Select Reading Question Responses (3/19)

Which is a more powerful or useful statistic in understanding correlation: R or R^2?

Well, R^2 tells you about how linear the data is, while R tells you about how linear the data is *and* the sign of the slope of the linear relationship. By that measure, one could say that R is the more "powerful" statistic. That being said, usually it's pretty easy to just eyeball the sign of the slope from a scatterplot, so R^2 isolates the "important" or "more useful" part of the information: how linear the data is.

Is "$\Pr(>|t|)$" just the p-value? What exactly does the notation mean?

Can you elaborate on the relationship between the p-value and R^2?

Yes, $\Pr(>|t|)$ is just the p-value. In this notation, t refers to a certain test statistic that you calculate based on your sample. The notation $\Pr(>|t|)$ is then asking: assuming that the explanatory variable "does not explain" the response variable at all (ie, that the slope of the "true" best fit line is 0), what is the probability of seeing a sample whose test statistic would be more extreme than t (ie, more extreme than the test statistic that you actually calculated from your sample)?

R^2 tells you about the strength of the relationship between two variables. The p-value tells you about how likely it is to see a relationship that's more extreme than the one that you saw if the data truly comes from population in which the relationship is linear with slope 0.

There are situations where these two numbers could be very different. For example, suppose I collect a very big sample and find that all of the points on my scatterplot all *exactly* lie along a line of slope 1. Then I would expect a R^2 of 1 (the linear relationship is *very strong*), but a p-value of essentially 0 (the probability of seeing such a data if the explanatory variable was not doing any explaining would be infinitesimal).

In what ways is a prediction interval different from a confidence interval?

Let's say we're interested in the mean of some numerical variable, eg, the average height of American women. I might decide to study this question by collecting a simple random sample of 100 American women. There are then (at least) two intervals I can calculate using my sample.

- I can calculate a 90% confidence interval. The way to interpret this interval is: there's a 90% chance that the true population mean (ie, the average height of *all* American women, not just those in my sample) is contained in my 90% confidence interval.

- I can also calculate a 90% prediction interval. The way to interpret this interval is: if I randomly sampled one more American woman, there's a 90% chance that her height would be contained in my 90% prediction interval.

These interpretations are slightly different; correspondingly, the formulas used to construct prediction intervals are slightly different than those used for confidence intervals.