Explanation of R^2, and Other Stories

I accidentally gave an incorrect formula for R^2 in class. This summary was initially just going to be me correcting my error, but I've taken the opportunity to clarify every little thing I touched on. Now it functions as an outline (and expansion) of what I said in that lecture, including the corrected version of the R^2 formula; as a review of some material you've had in previous courses and may or may not remember well; and as a hopefully clear first statement of some of the concepts and terminology that will come up through the rest of the semester and in other courses (which I really wish I had had as a beginning grad student!).

In class I was describing R^2 as an example of an Effect Size measure, in contrast with p-values which don't tell you how much of an effect your treatment had. R^2 represents "the proportion of variance in the DV (Y) that is accounted for or explained by the IV (X) or set of IVs." Might as well point out right now: In the phrase "the proportion of variance accounted for or explained", the term "variance" is not being used in its technical sense, but rather, synonymously with "variability" or "variation" which would both be better terms, and in fact it's really the proportion of the Sum of Squares (see below) that's being accounted for or explained. But people say "proportion of variance" all the time in this context, so I will too.

My correction is simple (scroll way down to see it) but I want to review the explanation as well. Since the formula uses the abbreviation / symbol "SS" I will first remind you what SS means: the Sum of the Squared deviations from the mean. It's ubiquitous throughout all of statistics as, variously, an expression of random variation / lack of knowledge / departure from prediction / spread of numbers.

EXPLANATION OF SS

If I tell you Group 1 has a mean score of 8 on some dependent variable Y, and I tell you Subject S1 (not their real name) is in Group 1, what's your best prediction about Subject S1's score? It better be 8, cause you have no other basis for guessing anything else. Consider what you DON'T know about Subject S1 -- what other factors beside your treatment affected their score, how reliable that measurement is, how they differ from the other subjects in that group. That's why their score probably isn't 8 even though that is the best guess you can make. So when we measure the DEVIATION from the mean (i.e., the distance from S1's actual observed score to the mean, or Y-M), we're measuring what we CAN'T explain or account for: the measurement error, the individual differences -- sometimes summarized in other contexts with the letter e as a symbol for "error".

(BTW, you might be used to equations using X to represent scores instead of Y. Andrew Gelman has a blog post from 2007 (http://andrewgelman.com/2007/11/01/how_to_tell_the/) titled "How to tell the difference between a theoretical statistician and an applied statistician" that says in its entirety, "The theoretical statistician uses x, the applied statistician uses y (because we reserve x for predictors)." Now you know what kind we are, I guess.)

The average of those deviations would be informative about how spread out the data are around the mean. But what's the average, or mean, of the deviations from the mean? Zero, always. That's what the mean IS: the number that is closest to every other number. Add up all the positive and negative deviations from scores above and below the mean, they cancel out and you get zero; divide by the number of scores, still zero. Pointless. You could alternatively make sure the deviations come out to something positive by finding the mean of the absolute values of the deviations from the mean (the mean of all the $|Y-M|$) but as it turns out, that would be useless mathematically, so we don't bother with that.

Instead, guarantee that the average will be positive by finding each score's deviation from the mean and SQUARING it: $(Y-M)^2$. That's a SQUARED deviation from the mean, which you do for each score. Then add

them up and divide by the number of scores to get the average, or mean, squared deviation from the group mean. Adding them up gives you the Sum of the Squared deviations from the mean, or Sum of Squares, or SS, which as a formula is

$\Sigma(Y-M)^2$

Dividing by the number of scores gives you the average, or Mean, Squared deviation from the mean, or Mean Square, or MS, which as a formula is

$\Sigma(Y-M)^2/N$

-- more familiarly known as the variance (see next section for details about that formula though). The abbreviation / symbol MS for the variance is used in ANOVA and regression, so you'll see that plenty.

The mean is defined as the number producing the smallest SS to all the other scores -- the sum of the squared deviations from the scores to any number OTHER THAN the mean will be larger than the sum of the squared deviations from the scores to the mean. The criterion of producing the smallest possible SS is what's known as a Least Squares solution for a lot of statistics, not just the mean. For instance, when you do regression (in the Spring stats course) you try to come up with the equation predicting Y scores from X scores, which will be a line that's the best fit to all the data, like you're drawing a line through a cloud of (x,y) points on graph paper to capture the overall orientation of all the points. The "best fitting" line is the one that has the smallest sum of squared deviations from the actual data points to their predicted values on the line you draw. That ordinary kind of regression is called Ordinary Least Squares, or "OLS", regression.

VARIANCE AND STANDARD DEVIATION

Quick note about sample size in this context: For the moment I'm being casual about what N means, using it as the number of scores regardless of whether we're talking about the population or the sample. Sometimes people use lower case n for the sample size; the population size N could be in the hundreds, or millions, or in principle, infinite, but maybe the sample you took has n = 15. IN THIS COURSE, we'll soon use lower case n as the sample size of one group -- so if we have 3 groups of 5 each, n = 5 even though the total sample size is 15. But for now I'm staying imprecise about it and just using N.

A variance is a Mean Squared deviation from the group mean -- but that's only technically true if you're talking about the whole population of Treatment 1 scores that Group 1 was sampled from. And if we're talking about the population, let's be accurate about symbols. The sample mean is a statistic symbolized by M, and the population mean is a parameter symbolized by μ. So put the μ in the formula above and the population variance symbolized by σ^2 is:

$\sigma^2 = \Sigma(Y-\mu)^2/N$

which is literally the Mean Squared deviation from the mean -- you add up the squared deviations and divide by N -- with N being the number of scores in the entire population. Yeah right, like we're ever gonna be able to calculate that.

The SAMPLE variance formula is different -- it's symbolized by s^2 and uses the sample mean M not the population mean μ and has the (familiar?) different denominator of N-1 (the number of scores in the sample minus 1):

$s^2 = \Sigma(Y-M)^2/(N-1)$

That's because if we just put N in the denominator, the variance calculated from any individual sample would systematically underestimate the true population variance. And by systematically, I mean that the long run average value of that wrongly calculated sample variance -- its "Expected Value", as it's called -- would come out too small. That is, if you took thousands and thousands of different samples, the average of the thousands and thousands of sample variances calculated using the N denominator would come out to be less than the real population variance. This can be demonstrated through large computer

simulations (called "Monte Carlo studies") and also proven mathematically which is beyond most of us though it's on the web page just for kicks.

If we use the variance formula with a denominator of N-1, those thousands and thousands of sample variances would indeed average out to give exactly the true value of the population variance, which is what we'd want. We call that sample variance an "unbiased" statistic or estimate, and any statistic whose Expected Value (long run average) is actually the population parameter value is called unbiased. (The long run average or Expected Value of the variance or Mean Square in the population is usually abbreviated as Expected MS or EMS, which will come up later in this course; its value here is σ^2 but there will be other MSs to consider, you'll see.)

Notice if we're dividing the sample SS by N-1 instead of N, it's not literally the "Mean" Squared deviation from the mean anymore -- that would require dividing by N. But what we have is the unbiased sample estimate that is our best estimate of the population variance, because that's its long run average value. So I'm okay with referring to it as basically an average of the squared deviations, since that's conceptually true for the population if not literally for the sample.

The N-1 number is referred to as the degrees of freedom, or "df", of that variance. Degrees of freedom is also a ubiquitous concept and we'll say more about it later, in terms of it being "the number of observations that are free to vary" and "the number of parameters being estimated from the sample" (yikes!). For now just note that if using N in the denominator causes us to underestimate the population variance, it makes sense that using N-1 makes the denominator a little smaller, which makes our variance estimate a little bigger like it should be. It's kind of weird that that's the EXACT RIGHT AMOUNT of bigger it needs to be -- why not decrease the denominator to N-2? or N-3? Those would make the variance bigger too. But there's math that proves the df is the right number to use in the denominator.

The general form of a variance, then, is MS = SS/df, which is just what the formula $s^2 = \Sigma(Y-M)^2/(N-1)$ says less compactly.

Finally, the standard deviation, whether σ for the population or s for the sample, is just the square root of the corresponding variance. It's useful because if we say a sample of men's heights has a mean of 70 inches and a variance of 4 SQUARE INCHES, that's a little odd to interpret. But that variance of 4 means the standard deviation is 2 inches (not square inches), and it's more interpretable to say the typical difference from the mean is 2 inches. The standard deviation is nice for description but it's not as useful as the variance mathematically -- for instance, unlike the variance, the Expected Value of the sample standard deviation formula is NOT the population standard deviation. In case you're curious, the reason is related to the fact that the mean of the square roots of a set of numbers isn't the same as the square root of their mean. For example: 1, 4, 9, 16 have square roots 1, 2, 3, 4, whose mean is 2.5. But the mean of 1, 4, 9, 16 is 7.5, whose square root is 2.739. Likewise the long run average of thousands of samples' standard deviations, which are the square roots of their variances, isn't the same as the square root of the population variance, which is what the population standard deviation is.

SS AS A MEASURE OF WHAT IS NOT KNOWN

If a SS ($= \Sigma(Y-M)^2$) is suggesting how far the scores are from the mean, it's quantifying how much you DON'T KNOW about those scores even after you know their mean (measurement error, individual differences, general inadequate prediction). A large SS means there's a lot you don't know, and a small SS means there's just a little you don't know, considering your best guess at someone's score is their group's mean and the SS indicates how far or near most scores are to that mean. How can we quantify the change in our knowledge, that is, the amount of knowledge we gain by knowing that a subject is in Group 1, as

opposed to if we didn't know which group they're in? We can look at how much the SS decreases when we know group membership.

Look at these numbers as an example:
Group 1 has scores of 4, 6, and 8
Group 2 has scores of 5, 9, and 13
If we didn't know which group someone is in, our best guess as to their score would be the mean of all six scores, which is 7.5. The total SS would capture the variability of all the scores around that overall "grand" mean (not the "variance", which is a specific number, but the variability more generally, not meant as a technical term).
$SS_{total} = (4-7.5)^2 + (6-7.5)^2 + (8-7.5)^2 + (5-7.5)^2 + (9-7.5)^2 + (13-7.5)^2 = 53.5$
If we DO know which group someone is in, our best guess as to their score would be THEIR OWN GROUP'S mean (e.g., for Group 1 it would be 6), not the overall mean of 7.5. The Group 1 SS would tell us how much we DON'T know about the Group 1 scores beyond what their mean is:
$SS1 = (4-6)^2 + (6-6)^2 + (8-6)^2 = 8$
Likewise if they're in Group 2, we'd guess their score is the Group 2 mean (9), not the overall mean of 7.5. And the Group 2 SS would tell us how much we DON'T know about Group 2 scores beyond what their mean is:
$SS2 = (5-9)^2 + (9-9)^2 + (13-9)^2 = 32$
To quantify the total of what we still don't know, just add those together: in Group 1 there's a SS of 8, and in Group 2 there's a SS of 32 -- so we have a Sum Of Squared Deviations from each score's OWN GROUP MEAN that is 8 + 32 = 40. That is, that new SS is based on how far the Group 1 scores are from THEIR mean and how far the Group 2 scores are from THEIR mean. (When we do ANOVA, this combined SS from within each group will be known as the $SS_{within\ groups}$.)

We started out with a total SS of 53.5 representing our ignorance when we didn't even know group membership and had to guess 7.5 as the best guess for everyone's score. But knowing group membership, we can make better guesses based on which group someone's in, and our ignorance has been reduced to SS1 + SS2 = 40. How much has our ignorance been reduced by? 53.5 - 40 = 13.5, and as a proportion of the initial SS of 53.5, that would be 13.5 / 53.5 = .252, or 25.2% of the initially unknown variability being "explained" or "accounted for" by knowing group membership. THAT IS WHAT R^2 IS!
$R^2 = [SS_{total} - (SS1+ SS2)] / SS_{total}$
$R^2 = [53.5 - (8+ 32)] / 53.5 = 13.5 / 53.5 = .252$

THE CORRECTION TO MY IN-CLASS ERROR

I had a feeling I got the formula for R^2 wrong when I wrote
$R^2 = (SS1+ SS2) / SS_{total}$
And immediately after class I knew that was wrong because in fact that would actually be the proportion of variance that is NOT accounted for or explained.
That is, it would give you (8+ 32) / 53.5 = .748, which is actually "1 - R^2 ".

Correct your notes -- it should be
$R^2 = [SS_{total} - (SS1+ SS2)] / SS_{total}$
The numerator isn't SS1+SS2 -- it's how much smaller SS1+SS2 is than SS_{total} -- that is, SS_{total} - (SS1+SS2).
The denominator was right.

Some simple arithmetic would also show you that all I had to do to make my mistaken formula correct was change it to "1 - what I wrote", or
$R^2 = 1 - [(SS1+ SS2) / SS_{total}]$
which is what I mixed up with the incorrect version above.

WHY IT'S CALLED "R^2"

The symbol R^2 is the square of the "multiple correlation coefficient" R, which expresses how much a set of variables "goes together" just like the familiar (Pearson product-moment) correlation coefficient r does when there are only two variables. The upper case R^2 is general and could be used in place of r^2 even when there are just two variables, but r^2 is maybe more common in that case. The simple two-variable correlation coefficient itself is notated as r rather than R -- the upper case letter being reserved for correlation among three or more variables. Since the correlation coefficient ranges from -1 to +1, the value of R^2 ranges from 0 to +1, which is why it lends itself to interpretation as a proportion (or percentage when multiplied by 100). R^2 is known as the "coefficient of determination," though that term doesn't come up a lot.

It's not readily apparent WHY the square of the correlation coefficient should be equivalent to the SS formula for proportion of variance accounted for, but it's pretty easy to show that it's the case. We were trying to quantify how much we knew about DV scores based on knowing group membership. Let's rearrange the data to make that explicit: now one variable indicates group membership, and a second one indicates DV scores. A correlation between X and Y expresses how much we know about Y based on knowing X, so here it says how much we know about the DV based on knowing GP. It's the exact same information as listed for the groups above, but arranged as two variables that are paired columns of numbers:

GP	DV
1	4
1	6
1	8
2	5
2	9
2	13

With that data you could calculate the correlation coefficient r, and it comes out to be r = .502. You could test the significance of that value given the null hypothesis that the true population value is 0, i.e., that there's no relationship between group membership and DV score; and the p-value you get would be exactly the same as the p-value you'd get from doing a t-test on those two groups with the null hypothesis that the population means for those treatment conditions are equal. It's exactly the same question, logically and mathematically. And if you haven't noticed, if you square that r = .502, you get r^2 = .252, which is what the SS version of the R^2 formula above said.

The correlation formula is simple, so if you're interested, here's a preview / review with no other explanation for the time being. It uses the covariance (cov) and the Sum of Products of deviations from the mean (SP) instead of the Sum of Squared deviations from the mean:
$r = cov_{xy} / (s_x * s_y)$, where $cov_{xy} = SP_{xy} / (N-1)$, and $SP_{xy} = \Sigma(X-M_x)(Y-M_y)$
$SP_{xy} = (1-1.5)(4-7.5) + (1-1.5)(6-7.5) + (1-1.5)(8-7.5) + (2-1.5)(5-7.5) + (2-1.5)(9-7.5) + (2-1.5)(13-7.5) = 4.5$
$cov_{xy} = 4.5 / (6-1) = .9$
$r = .9 / (.548*3.271) = .502$

Incidentally I mentioned that you could use any two numbers at all for your GP variable (the group labels) to do the correlation. The calculation will come out exactly the same with the Groups labeled as 1 and 2, or as 0 and 1, or as 12 and 57, or even as -17 and 43.4. TRY IT, NERDS. And if you reverse the labels so the higher scoring group (Group 2) is labeled with the smaller number, e.g., the 1, or the 0, or the 12, or

the -17, the value of the correlation is still the same, but the sign is changed because now the higher the Group number is, the lower the DV score is: it'll be r = -.502. This works because with only two numbers, the only property they have is that they're different from each other; they're not really on a scale till you add in a third number, which is to say a third group. (The numbers 1 and 2 are simply different, but throwing in a 3 means the 3 is twice as far from 1 as the 2 is -- suddenly there's scale.) But if you do that, you can no longer have a single column representing the Groups, using 1, 2, 3 or any other numbers. In fact you'd need more than one column to represent group membership -- but then you're talking about "dummy coding" etc., which, for once, is something beyond the scope of this discussion.

Let's note that there are MANY equivalent expressions for R^2 which capture different uses of it and concepts underlying it, some of which were found in class, and which seemed superficially similar to what I wrote incorrectly. Equivalent forms include

$R^2 = SS_{regression} / SS_{total}$ -- for regression
$R^2 = SS_{between\ groups} / SS_{total}$ -- for ANOVA, usually referred to as η^2 "eta-squared"
$R^2 = t^2 / (t^2 + df)$ -- for t-tests, which is the least intuitive of all
$R^2 = r^2$ -- for correlation, which is trivial but I figured I'd mention it
There is also "partial R^2" which comes up in regression, and "partial η^2" which is SPSS's option for effect size in ANOVA. The meaning of "partial" will be addressed later.

And there are a bunch more. Why? Because it's a useful measure of EFFECT SIZE (as opposed to merely reporting p-values, which don't tell you how much of an effect your treatment had) and so it appears in almost every version of the General Linear Model. Even in techniques where it doesn't literally apply, you might see a version of it anyway, such as the "pseudo-R^2" in logistic regression (my personal favorite, FYI).

And in multivariate statistics, Wilks's lamdba (Λ) represents a related idea -- the proportion of variance NOT accounted for. Which means when I made my mistake, I was actually giving you a formula for Wilks's lambda. Wow, even my mistakes are smart.

BRIEF NOTE ON CONFIDENCE INTERVALS OTHER THAN 95%

Apart from R^2, another thing discussed in class was the confidence interval. I described confidence intervals with the example of the 95% CI. Applied to a sample, say you had a sample mean of 10 and a CI with confidence limits from lower bound 7 to upper bound 13, and say you could do a t-test to see if that 10 was significantly different from any hypothesized population value of the mean μ that you choose. We don't HAVE to test whether 10 is significantly different from 0; maybe we want to see if it's significantly different from 5, or 19, or whatever else we have a reason to be interested in. The CI tells you that if you test the sample mean of 10 against any hypothesized population value INSIDE the 7-13 range, you'll get a $p > .05$ (conventionally non-significant), and if you test it against any value OUTSIDE the 7-13 range, you'll get a $p < .05$ (conventionally significant). If you test that M = 10 against either $\mu = 7$ or 13 exactly, you'll get $p = .05$ exactly. You could also make a 99% CI, which would give you a wider interval with a .01 p-value cutoff in place of the .05 above, or a 90% CI which would give a narrower interval with a .10 p-value cutoff, or any other value you like -- though in practice, those are the ones people mostly use in real life, and of the three, it's the 95% CI that's overwhelmingly most popular.

For the population, the interpretation is that there's not a 95% probability of the true value of the population mean falling between 7 and 13, because the next sample might give a CI from 8 to 16, and the next might give one from 5 to 9, and they can't ALL have a 95% probability of containing the population mean. Instead we say when 100 of those confidence intervals have been calculated from 100 samples, we

expect 95 of them to contain the population mean. (That might not sound all that different, but it is). With a 99% CI, we expect 99 out of 100 calculated CIs to contain the population mean, and with a 90% CI we expect 90 out of 100 to contain it. If you think about it, it's obvious that greater confidence (e.g. 99%) goes along with wider (and therefore less precise) intervals, whereas we can get narrower (more precise) intervals if we're willing to settle for only 90% confidence. The confidence, again, is not in a single particular interval but in the technique for calculating the intervals.

Even though they use the same information as a t-test and p-value, CIs are still a bit more useful as a descriptive statistic since they tell you what the sample mean is (e.g., 10) and something about the precision of that estimate (a narrower interval from 7 to 13 is more precise than a wider interval from 2 to 18). These two pieces of information are referred to respectively as a point estimate and an interval estimate.