

Statistics Primer for Lab Animal Researchers

July 2005

Instructor: Cara Olsen, MS MPH
colsen@usuhs.mil

1. Objectives

- a. Define common study designs
- b. Recognize and describe different types of data
- c. Choose and apply appropriate statistical tests based on type of data and study design

2. Study Designs

Most lab animal studies are randomized controlled trials (RCT's). These studies have three important aspects:

- *Randomization*: Animals are randomly assigned to treatment groups. Randomization is important because it increases the likelihood that there is no difference among treatment groups at the beginning of the study, and therefore that differences among the groups at the end of the study are the result of the treatment. Randomization does not ensure that treatment groups are exactly comparable in every study, only that they will be comparable on average. There is still a possibility that the groups will differ by chance alone, and randomization makes it possible to measure the likelihood of differences due to chance. This is addressed by the concept of statistical significance, which will be discussed below.
- *Control*: The study compares outcomes for animals receiving the treatment(s) of interest to outcomes for animals in a comparison group that is similar in all respects except the treatment. Typically the comparison group receives a placebo or the current standard of care. This is necessary because often animals will get better (or worse) on their own, and it is important to know how the treatment affects them beyond what would have happened in the absence of treatment.
- *Blinding*: The experimenter does not know which animals are receiving which treatment. This is important to avoid problems such as providing better care or applying different standards (even without being aware of it) to one group.

Some common designs for randomized controlled trials are

- *Parallel groups*: This common study design starts with a single group of animals. Each animal is randomly assigned to receive one and only one treatment.
- *Blocked design*: Sometimes animals are naturally organized into blocks, or groups that may differ from each other, such as litter, cage, or rack. It would not be appropriate to apply treatment A to the first rack, treatment B to the second rack, etc., because it would be impossible to determine whether any differences were due to the treatment or to different conditions in the different racks. In this situation, it is common to randomize the animals within each rack. So, if there are two treatments, half of the animals in the first rack would be assigned to treatment A, and half would be assigned to treatment B. The same randomization scheme would be carried out in the other racks.
- *Matched design*: Animals are paired based on characteristics such as sex, age, and genetics. For example, a study may use 10 sibling pairs of female newborn mice. One animal from each pair is randomly assigned to treatment A, and the other animal is assigned to treatment B.

- Paired design: Each treatment is applied to a different part of a single animal's body. For example, treatment A may be applied to the right eye and treatment B to the left eye of the same animal.
- Crossover design: Two or more treatments are applied sequentially to the same animal. Animals are randomly assigned to receive either treatment A or treatment B first, in case the order in which treatments are received affects the outcome. Each animal gets one treatment for a specified period of time, then after a recovery or "washout" period with no treatment, receives the second treatment.

3. Types of data

Statistical analysis depends on the type of data collected. There are many different ways to categorize types of data, and here is one approach:

- a. Categorical: Data obtained by assigning each observation to a group
 - i. Nominal: Data can be divided into two or more groups that have no natural rank order.
 - alive/dead
 - male/female
 - blood type
 - ii. Ordinal: Data can be divided into three or more groups that can be naturally ranked from low to high
 - tumor grade
 - better/same/worse
 - any rating scale
- b. Quantitative: Data measured on a scale with equal units
 - i. Count: Data that represent the number of items observed; can be a whole number greater than or equal to zero
 - number of affected cells
 - number of tumors
 - ii. Continuous non-normal: Data that can take on values other than positive whole numbers, but that are not normally distributed (see below)
 - ratios
 - percents
 - titers
 - iii. Continuous normally distributed: Data that follow a bell curve. This is important because many common statistical tests, including *t*-tests and analysis of variance, are based on the assumption that the data are normally distributed.
 - weight
 - length
 - volume

Quantitative data may be arbitrarily divided into categories (e.g. weight < 5g vs. weight ≥ 5g). This may be useful for describing the data, but results in a loss of precision when conducting statistical tests.

4. Describing data

A first step in any statistical analysis is to summarize and describe the data. Most researchers are familiar with how to calculate an average or mean, but sometimes a different summary is more appropriate. The best choice depends on the type of data being summarized.

- a. Categorical data: Report the number and percent in each category.

For example, "24 of 40 animals (60%) improved over the course of the study, 12 animals (30%) stayed the same, and the remaining 4 (10%) got worse."

- b. Quantitative data: Quantitative data should be described by both a measure of location that describes the center of the distribution, or what is a typical value; and by a measure of dispersion that describes the spread of the data, or how far most observations lie from the center of the distribution. Common measures of location include mean, median, and geometric mean. Common measures of dispersion include standard deviation, confidence interval and interquartile range (IQR). Formulas are given in section 7 below, and the best choice for a particular analysis depends on the type of data being summarized.
 - i. Count: Median and IQR
 - ii. Continuous non-normal: Median and IQR, or geometric mean and confidence interval. The geometric mean and confidence interval are often used for ratios, titers, and other measurements that are best viewed and analyzed on a log scale.
 - iii. Continuous normally distributed: Mean and standard deviation

5. Statistical tests

The choice of an appropriate statistical test depends on the research question, the study design, and the type of data. Three common research questions are

- How do two groups differ? (e.g. compare treatment with placebo)
- How do more than two groups differ? (e.g. compare several doses of a drug, or several different methods for treating the same condition)
- How well do two measurements agree? (e.g. do animals that receive a higher dose have better outcomes, or do two different ways of measuring depth of sedation yield similar results?)

For the purpose of this lecture, study designs will be divided into two broad categories:

- Parallel groups, including blocked designs: Each animal receives one and only one treatment. These studies yield *independent* data, in which knowing the outcome for one animal does not provide any information about the outcome for another animal.
- Paired, matched, or crossover designs: Each animal, or matched pair of animals, receives all of the treatments, and therefore can be compared with itself. These studies yield *dependent* data, in which two measurements taken on the same animal (or matched pair of animals) may similar to each other in ways unrelated to the treatment, meaning that knowing the first outcome for an animal (or matched pair of animals) provides information about the second outcome.

The tables below list the most common statistical tests for each research question, study design, and type of data discussed here. Selected formulas for the most common tests are in section 7 below; for other tests, see one of the references. Keep in mind that this is only a small subset of possible statistical tests. If none of these seems to fit your situation, or if you aren't sure which to choose, consult a statistician or a textbook.

a. Parallel groups, independent data

Distribution of Data	Comparing 2 groups	Comparing >2 groups
Categorical	Chi square or Z test (Fisher's exact test for small samples)	Chi square test
Count or continuous non-normal	Mann-Whitney U / Wilcoxon rank sum test	Kruskal-Wallis test
Continuous normal	Independent samples t-test	Analysis of variance (ANOVA)

b. Paired, matched or crossover trials; dependent data

Distribution of Data	Comparing 2 treatments	Comparing >2 treatments	Agreement or association between 2 measurements
Categorical	McNemar's Test	Cochran's Q	Kappa
Continuous normal	Paired t-test	Repeated measures analysis of variance (ANOVA)	Pearson product-moment correlation
Count or continuous non-normal	Wilcoxon signed ranks test	Friedman ANOVA	Spearman rank correlation

Example: Suppose you randomize 10 mice to receive a treatment and 10 mice to receive a placebo. You want to compare the average body weight in the two groups after 7 days. Weight is measured on a continuous scale, and upon inspection (by plotting a histogram), appears to be normally distributed. This study has the following characteristics:

- Parallel group design
- Research question involves comparing 2 groups
- Type of data is continuous, normally distributed

In this example, the appropriate test is the independent samples *t* test.

Interpreting tests: Most statistical software will provide a test statistic and a p-value. The test statistic can be interpreted with respect to a reference value from a table, but doesn't mean much by itself. The p-value indicates the probability, if there is truly no difference among the groups being compared, of observing by chance a difference among groups at least as large as the difference observed in this study. A small p-value, therefore, indicates that the observed result is unlikely to be due to random differences among the groups, and therefore is likely to be due to the treatment. Typically, p-values less than 0.05 are interpreted as indicating a statistically significant difference among groups.

6. Sample size calculations

Animal studies should be designed to use the minimum number of animals required to achieve the objectives of the study. The appropriate number of animals depends on the following factors:

- Effect size (e.g., difference in means between two groups)
- Variability of data (e.g., standard deviation)
- Desired significance level (probability of finding a significant result by chance when there is really no effect, usually 5%)
- Desired power (probability of finding a significant result when it actually exists, usually set at 80% or 90%).

Often, researchers do not know the expected effect size or variability of the data when they are planning a study. These can be estimated from previous research or from a small pilot study. In general, the larger the effect size and the smaller the variability of the data, the smaller the required sample size.

Sample size formulas are complicated, so consult a statistician, a textbook table or an online sample size calculator such as <http://calculators.stat.ucla.edu/powercalc/>. The two formulas below provide rough rules of thumb for estimating sample size in two common scenarios.

- Comparing two means: sample size per group =

- Comparing two proportions: sample size per group = $\frac{4 * \text{standard deviation} / (\text{group 1 mean} - \text{group 2 mean})^2}{16 * p * (1-p) / (p_1 - p_2)^2}$, where
 p_1 = proportion in group 1
 p_2 = proportion in group 2
 $p = (p_1 + p_2) / 2$.

7. Formulas and Definitions

Sample size: n = number of observations

Mean: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$

Median: middle value; half the values are greater than the median, and half are less than the median. For data ranked in increasing order $x_1 \dots x_n$, median (x) = $x_{(n+1)/2}$ if n is odd, and the average of $(x_{n/2}, x_{(n+1)/2})$ if n is even.

Geometric mean: $\text{antilog} \left(\frac{1}{n} \sum_{i=1}^n \log(x_i) \right)$

Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Standard deviation: $s = \sqrt{s^2}$

Standard error of the mean: $SEM = s / \sqrt{n}$

95% confidence interval for the mean: $\bar{x} \pm t^* SEM$, where t^* is the t -table value for $n-1$ degrees of freedom and probability 0.025 (for a two-sided interval). On average, if an experiment is repeated 100 times, and 100 means and confidence intervals are calculated, 95 of the calculated confidence intervals will contain the true mean.

For t -tests, find the critical value of t from a table. You will need to know the degrees of freedom, the significance level of the test, and whether the test is 1- or 2-tailed. Most researchers use the 5% significance level, indicating a 5% probability of observing a significant difference by chance when the two groups are actually the same. A 2-tailed test is appropriate if you want to show that the two groups are different, and a 1-tailed test may be used if you have a specific hypothesis about the direction of the difference. If the absolute value of the calculated value of t is greater than the critical value of t from the table, the difference between the groups is considered statistically significant.

Paired samples t -test of whether the average difference d between two quantities x and y is equal to zero:

let $d_i = x_i - y_i$, then $t = \frac{\bar{d}}{s_d / \sqrt{n}}$, with degrees of freedom = $n-1$

Independent samples t-test of whether the means of two independent samples are equal to each

other:
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with approximate degrees of freedom = smaller of n_1-1 and n_2-1

Chi square test:
$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$
,

where
$$\text{expected count} = \frac{\text{rowtotal} \times \text{columntotal}}{n}$$
,

Compare the calculated chi square statistic with the critical value from a chi square table with degrees of freedom = (number of rows - 1)*(number of columns - 1)

Pearson product-moment correlation: given two variables, x and y ,

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

r lies between -1 and $+1$. Positive values of r indicate a positive linear relationship between x and y ; negative values of r indicate a negative linear relationship between x and y , and $r = 0$ indicates no linear relationship between x and y .

Spearman rank correlation: given two variables, x and y , first rank the values of each variable in increasing order, then compute the Pearson product-moment correlation between the two sets of ranks. Because the Spearman rank correlation is based on the ranks of the data instead of the actual data, it requires only that the variables be measured on an ordinal scale.

8. References

- Armitage, P. (1971). Statistical Methods in Medical Research. New York, John Wiley and Sons.
- Hulley, S. B., Cummings, S.R. (1988). Designing Clinical Research: An Epidemiologic Approach. Baltimore, MD, Williams & Wilkins.
- Moore, D. S., McCabe, G. P. (1999). Introduction to the Practice of Statistics. New York, W. H. Freeman and Company.
- Swinscow, T. D. V. (1997). Statistics at Square One. BMJ Publishing Group. Accessed at <http://bmj.com/collections/statsbk/>, June 30, 2003.
- Van Belle, G. (2002). Statistical Rules of Thumb. New York, John Wiley and Sons.