;123:430–435.
28. Vetter TR, Mascha EJ. Bias, confounding, and interaction: lions and tigers, and bears, oh my! *Anesth Analg*. 2017;125:1042–1048.
29. Department of Statistics Online Programs. Comparing two population proportions with independent samples. *STAT 500* 2017. Available at: https://onlinecourses.science.psu.edu/stat500/node/55. Accessed September 29, 2017.
30. Voinov V, Nikulin M. Chi-square goodness-of-fit tests: drawbacks and improvements. In: Lovric M (ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011:246–250.
31. Vetter TR, Boudreaux AM, Papapietro SE, Smith PW, Taylor BB, Porterfield JR Jr. The perioperative management of patients with coronary artery stents: surveying the clinical stakeholders and arriving at a consensus regarding optimal care. *Am J Surg*. 2012;204:453–461.e2.
32. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist*. 1947;18:50–60.
33. Mendenhall W, Beaver RJ, Beaver BM. Nonparametric tests. In: *Introduction to Probability and Statistics*. Boston, MA: CL-Wadsworth, 2013:606–654.
34. Field A. Non-parametric models. In: *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll*. Los Angeles, CA: Sage, 2013:213–261.
35. Neuhäuser M. Wilcoxon–Mann–Whitney test. In: Lovric M (ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011:1656–1658.
36. Hodges JL, Lehmann EL. Estimates of location based on rank tests. *Ann Math Statist*. 1963;34:598–611.
37. Divine G, Norton HJ, Hunt R, Dienemann J. Statistical grand rounds: a review of analysis and sample size calculation considerations for Wilcoxon tests. *Anesth Analg*. 2013;117:699–710.
38. Naik BI, Nemergut EC, Kazemi A, et al. The effect of dexmedetomidine on postoperative opioid consumption and pain after major spine surgery. *Anesth Analg*. 2016;122:1646–1653.
39. Rey D, Neuhäuser M. Wilcoxon-signed-rank test. In: Lovric M (ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, 2011:1658–1659.
40. Gruenewald M, Ilies C, Herz J, et al. Influence of nociceptive stimulation on analgesia nociception index (ANI) during propofol-remifentanil anaesthesia. *Br J Anaesth*. 2013;110:1024–1030.