Int J Physiol Pathophysiol Pharmacol 2017;9(5):157-163
www.ijppp.org /ISSN:1944-8171/IJPPP0066082

Review Article

Statistical methods and common problems in medical or biomedical science research

Fengxia Yan[1], Mayberry Robert[1], Yonggang Li[2]

[1]*Department of Community Health and Preventive Medicine, Morehouse School of Medicine, Atlanta, GA, USA;* [2]*ICF, Atlanta, GA, USA*

Received September 22, 2017; Accepted October 16, 2017; Epub November 1, 2017; Published November 15, 2017

Abstract: Statistical thinking is crucial for studies in medical and biomedical areas. There are several pitfalls of using statistics in these areas involving in experimental design, data collection, data analysis and data interpretation. This review paper describes basic statistical design problems in biomedical or medical studies and directs the basic scientists to better use of statistical thinking. The contents of this paper were based on previous literatures and our daily basic support work. It includes the sample size determination and sample allocation in experimental design stage, numerical and graphical data summarization, and statistical test methods as well as the related common errors at design and analytic stages. Literatures and our daily support works show that misunderstanding and misusing of statistical concept and statistical test methods are significant problems. These may include ignoring the sample size and data distribution, incorrect summarization measurement, wrong statistical test methods especially for repeated measures, ignoring the assumption for t-test or ANOVA test, failing to perform the adjustment for multiple comparison. This review intends to help the researchers in basic medical or biomedical areas to enhance statistical thinking and make fewer errors in study design and analysis of their studies.

Keywords: Statistical test methods, sample size determination, data summarization

Introduction

Statistical thinking is commonly used in public health, clinical research and community studies. In biomedical area, the studies usually involve animals or cell lines and the studies typically use the experiments which introduce more challenge for data analysis and statistical method utilization as compared to other areas of studies. Biomedical research, including study planning, experimental design, sample size and power determination, data collection, data analyses and interpretation as well as manuscript preparation, all require statistical support. While utilization of statistical thinking in biomedical and biological research are discussed by several researchers [1-5], there are still a lot of misunderstood and misinterpretation of statistical concepts. In this review paper, we will explain the related statistical issues in basic science research based on our daily statistical support activities. First, we will explain how to perform the experimental design including sample size determination and sample allocation. Then, the appropriate descriptive statistical summaries will be illustrated. Lastly, the most efficient statistical inference test methods will be described.

Experimental design stage

In the experimental design stage, the major neglects include two parts: 1) sample size determination and 2) allocation of the sample to different groups using randomization and defining the replication rate.

Sample size determination

Correctly defining the appropriate sample size is critical for clinical and community studies as well as the basic biomedical studies including animal study. In basic biomedical science, the researchers usually use the sample size that most similar studies use which would be a limitation for the study. Charan [6] proposed method to calculate the sample size for animal study based on power analysis which is very similar to

Table 1. Sample size calculation formulas [7]

	Formula	Note
Continuous outcome		
One sample	$n \geq \frac{\sigma^2}{\delta^2}(Z_\alpha + Z_\beta)^2$	δ is the detected difference σ is the population standard deviation
Two independent samples with common standard deviation	$n \geq \frac{4\sigma^2}{\delta^2}(Z_\alpha + Z_\beta)^2$	δ is the relevant difference in means σ is the population standard deviation
Two paired samples	$n \geq \frac{\sigma_d^2}{\delta_d^2}(Z_\alpha + Z_\beta)^2$	σ_d is the standard deviation of the mean difference δ_d is the mean difference
Categorical outcome		
One sample	$n \geq \frac{(Z_\alpha \sqrt{p_0(1 - p_0)} + Z_\beta \sqrt{p_1(1 - p_1)})^2}{(p_1 - p_0)^2}$	p_0 is the success rate under null hypothesis p_1 is the success rate under alternative hypothesis
Two independent groups	$n \geq \frac{p_C(1 - p_C) + p_E(1 - p_E)}{\delta^2}(Z_\alpha + Z_\beta)^2$	p_C is the success rate in control group p_E is the success rate in treatment group δ is $- p_E - p_C$
Two paired sample	$n \geq \frac{(Z_\alpha + Z_\beta)^2 f}{d^2}$	f is the proportion of discordant pairs d is the proportion difference

Where α is type I error level and β is the type II error level.

the clinical and community study sample size determination. If the research only has fixed number of subjects, then post hoc power analysis needs to be provided to show ability of the statistical tests. Five questions need to be considered in advance to determine the minimum required sample size to test a hypothesis.

1. What is the primary outcome? The primary outcome is the measurement of the main study purpose. In basic biomedical science, there may be more than one primary outcome. Sample size can be calculated for each outcome and the most conservative one can be applied as the study sample size. The caveat, however, is to make sure the sample size for the primary outcomes is well within this conservative estimate.

2. How to measure the primary outcome? Is it a continuous variable or a categorical variable? If the primary outcome is a continuous variable, what is the mean and variance in the general population and what is the expected difference between the control and intervention group? If the primary outcome is a categorical variable, what is the proportion in control group and what is the expected difference between the control group and intervention group?

3. What is the study design and what kind of statistical methods are used for the data analysis? In the basic biomedical science area, most of the studies are randomized-control designed studies. Hypothesis testing is involved most of the time.

4. What is the type I error and type II error levels. Type I error is the false positive rate which occurs when the researchers concluded that there was a significant effect when there is none. The most commonly used value for type I error is 5% which means only by chance the researchers may have 5% positive findings. Conversely, type II error is false negative rate which means the researchers fail to make positive conclusions when there are real positive findings. Type II error is used to define the study power which is the ability to detect positive findings when there is any. Appropriate defined sample size can minimize type I error and increase the power.

5. What is the expected attrition, e.g., death of animals in the study sample. A lot of basic biomedical science, especially animal study involved the death of the animals or cells, or the lost sample which require additional consideration to adjust the required sample size. For example, if power analysis show that the minimum required sample size is 20 and the study expects 10% attrition rate, the final sample size should be 20/0.9, which needs 23.

The sample size calculation formulas for different statistical tests are listed in **Table 1**.

While the power analysis method is the most powerful way to determine the sample size, it is not always possible to have the required information, such as standard deviation or effect size. A method called “resource equation method” [8] can be applied in this situation. Here is the formula based on the decided sample size.

E = Total number of animals - Total number of groups

Where E is a value to represent if the sample size is optimum. If E is less than 10 it means the sample size is not large enough. If E is greater than 20 it means the sample size is too large. For example, if the researcher has developed 5 groups to perform the intervention in an animal study, the total number of animals should be from 15 to 25. This crude method should only be used when sample size calculation cannot be done by power analysis method.

Sample allocation and replication

Randomization is the main principle for sample allocation. Randomization can greatly reduce the unintentional bias and confounding effects which may exist between the control group and the intervention group. Randomization ensures later use of probability theory to perform the statistical analysis [5]. It is important to make sure the control and treatment have same conditions in various aspects, such as the time of the day and temperature. For example, if 20 animals were required to perform the stroke model, these 20 animals should be randomized into two groups: 10 controls and 10 interventions. If the investigators only can do 4 animals per day, the morning and afternoon should also be considered for the randomization.

Data summarization

Summarizing data includes numerical summary and graphical summary. The purpose of data

Table 2. Description of numerical data summarization

	Measure of center	Measure of spread
Continuous numerical normal	Mean	Standard deviation
Continuous numerical skewed	Median	Interquartile range
Nominal categorical	No center and spread, use frequency with percentile	
Ordinal categorical	Median	Interquartile range

Table 3. Description of graphical data summarization

Variable type	Graph type
Nominal categorical	Bar graph, pie chart
Ordinal categorical	Bar graph, pie chart
Continuous numerical	Histogram, scatter plot, box-plot
Discrete numerical	Box-plot, stem-leaf plot

summarization is to describe the center of the data and how the data spread from the center. The data description in the first step can show some basic information for the readers including the overall sample size, sample size in each group as well as variable summaries. Appropriate measurements need to be chosen based on the properties of the variables. For continuous numerical variables, if the sample size is large enough and the data is normally distributed, mean would be the best way to represent the center of the data and standard deviation will be used to describe the variables dispersion. The most incorrectly used concept in basic science area is the standard error of the mean (SEM) [9] which equals sample standard deviation divided by square root of the sample size. The SEM measures the variation of different sample mean while the standard deviation measures the variation of the current sample. The purpose of using SEM is to construct confidence interval for the current sample mean or to perform statistical test. Marcel [10] described that all original journals published in 2012 in Cardiovascular research, Circulation, Heart Failure and Circulation Research had inappropriate use of SEM. In the same paper, the author stated that basic science studies had a 7.4-fold higher level of inappropriate use of SEM compare to clinical studies.

For continuous numerical variables, if the sample size is not large enough to determine the data distribution or the data distribution is not normal, the median with interquartile range (IQR) would be a better measure to describe the variable compare to the mean with standard deviation. For ordinal categorical variables, such as the Likert data or disease stage data, the median with IQR would be a good way to describe. Frequency with percentile will be used to describe the categorical variables. Another point of the data summarization related to the later statistical test method is, for example, if the statistical method the researcher chosen for the data analysis is nonparametric method, the median with IQR should be the correct way to summarize data instead of using mean with standard deviation. Different ways of summarization methods were listed in **Table 2**.

Graphics can provide informatively display of the variables. Bar charts or pie charts are usually used for categorical variables. Histogram can be used to describe continuous variables and show the distribution and potential outlier of the variables. Scatter plot with regression line often will be used for two continuous variables. Box-plot would be used to describe the ordinal categorical variables or non-normal distributed continuous variables. If the researcher chose to use median and IQR to describe data and use nonparametric statistical test later, the box-plot would be the correct figure to use. All these figures can show the distribution of the data and make intuitively comparison between groups or over time periods. The description of graphical summarization was listed in **Table 3**.

Statistical analysis tests

The study design, research hypothesis, the type of the variables as well as the data distribution defines the statistical analytical tests. The major tests method and related variable types were described in **Table 4**.

In basic medical or biomedical research, researchers often make mistakes in the following aspects: 1) two sample t-test and ANOVA, 2) repeated measurements, 3) non-parametric test, and 4) multiple comparisons.

Two sample t-test and ANOVA

In basic medical science area, continuous measurements are the most common outcomes,

Table 4. The basic statistical test methods for different type of variables

Outcome	Assumption	Statistical test
Numerical data		
One sample mean	Normal distribution	One sample t-test
One sample median	Not normally distributed	One sample median test or sign test/signed rank test
Two independent means	Normal distribution	Two sample t-test
Two independent means	Not normally distributed	Wilcoxon-Mann-Whitney test (Wilcoxon rank sum test)
Two correlated means	Normal distribution	Paired t-test
Two correlated means	Not normally distributed	Wilcoxon signed rank test
Independent more than two means	Normal distribution	ANOVA test
Independent more than two means	Not normally distributed	Kruskal Wallis test
Correlated (or repeated) more than two means	Normal distribution	Repeated measure ANOVA
Relationship between two numerical variables	Normal distribution	Pearson correlation test
Relationship between two numerical variables	Not normally distributed	Spearman correlation test
Categorical data		
One proportion test		Binomial test
Relationship between two categorical variables		Chi-square test
Relationship between two categorical variables, but one or more cells have expected value less than 5		Fisher's exact test
Test same categorical outcome on matched pairs		McNemar test
Binary outcome measured repeatedly		Repeated measure logistic regression

such as protein, DNA and RNA measurements. Two samples t-test is the way to compare two sample means and ANOVA would be the best way to compare more than two group means. There are two assumptions for t-test: the data should be normally distributed and identical independent. There are three assumptions for ANOVA: the data should be normally distributed with independent measurements and groups have equal variances. Several methods can be used to examine if the data is normally distributed. The graphic illustration is not good for small sample size. The most commonly used graphic methods are Q-Q plot and P-P plot. Statistical test is more accurate to test the normality which involves several different methods like K-S test, or S-M test. For basic science study, the sample size 5-6 is not enough to perform the normality test using any of these tests because small sample size is not large enough to provide enough power to do the test [7].

Repeated measurements

In basic biomedical science area, measurements of same subjects in repeated times are very common study design features. If the study design is the pre-post design and the outcome is continuous variable, then paired t-test should be applied instead of two independent sample t-test. If the outcome is categorical variable, the McNemar test would be the correct test to consider the inner correlation. If the time periods are more than two and the research would like to examine the outcome changes following different time periods, the correct statistical test should be one-way repeated measure ANOVA. Some investigators chose two-way ANOVA to consider the time effect which is not correct because the measurements in different time points are correlated which violated with the independent assumption of ANOVA test. If other risk factors may have an effect on the outcome, mix can be applied.

Non-parametric test

If the underlying distribution is not normal or the sample size is not large enough to determine the underlying distribution, then non-parametric test should be considered. Non-parametric tests do not depend on the underlying distribution which would be more robust compared to the parametric test. But, because the non-parametric tests ignore the original data value and perform the test based on the rank, some information will be lost which makes the non-parametric tests have less power compared to the related parametric test. The researchers need to choose a good balance point between parametric and non-parametric test. In basic biomedical research, small sample size is a common problem due to several factors including budget consideration. A small sample size makes it difficult to determine the underlying distribution for most parametric tests. Related non-parametric methods can be used. For example, if we would like to compare one specific protein level between control and treatment groups and we only have 5 sample for each group, non-parametric test called Wilcoxon-rank-sum test can be performed instead of two sample t-test.

Multiple comparisons

Multiple tests need to be performed if the researcher has multiple groups to compare or if the research has multiple outcomes to examine between the same group. For example, ANOVA test is to test group means for more than two groups comparison. If one is interested in more than two comparisons, there is a multiple comparisons problem which requires the multiple comparisons adjustment. For genetics study, multiple gene expressions between groups will be examined which also involves multiple tests. When multiple tests are conducted, the overall alpha level cannot be applied to all performed tests. There are a number of ways to perform the multiple comparisons adjustment. Bonferroni adjustment is to use proposed type I error level divided by the total number of tests which is the most conservative and easiest method. For example, if 5% type I error was chosen for the significance level and there are total of 5 tests, then 0.05/5=0.01 would be the new significance level. Multiple comparison adjustment can reduce the type I error which enlarges our test power. For example, 100 genes were tested between control and treatment group, there will be 5 positive expressed genes only due to the type I error if significance level of 5% was used.

There are several ways to perform the ANOVA post hoc test [11]. Tukey's HSD is good for all-pairwise comparisons while the Dunnett's procedure is appropriate for many-to-one comparisons because the Dunnett's procedure only

considers k-1 tests (k is the comparison group number) whereas Tukey's HSD assumes k*(k-1) tests which increases the false positive rate. Another method called Fisher's least significant difference procedure (LSD) you can choose for three groups pairwise comparison even though Meier stated that LSD has potential loss of power. Main pairwise comparisons for non-parametric Kruskal-Wallis test included Dunnett's procedure and Bonferroni adjustment. Dunn's z-test statistics approximates exact rank-sum test statistics by using the mean ranking of the outcome in each group from preceding Kruskal-Wallis test. Some researchers choose to use Mann-Whitney test for each two groups and perform Bonferroni adjustment which is acceptable whereas it may introduce more bias due to using different mean rank values from the Kruskal-Wallis test. False discovery rate (FDR) was widely used for genetics study with multiple comparison adjustment.

Conclusion

Statistical thinking is critical in basic biomedical research area including study planning, sample allocation, data description, data analysis and interpretation. Investigators from different areas of basic science already noticed the importance of applying correct statistical thinking into their research. This review may help the basic science researchers to understand some basic statistical concepts to avoid significant and routine errors in data summarization and statistical test methods for simple study designs. If the basic science researcher's study involves a complex study design, we encourage the researcher to consult with the statistician from the study conceptualization phase to better process the study.

Acknowledgements

The project described was supported by the National Institute on Minority Health and Health Disparities (NIMHD) and National Institute of Allergy and Infectious Diseases (NIAID) Grant Number U54MD007588, a component of the National Institutes of Health (NIH) and its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIMHD, NIAID or NIH, and The project described was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Fengxia Yan, Department of Community Health and Preventive Medicine, Morehouse School of Medicine, 720 Westview Dr. SW, Atlanta, GA 30310, USA. Tel: 404-752-1153; Fax: 404-752-1154; E-mail: fyan@msm.edu

References

[1] Binu VS, Mayya SS, Dhar M. Some basic aspects of statistical methods and sample size determination in health science research. Ayu 2014; 35: 119-23.
[2] Sprent P. Statistics in medical research. Swiss Med Wkly 2003; 133: 522-9.
[3] Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. Indian J Anaesth 2016; 60: 662-669.
[4] Bajwa S. Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. J Anaesthesiol Clin Pharmacol 2015; 31: 547-53.
[5] Sullivan LM, Weinberg J, Keaney JF Jr. Common Statistical Pitfalls in Basic Science Research. J Am Heart Assoc 2016; 5.
[6] Charan J, Kantharia ND. How to calculate sample size in animal studies? J Pharmacol Pharmacother 2013; 4: 303-6.
[7] Rosner B. Fundamentals of biostatistics, seventh edition. Boston: Brooks/Cole: Cengage Learning; 2011.
[8] Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 2002; 43: 244-58.
[9] Barde MP, Barde PJ. What to use to express the variability of data: Standard deviation or standard error of mean? Perspect Clin Res 2012; 3: 113-6.
[10] Wullschleger M, Aghlmandi S, Egger M, Zwahlen M. High incorrect use of the standard error of the mean (SEM) in Original Articles in Three Cardiovascular Journals Evaluated for 2012. PLoS One 2014; 9: e110364.
[11] McHugh ML. Multiple comparison analysis testing in ANOVA. Biochem Med (Zagreb) 2011; 21: 203-9.