

We are working on the problem of calculating the correlation between X and Y, where each Y (n=20) has ten X values associated with it. A simple answer would be to take the mean of the ten X values and use the 20 pairs to find the correlation. Is there a more sophisticated way to do this that would retain more of the information in the X values?

For those of you who want more information, Y = nurse satisfaction and X = patient perception of nurse care for each of ten patients cared for by a given nurse.

Beverly

---

You might consider replicating the Y values 10 times and associate with each of the ten X values.

---

You could try a repeated measures ANOVA.

---

Replicating the Y's 10 times for each set of 10 X observations and finding the correlation between the 200 values of X and the 200 values of Y should work. Not sure what the properties of the estimator are.

---

As I understand, you have 20 observations on each Y and corresponding 10 dimensional X vector. You could consider doing principal component analysis for X and then select the first principal component with maximum variance to represent X vector. So that you will have a single independent variable, representing patient's perception for each Y. You can then calculate the correlation. Another choice is, to calculate the multiple correlation coefficient between Y and the 10 dimensional vector X.

Both methods will involve lot of calculations and assumptions of normality of observations and nonsingularity of the variance co-variance matrix.

---

Are the ten patients the same for each nurse? I wouldn't expect this to be the case, but if so, you could simply run a multiple regression of Y on the X's (where  $X_i$  = patient i perception), and use the positive square root of the model  $R^2$ . This is the correlation of the observed Y's with the model predicted Y's (which itself is a weighted function of the X's).

---

---

I wish to add that if the sample is such that the nurses come from different departments in the hospital, you may subset the data and use weighted means (especially if there are unequal representations). Thus, you would need to know a nurse's department. I am thinking of a situation where the level of satisfaction depends on the nurses' area of work.

---

I might be not understanding the problem but wouldn't you need to define the parameter of interest under some conceptual model that generates such data, and then (and only then) find a good estimator?

---

...Think about what question you want to answer.

Correlating  $Y$  and  $\bar{X}$  gives you the correlation between the nurse and the **mean** patient satisfaction.

In this case you don't have to worry about the correlation structure of the patients.

I'm not sure what question is being answered by correlating the  $Y$  with the individual  $X$ 's, but because of the nested structure of the  $X$ 's, this has to be done carefully.

Replicating the  $Y$ 's 10 times will change the structure of the problem, unless you make some very strong assumptions that I expect will affect your conclusions.

You might be interested in estimating the mean and the variability of the patient responses among nurses. Which would make the correlation not a good answer to this question.

I hope this helps; I often find that refining the question helps me determine the appropriate method.

---

Are the  $Y$  variables an average overall satisfaction score from the 10 different patients who rated that specific nurse?

---

**Clarification requested:** The  $Y$  values are the satisfaction scores - each nurse completed one questionnaire about his/her overall satisfaction with the job, supervisor, etc. Each of the nurse's 10 patients filled out a questionnaire about the care he or she received from that nurse. It is of interest to see if the level of satisfaction felt by the nurse is correlated to the level of care received by the patient. The patient care scores are nested within nurse satisfaction scores.

---

---

Sounds more like the patient care scores are the Y and nurse satisfaction is the X.

I would do a repeated measures anova modeling mean patient care scores as a function of nurse satisfaction, adjusting standard error estimates based on the intra-nurse correlation among measurements.

---

Sounds like you need to do a regression in that case. One of the previous responses mentioned a repeated measures ANOVA, which I think would be essentially the same thing in this case.

In my opinion, you don't want to just replicate the Y's 10 times, because that would be artificially inflating the sample size of the data, or something along those lines.

---

We still need a little more information. Why are there 10 care scores per patient? Is that the same type of care measured 10 times per patient (e.g., on 10 separate days) or is it, for example, the care score in each of ten different areas of care--such as delivering medications, changing dressings, etc.?

---

I think you should just take the average of the X values within a nurse, then correlate X-bar with Y. Y is the nurse's overall satisfaction with her job, integrated (in her mind) over all patients she has seen recently. Y is not her satisfaction with a particular patient interaction. I would frame the question as how strong is the correlation between the nurse's job satisfaction and the average patient satisfaction with care, averaged over the population of all patients the nurse has cared for recently. You don't know the mean for that population, but you can estimate it from the sample of 10 patients from that population. There is of course some measurement error because X-bar is only an estimate of the true mean, but I would not worry about that - there is also measurement error in Y which could be bigger.

---

I think you could easily treat this as a multilevel model (nurses are level 2 units, patients are level 1 units). The outcome is a patient-level variable, so you would look at using nurse satisfaction as a Level 2 predictor of the random intercept.

---

If I understand the data set-up, multiple Y's and multiple X's, then consider a canonical correlation and/or a principal components analysis

---

---

Have you considered looking at the correlation between the satisfaction levels and a measure of the dispersion of the x variables? Maybe a regression with Y against the median X and the lower semideviation of each group of Xs would be informative. (In terms of other variables, is there a measure of how long each patient experienced care from the given nurse? How long the total hospital stay was? Some ranking of severity of condition? It sounds like an interesting marketing perception problem!)

---

If I understand the problem correctly, the data are ordinal categorical in nature. Talking about correlation coefficient and regular regression may not be the best approach. Subjective rating or scoring could be analyzed with discrete data analysis. With only 20 nurses, there may not be many distinct frequency counts in your categorical data model. Collapsing the score categories may be needed.

---

Bootstrap? Randomly select one patient observation from each of the 10 associated with a nurse's score - compute the correlation coefficient across the 20 nurses, store the value and repeat the resampling (with replacement) until a stable density plot is obtained - if skewed use the median, if normal use the mean to calculate the overall correlation between nurse satisfaction and patient satisfaction, you also get standard errors and percentiles as well this way. This can be done in R or a spreadsheet program that has a RAND function.

---

If I were the OP, I'd be interested in more than the effect of the central tendency. What if it is the lowest rating a given patient gives that affects a nurse's rating (i.e. you get one complaining patient and it ruins your whole day). Or maybe it's the highest one (at least you got something right that day) or maybe it's the range (consistency). The problem seems to me to beg for solution that allows for testing the different possible models.

---

I've found this discussion thread very stimulating and close to worthy of the topic's being a doctoral orals question in a research methodology program! However, it also reminds me of the gag, if you get 12 statisticians in a room to discuss something, you'll get 13 opinions!

---

You are being modest. More likely, out of 2 statisticians you will get 3 opinions.

---

Variance (noun): What any two statisticians are at

---

---

As this is the statistical consulting section I think we should concentrate on what constitutes good consulting when a problem is poorly posed. This is one of the most common experiences we have in consulting. First Beverly seems to have a real problem which is something to do with whether the nurses are satisfied with their performance compared to the judges (ten patients that they treated). But Beverly posed this in purely statistical terms as a problem of correlating the nurses satisfaction score with ten other scores. Should we summarize the ten response by an average and then correlate that with the nurses response. So some of us took it purely as a mathematical/statistical question and proposed answers. Some worried about ordinal vs continuous measurement. Others thought about changing their analysis to principal components or multiple regression. One person even suggested bootstrap. Well we all have our favorite ways of analyzing data and that seems to be entering into the discussion because the question is so open-ended to allow it.

A few of us saw this as a consulting exercise and did the proper thing. First let's not commit the type III error the cardinal sin of consulting (i.e. apply a perfectly brilliant solution to the wrong problem). I don't think that Beverly's initial message gave us a well posed problem. Whether it did or not it did not describe the application. Why are the nurses and patients being surveyed? Are we trying to see if there is agreement on individual nurses by the 10 raters. That would involve just a measure of level of agreement between the raters. That is a problem we know how to solve. It might be one of many problems that Beverly is trying to address. We know measure of interrater agreement and how to estimate them. But the nurses are also providing their own rating and it appears more likely that Beverly is more interested as to whether the nurses agree with the patients related to their performance. In that case maybe we want to make a pairwise comparison between the nurse and each of her ten patients. Some of us have asked some of these questions but we have not gotten all the answers. So we are not in the position to give good consulting advice yet. We need to hear the answers to all our questions and these answers will likely raise additional questions. My advice is to be a good consultant and make sure you know the "real" problem before you jump at the solution.

There is some joking going around now as we read the myriad of solutions often to different problems than what Beverly may have intended. In consulting there is no one "right" solution but there are good ones and there are "bad" ones. I am sure that everyone responding with solutions is a well trained statistician and all the solutions presented would be good solutions if they address the right question. But I don't think we have enough background as to what the real problems is. My guess is that most of us are repeatedly committing the type III error!

Michael Chernick  
Director of Biostatistical Services  
Lankenau Institute for Medical Research

---

---

I agree with Michael's insightful comments. Here are some additional reflections in this domain: I think that there is very often a question behind the problem that is more general and, in many cases, more interesting. An important job of the consulting statistician is to help translate from the more general to something that can be addressed with out methods. This challenge can be viewed as a poorly posed problem by an inadequately trained client or as an opportunity for an interesting collaboration. Many of us have seen examples of each. In the present setting, the more general question might be something like: do workers who have greater job satisfaction perform better? It gets refined to nurses in a particular setting, self-reported job satisfaction, self-reported patient perception of care provided by the nurse. The discussion has revealed that we, as statisticians, have something to add to process of formulating the specific question to be addressed by statistical methods. Most people see mu while we see sigma; it has been noted that the variability of the patient responses might be interesting and important. Given this idea, one go one step further to look at the proportion of failures, scores in the unacceptable range which might require some retraining or corrective action.

George McCabe  
Purdue University

---

When I started consulting in 72, I framed the process as moving from the presenting question to the underlying question. I my whole career I have had 3 times the presenting question was the underlying question. One time was a researcher who had come before for several projects. One was from a staffer for Senate Foreign Relations. The last was from a staffer from House Armed Services. I had been consulting on stat and methods for 18 years before my first encounter with the presenting question being the same as the underlying question.

This was less than 1 percent of the time.

Arthur Kendall  
Social Research Consultants

---

I think I'm going to disagree with Michael. I think that Beverly's question was fairly well-posed and rather clearly stated. Michael's concern seems to be about whether a correlation coefficient is the best thing to be chasing after for Beverly's practical problem...which, I'll grant, is a legitimate concern.

Eric

---

I think you miss the point Eric. It is not whether or not Beverly has a well posed statistical problem. I may think there is a little ambiguity but I won't quibble on that. Suppose she just wants to know that there is a statistically valid way of computing an estimate of bivariate

correlation when one nurse's score is to be compared to her ten patients. Well there may not be a standard answer in the literature. So some of us have been creative and come up with a variety of different potentially reasonable approaches.

The problem is that I think this is a real problem searching for a solution. I can accept that to average the ten responses to create a bivariate pair might be a legitimate solution to the problem that was posed. But an experienced consulting statistician would not believe that giving that answer is solving the underlying problem!

That is why so many excellent consultants came out of the woodwork when they saw my response. To them what we were doing for Beverly was a bunch of pointless exercises. We just heard one consultant say in his experience only 1% of the problems posed to him initially were the underlying question. So what has changed the direction of the discussion is that first and foremost the consulting statistician must sit down and talk with the client to discover the underlying question and then to see if there is a way to use his statistical expertise to solve that problem.

Before that the discussion was mostly silly and uninteresting. It may be interesting mathematically or theoretically but I would have no confidence that any of the suggestions posed no matter how well thought out really provides a good service to the client because I get the strong feeling that I haven't heard the whole story. Remember Beverly invited questions but only a few of us took her up on it. Maybe I haven't tracked the discourse well enough but I didn't hear answers that would lead me to say aha I understand your problem.

Michael Chernick

---

My only comment: A problem well-stated is a problem half solved ( maybe more than half like 2/e or whatever). I forgot how this is supposed to go but I think it is what a lot of these emails are saying.

---

I want to say that I was not criticizing Beverly at all for the way she posed the question. She specifically encouraged us to ask clarifying questions. My only critique was the way a number of us jumped into to provide answers without asking any questions when questions were warranted. I think this description gives better detail and background. I was more interested in the way we approached the problems consultants rather than how you should best address the client. I always get worried when a client says we analyzed it this way and didn't get a significant result. Is there a better way to analyze it? My worry is not that the appropriate answer to the question might be yes but rather that had the result been significant would the client have pursued a different form of analysis? Might this be a case where the client wants a significant result and is willing to look at the data upside down and backwards until one is found. I find that to be another one of the hazards of consulting. The research is so interested in a significant result that they test the data several times and in several different ways until they get a significant result. Unfortunately as we know without adjustment for multiplicity a significant raw p-value is likely to pop-up. It is part of the job of the consulting statistician to insist on multiplicity adjustment or at least provide fair warning of the danger of an erroneous conclusion.

Most of all inspite of the temptation to publish I wouldn't put my name on a paper where important consulting advice of mine was ignored.

Michael Chernick

---

" the important first step in consulting is to make the client express explicitly and unequivocally exactly what he want to find out; goals lead to design, goals plus design lead to data, and goals plus design plus data lead to analysis." This is very close to what I was saying.

I would change "to make" to say "to help". Turning tacit knowledge to explicit knowledge is not as easy as it seems. After all, a client has one discipline specific culture/dialect background, and the consultant is from another background usually one of the methodological/statistical cultures.

" the important first step in consulting is to help the clients express as explicitly and unequivocally as possible exactly what they want to find out; goals lead to design, goals plus design lead to data, and goals plus design plus data lead to analysis."

In my experience, it is important to have all of the key players involved in the discussion of the goals and that all of them understand the approach taken and what kinds of statements might be possible on the basis of the work. Very often the person who will make the decision sends a very junior member of the team to be an intermediary. Watch out for this. Also, it is often helpful to have the research team agree on drafts of "A Memo of Understanding" from the meetings with the consultant.

Arthur Kendall

---

What's worse is that statisticians and others who should know better are only interested in producing statistically significant results, no matter what the data say. Part of this is due to publication bias and part is due to incompetence. Perhaps we need a Journal of Insignificant Results for important findings of insignificance.

I hope that not too many statisticians fall into the trap you mention. I think most of us know better but some people do get intimidated in the work environment. I think ASA should protect them better and perhaps statistical ethics as defined and hopefully emphasized by the ASA could help against such intimidation.

Michael Chernick

---

I find this discussion group plays a role in reminding many of us that it is important to listen to the client and not assume anything about the problem at hand, and keep asking questions until we reach agreement about their objectives and their statistical needs, and the suitability of our proposed solution.