Reporting Associations

Scott S. Emerson, M.D., Ph.D.

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

January 5, 2015

Abstract

Many statistical analyses are performed in order to detect associations between two variables. In this document I describe a general format for a formal report of such an analysis.

1 Associations Between Variables

Two random variables are said to be associated if the probability distributions for the two variables are not independent. Independence of probability distributions is defined mathematically. For instance, if we have two ordered variables such as age (Age) and systolic blood pressure (SBP) we would say that these random variables are independent based on the relationships between their *joint probability distribution* (on the left side of the following equation) and their *marginal probability distributions* (on the right side of the following equation):

$$Pr[Age < a \text{ and } SBP < s] = Pr[Age < a] \times Pr[SBP < s] \quad \text{for every choice of } a \text{ and } s.$$

Equivalent definitions can be based on conditional probability distributions which look at the distribution of one variable while holding the other constant. For instance, we might look within a group of people all having the same systolic blood pressure and consider the probability distribution for age. If the variables are independent, we would have the conditional distribution of age for every value of SBP equal to the marginal distribution for age. Alternatively, we could look within a group of people all having the same age and consider the probability distribution for systolic blood pressure. If the variables are independent, we would have the conditional distribution of systolic blood pressure for every value of age equal to the marginal distribution for systolic blood pressure. Hence, for independent random variables, both of the following equations must be true:

$$\begin{aligned} Pr[Age < a \,|\, SBP = s] &= Pr[Age < a] \quad \text{for every choice of } a \text{ and } s \\ Pr[SBP < s \,|\, Age = a] &= Pr[SBP < s] \quad \text{for every choice of } a \text{ and } s \end{aligned}$$

These equations also dictate that we can compare conditional distributions to each other,:

$$\begin{aligned} Pr[Age < a \,|\, SBP = s_1] &= Pr[Age < a \,|\, SBP = s_2] \quad \text{for every choice of } a,\ s_1, \text{ and } s_2 \\ Pr[SBP < s \,|\, Age < a_1] &= Pr[SBP < s \,|\, Age < a_2] \quad \text{for every choice of } a_1,\ a_2 \text{ and } s \end{aligned}$$

(Analogous definitions hold when our random variables are unordered, the notation for the probability distribution just always involves equality rather than inequality.)

Proving independence thus seems an impossible task:

- For each choice of a and s, we would have to have an infinite sample size in order to have the precision to be sure exact equality holds in the above equations, and
- (if that were not bad enough), we would have to perform such an analysis for all the possible ages (there are uncountably infinite different choices) and all the possible values of systolic blood pressure (again, an uncountably infinite number of different choices).

In any case, NIH will probably not fund the study. And even if they did, the publication could not come out in time to get tenure.

Most science therefore revolves around disproving independence, thereby establishing an association. If we focus on trying to establish associations, we have an easier task. For instance, if we can find <u>any</u> combination of a and s such that the above equations do not hold, we have established an association. And because independence demands that the conditional distributions be equal to each other as described above, it is sufficient to show that an association exists by showing that some summary measures of the distributions (e.g., mean, median) are different. This is the approach taken in most science.

2 Recipe for Investigating Associations

We can thus describe a "recipe" for finding an association between two variables:

1. Choose some variable Y as the "response" variable. We will estimate some summary measure of the distribution of Y within groups.
2. Choose some variable X as the "predictor of interest" (POI) variable. We will divide the population into subpopulations based on this variable. Each individual in a given subpopulation will have the same value of X as all other individuals in that subpopulation.
3. Choose some summary measure θ of the distribution of Y. This will have to depend on the type of variable (e.g., binary, unordered categorical, ordered categorical, continuous). Common choices for continuous random variables include
   (a) Mean
   (b) Geometric mean (providing Y is always positive)
   (c) Median (or, less commonly, some other quantile such as 25th or 75th percentile)
   (d) Probability that $Y > c$ for some specified, scientifically relevant value of c
   (e) Odds that $Y > c$ for some specified, scientifically relevant value of c
   (f) Hazard function (instantaneous rate of failure at a specified time, conditional on being at risk of failure at that time)
4. Estimate θ_x as the value of θ within the subpopulation having $X = x$. Do this for each value of x.
   (a) When sufficient data exists for each observed value of X, we might do this by estimating θ separately for each group (and not estimating θ for any group that we have no data on).
   (b) When our data is "sparse" (i.e., we have little or no data in some groups), we might do this by borrowing information across groups. For instance, regression models might presume that there is a linear relationship between θ_x and x.
      - In such cases we can interpolate to estimate θ_x for groups that between our minimum and maximum observed values of X, but for which we had no observations. This will be valid only if our method of borrowing information is exactly correct (e.g, a true linear relationship holds)

- More risky is to try to extrapolate outside the range of our data. This is almost always fraught with peril.
- (Most times, estimation of θ is not our true goal.)

5. Choose some "contrast" γ to compare θ across different populations. The most common choices are
   (a) differences: $\gamma(x_1, x_2) = \theta_{x_1} - \theta_{x_2}$, where $\gamma = 0$ corresponds to no association in θ
   (b) ratios: $\gamma(x_1, x_2) = \theta_{x_1}/\theta_{x_2}$, where $\gamma = 1$ corresponds to no association in θ
6. If there are more than two possible levels of x, choose some way to average across different comparisons (i.e., how we should combine $\gamma(x_1, x_2)$ and $\gamma(x_3, x_4)$ to come up with a "standardized" contrast

   - Most often we use a regression model to define some sort of weighted average across all of these comparisons.

Note that using this recipe, we will ultimately make one of two decisions

1. We can with high confidence state that the value of γ corresponds to their being an association between Y and X
   - We can of course be wrong, in which case we term this a "type I error"
2. We cannot with high confidence state that the value of γ correspond to their being an association between Y and X.
   - Reasons that we might come to this conclusion include
     (a) There is truly no association between Y and X, and the distribution of Y is the exact same no matter what the value of X.
     (b) There is truly an association between Y and X, but the value of θ_x is the exact same for every value of x.
     (c) There is truly an association between Y and X, and the value of θ_x is not always the same, but the method we chose for combining $\gamma(x_1, x_2)$ averages out to a value suggestive of no association.
     (d) There is truly an association between Y and X that corresponds to a standardize γ suggestive of an association, but we lacked sufficient precision to be sure of the value of that standardized γ.

3 Reporting Associations

When reporting associations, we will want to make sure that readers can judge the scientific importance, as well as the statistical evidence, or any association. Hence, we must make sure that our report makes clear

1. The variable whose distribution is being compared across groups.
2. The variable that is our primary definition of groups (our POI)
3. Any additional variables that were used to define groups of comparison (variables used to try to adjust for confounding or to provide additional precision)
4. The summary measure θ that is used for comparison (mean, geometric mean, ...)
5. The contrast γ that is used to compare θ across groups (difference, ratio, slope, ...)

6. Make clear the statistical methods used to make point estimates, confidence intervals, and p values (including whether one- or two-sided).

7. If there are only two groups, it is useful to provide descriptive statistics for those two groups. Include the sample size and the estimate of θ within each group.

8. Give a point estimate of the standardized γ across two "reference groups", making clear

   (a) the type of comparison of x (difference or ratio) that describes the reference groups (and generalizes to the way information was borrowed across groups)

   (b) units of measurement in scientific terms (e.g., males and females, not "group 1" and "group 2" unless those are the scientific terms)

   - For continuous predictor variables, it is generally a *very* good idea to make your reference groups reflect differences that are 1) scientifically relevant, and 2) commonly encountered.
   - For instance, when comparing ages of adults, you could consider a 1 year difference in age, because everyone understands such a difference. However, if you are considering a broad range of ages (e.g., 20 - 80 year olds), you might want to consider a 1 decade difference in age (i.e., 10 year difference in age) in order to capture a more scientifically relevant magnitude of the association. That is, if average blood pressure tends to differ by 0.5 mm Hg per year difference in age, that might at first seem an inconsequential amount. But for a 10 year difference in age, that corresponds to a 5 mm Hg difference in average blood pressure.
   - In any case, be especially careful in limiting significant digits when reporting the magnitude of the association. If you report the difference in mean blood pressure per 1 nanosecond difference in age, two significant digits may not be enough precision to judge the association per decade difference in age.
   - This latter point is especially true when reporting ratios such as odds ratios and hazard ratios. You need to report 2 to 3 significant digits *ignoring* the 1. If the odds ratio describing the association between cholesterol and prevalence of cardiovascular diseased is 1.0034 per 1mg/dL difference in serum cholesterol, you need to report it with at least that many significant digits, because we might be ultimately be interested in the OR associated with a 100 mg/dL difference in serum cholesterol: $1.0034^{100} = 1.40$. If all you reported was an OR of 1.003, that could result from an OR anywhere between 1.0025 and 1.0035. When that is exponentiated 100-fold, the range of OR is $1.0025^{100} = 1.284$ to $1.0035^{100} = 1.418$. It is of course even worse if you had only reported an OR= 1.0!!

   (c) the value of the estimate making clear the direction of any association (and see the comments above about significant digits)

   i. **A special note**; Our estimate of γ corresponds to a comparison of groups, either a difference or a ratio. Natural language (i.e., language spoken by humans) is quite imprecise about direction of comparisons, especially when talking about differences: Negative numbers were invented/discovered only about 2,000 years ago with natural language dating back substantially farther. Hence, when someone compares numbers of legs across animals, they might give the same answer of 2, no matter whether the question was "what is the difference in the number of legs for spiders and the number of legs for insects" or "what is the difference in the number of legs for insects and the number of legs for spiders". The same is less true for ratios, but we still need to be careful.

   ii. **Therefore**: When giving the value of the estimate for γ, you need to make certain that you define what a negative number or a positive number means in the real world. I usually try to avoid using negative numbers in text, instead using "lower" or "higher" (or similar wording) along with direct reference to the groups. So I might say "the males have a mean weight that is 26.9 pounds higher than the mean weight for women" or "the females have a mean weight that is 26.9 pounds lower than the mean weight for men".

iii. (If you are going to use a negative sign, make certain the reader knows what was subtracted from what. For instance, "The difference in mean weights (males minus females) was 26.9 pounds.")

9. Give an interval estimate that demonstrates the precision of inference. This is most often a 95% confidence interval. It is often useful to describe what a CI means.

   (a) As discussed above, negative numbers often cause confusion. Hence, when a confidence interval spans 0, you need to take extra pains to make certain that it is clear what a negative number means. I usually will describe a confidence interval using the words "lower" and "higher". So I might report, "Males were observed to have a mean SBP that was 0.424 mmHg lower than the mean SBP for females. The 95% confidence interval suggests that such an observation is not atypical of a setting in which the males' mean SBP might be anywhere from 3.27 mmHg lower to 2.43 mmHg higher than the mean SBP for females."

10. Give a p value. Give the exact value, because individuals may use different levels of significance or may be wanting to perform *ad hoc* adjustments for multiple comparisons. But do not as a rule report p values less than 0.0001, except to say $P < 0.0001$.

11. Optionally, include whether such an observation is "statistically significant". (There can be issues with multiple comparisons that might make me suppress this information.)