Interpretation of Statistics in Clinical Laboratory Studies

LISA W KAY, MARK A GEBERT

INDEX TERMS: data evaluation; statistics.

Clin Lab Sci 2006;19(1):18

Lisa W Kay PhD *is an Assistant Professor,* ***Mark A Gebert PhD*** *is an Assistant Professor, Eastern Kentucky University, Richmond KY.*

Address for correspondence: *Lisa W Kay PhD, Department of Mathematics and Statistics, 313 Wallace Building, Eastern Kentucky University, 521 Lancaster Avenue, Richmond KY 40475. (859) 622-1621, (859) 622-3051 (fax). Lisa.Kay@eku.edu*

Throughout the course of a research study that involves data, there are many opportunities for making errors. Poor survey construction, data collection methods that introduce bias, misuse of statistical analysis procedures, and misinterpretation of results can all lead to drawing incorrect conclusions. Researchers who are producing reports need to be able to collect and interpret data accurately, and clinical practitioners who are reading reports need to be able to assess the statistical content of such reports. While the technology of today makes it relatively easy to produce numerical summaries of data sets, the interpretation of the output still requires some careful thought. Correct interpretation of inferential procedures will be the focus of this article, and common misinterpretations will be described.

CONFIDENCE INTERVALS

One of the most commonly used inferential tools is the confidence interval. A confidence interval is used to estimate a parameter of interest and is of the form:
estimate ± margin of error.

An example of a fairly standard interpretation of a confidence interval might be the following: "We are 95% confident that the true mean age is between 25 and 35 years old." Here 95% tells us how confident we are in the *process* of constructing the confidence interval. If we were to randomly select samples of a particular size over and over again and construct a confidence interval for each one in this same manner, in the long run approximately 95% of the intervals would contain the true value that we are trying to estimate.

Example:

> All articles published in *Clinical Chemistry* over a six month period from January 2000 that contained data in the abstract were checked against the corresponding data reported in the article, including tables and figures. Data inconsistencies were classified as either data in the abstract being different from the data presented in the body of the article or the absence from the article of data presented in the abstract.
>
> Of 87 articles, 20 articles (23%; 95% confidence interval, 11%–35%) contained data in the abstract that were inconsistent with those reported in, or absent from the article.[1]

If we regard these 87 articles as a random sample of all articles in this journal containing data in the abstract, then the 95% confidence associated with the interval from 11% to 35% refers to the long-term 'chances' of many so-calculated intervals including the true proportion of all articles containing data inconsistencies with their abstracts.

We **cannot** make a probability statement about this particular interval that we have created. The ends of the interval are constants—they are fixed. Hence, either the population proportion (or other parameter of interest) is in the interval or it isn't. The level of confidence is based upon the confidence we have in the process that created the interval, not in the specific interval generated by a particular data set. Consider

The peer-reviewed Research and Reports Section seeks to publish reports of original research related to the clinical laboratory or one or more subspecialties, as well as information on important clinical laboratory-related topics such as technological, clinical, and experimental advances and innovations. Literature reviews are also included. Direct all inquiries to David G Fowler PhD CLS(NCA), Clin Lab Sci Research and Reports Editor, Dept of Clinical Laboratory Sciences, University of Mississippi Medical Center, 2500 North State St, Jackson MS 39216. (601) 984-6309, (601) 815-1717 (fax). dfowler@shrp.umsmed.edu

an analogy: we flip a 'fair' coin, catch it, and note which side is showing. Someone (who hasn't seen the coin) declares, "I am 50% confident that 'heads' is showing." This is a correct interpretation of confidence—on this particular toss, he is either 100% right or 100% wrong, but if the coin is 'fair' and if he continues expressing his confidence in the outcome 'heads', the long-term proportion of correct guesses will get very close to the 'true' value of 0.5 (50%).

Interpretation of *p*-values

Another commonly used procedure in statistical inference is the significance test or test of hypothesis. We generally state two hypotheses: a null hypothesis and an alternative hypothesis. The null hypothesis is the status-quo hypothesis or the hypothesis of no difference, while the alternative hypothesis is the research hypothesis. A small *p*-value in a test of hypothesis indicates that a difference has been detected. A *p*-value is the probability that we would get the statistic that we computed using our data, or something more 'extreme' that is supportive of the alternative hypothesis, if the null hypothesis were really true. Thus, a low *p*-value indicates that the data we have is not very believable if the null were really true. In statistics, we 'bet' on what is more likely; hence, if the data do not match the assumption of the null, then we bet that the alternative is more likely to be the truth. One common misinterpretation is to say that the *p*-value is the probability that the null is true, but this is not equivalent to the definition.

Example:

> Sodium values from capillary blood analyzed with the PCBA were higher ($p < 0.05$) than those from venous plasma analyzed with the CX5 as shown in Table 2 [of the original article]. Venous samples did not differ between methods. No outliers were identified, and the calculated TE for both comparisons were well within CLIA's EA of 4.0 mmol/L.[2]

The authors' statement means that if there were no difference between sodium as determined by the portable clinical blood analyzer (PCBA) and by the CX5 Chemistry Analyzer, and if many such samples were done and summary statistics calculated, the difference observed in their sample would happen in less than 5% of all samples. In simpler language, if there were no difference, what they actually saw wouldn't happen very often. As an analyst, we may conclude one of two things happened: choice one, there is no difference, and one of those 'rare things' happened; or two, there is a difference. Statisticians opt for concluding 'rare things don't happen very often', and declare the difference 'statistically significant'.

Note that while a smaller *p*-value is more significant, this only tells us that we are more certain that a difference (from the null value) exists. The size of the *p*-value is not indicative of the magnitude of the difference. In many cases, a confidence interval may be more informative than the corresponding test of hypothesis.

Even if the *p*-value is small enough to be considered significant, this does not necessarily mean that the difference really means anything. One thing to consider is that a large sample may allow you to detect a small difference. While the *p*-value may indicate that a difference exists, that difference may be so small that it does not really matter in practical terms. Again, a confidence interval may be more useful here in suggesting how large the difference might be.

Example:

> The only measured changes in the subject data during space flight (compared with before flight) were a slight (but statistically significant) increase in blood potassium as shown in Figure 5 [of the original article], and a slight (but statistically significant) decrease in ionized calcium. Glucose results were relatively more variable (compared with other data); however, fasting was not a constraint of this experiment. After flight, there were statistically (but not clinically) significant decreases in blood potassium, ionized calcium, and pH on R + 0 as shown in Figure 5 [of the original article] compared with preflight values.[3]

Note how the authors properly distinguish between statistical significance (differences too big to be due to 'chance', or sampling variability) and clinical significance (differences indicating an actual physiological change in the subject). Statistical significance reflects the likelihood the measure has changed while clinical significance reflects the magnitude of the measure's change, or whether or not a given drug affects the patient.

Any decisions based upon *p*-values that are 'borderline' or confidence intervals that indicate a small difference should be carefully considered using the researcher's best professional judgment. A *p*-value of .049 is really not that different from a *p*-value of .051. For such outcomes that are a little unclear, the researcher should take into account the cost of each choice and the consequences of making the wrong decision.

If a test has been performed in order to assess whether or not there is an association between two variables (the null

hypothesis is that there is no association) and the results are statistically significant, the researcher should be careful about the conclusions that are drawn. Here is a classic example: suppose that over a particular span of time in a particular city, the number of churches increases as liquor sales go up. Does this mean that the church members are buying alcohol? Of course not—the extraneous variable population size is causing both variables to increase. If an experiment has been performed, then one can make a case for causation. However, if the results come from an observational study, association does not necessarily imply causation, and other criteria need to be considered in order to establish a cause-and-effect relationship.

SOME COMMONLY USED STATISTICAL PROCEDURES

One-sample *t*-test

A one-sample *t*-test is used to compare a mean to a particular value. The null and alternative hypotheses are the following:

H_0: $\mu = \mu_0$ and

H_a: $\mu \neq \mu_0$ or H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$.

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

In "Segmented Neutrophil Size and Platelet Morphology in HIV/AIDS Patients" Bamberg and Johnson conclude that "Comparison of the sample mean of 15.1 to a reference mean for neutrophil diameter of 12.0 using a one-sample T-test demonstrated the HIV/AIDS patients' neutrophil diameter to be statistically larger than a reference population (T-test = 16.15, $p < 0.0001$)."[4] The small *p*-value indicates a significant difference but does not provide information about the magnitude of the difference. An accompanying confidence interval is often useful for such situations.

Paired *t*-test

A paired *t*-test can be used to make comparisons in data collected through matched pair designs, before-and-after observations made on the same subjects, etc. Here the mean difference is usually compared to zero. The null and alternative hypotheses are the following:

H_0: $\mu_d = 0$ and

H_a: $\mu_d \neq 0$ or H_a: $\mu_d > 0$ or H_a: $\mu_d < 0$.

The test statistic is

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

where and s_d are the mean and standard deviation of the sample differences, d_i.

In "An Investigation of Apoptosis in Androgenetic Alopecia", two scalp biopsies (frontal and occipital) were obtained from each of 16 men undergoing hair transplantation. Morgan and Rose conclude that "The paired *t*-test revealed a *t* statistic of 3.01 ($p < 0.05$) for TUNEL FSI vs OSI."[5] The paired *t* procedure is appropriate because of the natural pairing created by taking two measurements on each subject. The small *p*-value indicates that the mean difference in the frontal and occipital measurements is significantly different from zero.

One-way analysis of variance

The one-way analysis of variance (ANOVA) procedure is used to compare several means. The hypotheses are the following:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$ and

H_a: At least two means are unequal.

The test statistic is

$$F = \frac{\text{Between - groups estimate}}{\text{Within - groups estimate}} = \frac{BSS/(g-1)}{WSS/(N-g)}$$

where *BSS* is the *between sum of squares*, *WSS* is the *within sum of squares*, *g* is the number of groups, and *N* is the total sample size.

In "Comparison of Four Automated Hematology Analyzers", Koenn and others use one-way ANOVA several times to compare four automated hematology analyzers with respect to efficiency and sensitivity.[6] According to the authors, "Two of the DLC parameters displayed statistically significant difference ($p \leq 0.05$) from the manual differential." The authors go on to note that "Though statistically significant, the percent basophil variation would not be considered clinically

significant by most hematologists." In this example, the small *p*-value indicates statistical significance, but the authors' clinical expertise indicates that differences are not meaningful in a practical way.

Linear regression

When analysis includes examination of two quantitative variables for association, linear regression is the appropriate statistical method to use. The linear regression model is of the form

$$\hat{y} = a + bx$$

where *a* is the *y*-intercept and *b* is the slope. A statistical software package can easily produce the slope and intercept values as well as the correlation coefficient *r*.

Koenn and Ndah, in a method comparison study, used linear regression to quantify the comparability of results from a semi-automated PSA immunoassay and a manual uE_3 immunoassay to an automated analyzer.[7] Their results are excerpted here:

"PSA linear regression analysis for the two methods (ERA and Access 1) were

$y = 1.0008x + 0.0393$, $r = 0.9976$, SE = 0.1319, $n = 37$ …"

Examination of their accompanying graph will help us understand how these results led them to their conclusion that the automated analyzer method's output was acceptable (Figure1).

Examining their results, we see "$y = 1.0008x + 0.0393$"—this means, based on their experiment, the line that best describes the association between ERA PSA (regarded as "truth" about the immunoassay) and Access 1 PSA is the following:

Access IPSA = 1.0008 x ERA PSA + 0.393

Then to predict what value an analyst would get from the Access 1 method, we may take the value from the ERA method times 1.0008 and add 0.393.

Next, we see "$r = 0.9976$"—note that this value is positive, reflecting the fact that (as was expected) as ERA PSA increases, so does Access 1 PSA: a positive relationship. Also note that it is very near 1.0—if it had been 1, this would have indicated a *deterministic* relationship between the two PSA methods: no error in predicting one, using the other in a linear function. Often we see, rather than (or in addition to) *r*, its square, called r^2 = the *coefficient of determination*. This is interpreted as the proportion of the variability of the *y* variable that is accounted for in the linear association with the *x* variable. In this case, $r^2 = 0.9976^2 = 0.9952$, or 99.5%, a very high proportion of the variability. This reflects the fact that in the graph above, the points all fall very near the predicted line.

Finally, the statement "SE = 0.1319" describes the variability of the observed *y* (Access 1 PSA) values above and below the best-fit line: assumptions typically made when doing linear regression include that this variability (stated as a *s*tandard *e*rror) is uniform regardless of the *x* value (called *homoscedasticity*). In the picture, both the smallness and the uniformity of the variability in the *y*-values can be seen.

In a second example from their method comparison article, Koenn and Ndah compare manual uE_3 and automated Access 1 methods:

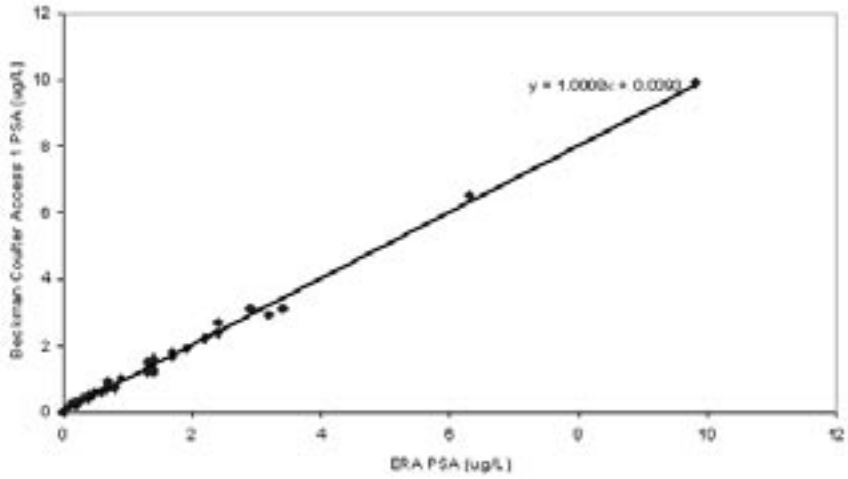
"uE_3 linear regression analysis for the two methods (RIA and Access 1) were

$y = 1.4105x - 0.3741$, $r = 0.8696$, SE = 0.8330, $n = 33$ …"

As we did in the first example, we include their graph to help us understand what these results mean. See Figure 2.

Their first result, "$y = 1.4105x - 0.3741$," indicates that to predict the Access 1 result from the RIA, we

Figure 1. Comparison of Beckman Coulter Access 1 PSA and ERA PSA by linear regression analysis



would multiply the RIA by 1.4105 and subtract 0.3741. If we regard RIA as 'truth', we may be concerned that we have to multiply by a number so different from 1 ($y = 1x + 0$ indicating the two measures are *identical*).

Next, we see "$r = 0.8696$"—a value that again is positive, however not as near 1 as in our first example. This is indicated in the graph by the higher variability of the y-values around the best-fit line. Consideration of $r^2 = 0.8696^2 = 0.7562$ reveals that far less of the variability of the Access 1 results are accounted for by the difference among the RIA values.

Finally, the "SE = .8330" reiterates the lower predictive value of the linear relationship in using RIA to predict Access 1 results: the bigger standard error impedes us in our goal of using Access 1 results in the stead of RIA.

Odds ratio

In their article analyzing incidence of thromboembolic events, Guirguis and others search for a relationship between these events and a significantly short aPTT (activated partial thromboplastin time).[8] Their analysis included the following: "Multivariate analysis revealed that patients with short aPTT have an odds ratio (OR) = 2.15, 95% confidence interval (CI) (1.27–3.64) ($p = 0.0042$), …" First, the odds ratio itself: note that the association being considered is that between two qualitative variables, namely whether a thromboembolic event occurred and whether the subject had a significantly short aPTT. If we are comparing the incidence of a particular characteristic in two groups of subjects, call p_1 the proportion of group one with that characteristic and p_2 the corresponding proportion in group two. Then

$$\text{Odds Ratio (OR)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Note that the numerator is the *odds for* a subject from group one *having* the characteristic of interest and the denominator, the same for group two. A property of the odds ratio, therefore, is that if this characteristic is equally prevalent in the two groups, the odds ratio will be 1. A greater prevalence in group one would lead to an odds ratio greater than 1. For our example study, the researchers found an odds ratio of 2.15, looking at incidence of thromboembolic events among those subjects with short aPTT compared to those without. They further quantify their stated association by pointing out that a 95% confidence interval for this OR is from 1.27 to 3.64—note that this interval does not include 1. They conclude their examination of this relationship by quoting a p-value of 0.0042, indicating that the sample result of 2.15, or greater, would happen in only 0.42% of experiments if the OR among all patients were actually 1. They conclude, statistically, that this is too rare to be due to chance, and decide that a short aPTT is a significant indicator of potential thromboembolic events.

Chi-square

If we are interested in examining the relationship between two categorical variables, a chi-square test is one possible approach. The hypotheses are the following:

H_0: The two classifications are independent.

H_a: The two classifications are dependent.

$$\chi^2 = \sum \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

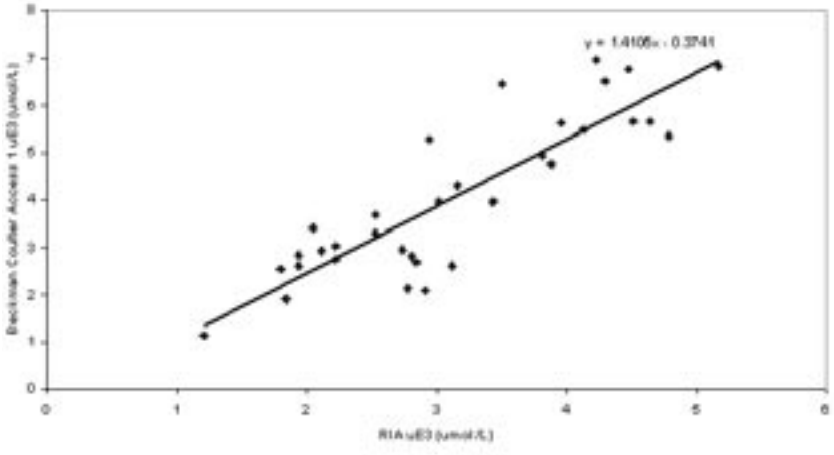
where

$$\hat{E}(n_{ij}) = \frac{(\text{row total})(\text{column total})}{n}$$

Note that this test should be used only when the expected cell counts are all at least five. When this is not the case, Fisher's exact test is a possible method of analysis.

Guiles and Tatum were considering the relationship among several qualitative variables in their 2002 study. They were curious about whether knowledge of certain skills, use of certain skills, and whether subjects were a clinical

Figure 2. Comparison of RIA uE3 and Beckman Coulter Access 1 uE3by linear regression analysis



laboratory science (CLS) or medical technology (MT) major in college were associated.[9] They reported results of 39 chi-square tests of association: like the odds ratio, the relationship among quantitative variables can be tested; but variables may have more than two levels. A chi-square test exhibiting significance indicates that the proportion of subjects at the various levels of one variable depends on (is affected by) the value of the other variable. For instance, in comparing problem-solving skills, 28 of the 34 laboratory professionals who were CLS/MT majors reported learning this skill compared to 7 of the 13 non-CLS/MT majors. These responses yield a chi-square test statistic of $\chi^2 = 4.019$ with a *p*-value = 0.045 (this, however, does not agree with their reported value of $\chi^2 = 8.672$; further, their reported *p*-value is not correct for this value of χ^2). As this value is less than 0.05, they declare that problem-solving skills are indicated by different proportions of CLS/MT and non-CLS/MT majors.

CONCLUSION

The technology of today provides us with the ability to quickly summarize and analyze data, but researchers and clinical practitioners are still responsible for interpreting computer output accurately. Thoughtful consideration of the results of confidence intervals and hypothesis tests can aid professionals in avoiding common pitfalls.

REFERENCES

1. Siebers R. Data inconsistencies in abstracts of articles in Clinical Chemistry. Clin Chem 2001;47:149.
2. Smith SM, Davis-Street JE, Fontenot T, and others. Assessment of a portable clinical blood analyzer during space flight. Clin Chem 1997;43(6):1056-65.
3. Ibid.
4. Bamberg R, Johnson J. Segmented neutrophil size and platelet morphology in HIV/AIDS patients. Clin Lab Sci 2002;15(1):18-22.
5. Morgan MB, Rose P. An investigation of apoptosis in androgenetic alopecia. Ann Clin Lab Sci 2003;33(1):107-12.
6. Koenn ME, Kirby BA, Cook LL, and others. Comparison of four automated hematology analyzers. Clin Lab Sci 2001;14(4):238-42.
7. Koenn ME, Ndah BV. Method comparison studies for prostate specific antigen and unconjugated estriol immunoassays. Clin Lab Sci 2003;16(2):94.
8. Guirguis NG, Eicher C, Hock L, and others. Thromboembolic risk factors in patients undergoing kidney transplant: implication of abnormally short activated partial thromboplastin time. Ann Clin Lab Sci 2003;33(4):396.
9. Guiles HJ, Tatum DS. The learning and application of generic skills by CLSs/MTs who have 'Left the Field'. Clin Lab Sci 2002;15(1):23.