USING PEARSON CORRELATIONS* TO TEACH OR LEARN STATISTICS

©
Thomas R. Knapp
2012

Preface

This paper summarizes one person's approach (mine) to the teaching and the learning of some basic concepts in statistics. One of the most common research questions in science is: "What is the relationship between X and Y?", where X is an independent (predictor) variable and Y is a dependent (criterion) variable. So why spend a lot of time on means, standard deviations, ANOVAs, etc. when you can get right to the heart of the problem by plotting the (X,Y) data, calculating the corresponding Pearson r, and (for sample data) making whatever inference from sample to population is justified? [I know that means, standard deviations, ANOVAs, as well as Pearson r's, are all part of the so-called "general linear model" (I hate the word "model") and many formulas for it** directly involve means and standard deviations, but I claim that the direction and magnitude of a relationship outweigh any concerns regarding location or dispersion.]

I have chosen for illustrative purposes to use a set of data collected a few years ago consisting of the ages, heights, weights, and positions (pitcher, catcher, first base, etc.) of the players on each of the 30 Major League Baseball teams (14 in the American League and 16 in the National League). This choice is not only based on (bad pun) my personal interest in the sport but also on people's general interest in height and weight. The total number of observations is 1034, although there is one missing weight (for a pitcher on the Cincinnati Reds). The data are available free of charge on the internet, but I have prepared Minitab, SPSS, and Excel versions that I will be happy to send to anyone who might be interested. (Some of you might have already gotten a copy of the data.) My personal preference is Minitab, and all of the analyses in what follows have been carried out by using that computer program (more specifically a 1986 version, with which I am perfectly satisfied).

*The full name of a Pearson Correlation is "Pearson Product-Moment Correlation Coefficient", but product-moment is a concept in physics that most people don't understand, and it's really not a coefficient since it doesn't necessarily multiply anything.
** See the Appendix for some of those formulas.

Getting started

Suppose you were interested in the question: "What is the relationship between height and weight?" for Major League Baseball Players. If you have handy the data I will be referring to throughout this paper, there are several prior questions that you need to ask yourself before you do any plotting, calculating, or inferring. Here are ten of them. (The questions are addressed to the reader, whether you are a teacher or a student.)

1. Do you care about all of the players (the entire population of 1034 players) or just some of them? If all of them, you can plot and you can calculate, but there is no statistical inference to be made. Whatever you might choose to calculate are parameters for populations, not statistics for samples.

2. If all of them, do you care about what teams they're on, what positions they play, or their ages; or are you only interested in the "over-all" relationship between height and weight, regardless of team membership, position, or age?

3. If you do care about any of those matters, when it comes to plotting the data, how are you going to indicate same? For example, since there are 30 teams, will you use some sorts of symbols to indicate which of the data points correspond with which of the teams? How can you do that without cluttering things up? [Even if you don't care, trying to plot 1034 points (or 1033 points...remember that one weight is missing) so you can see what is going on is no easy task. Sure, Minitab or SPSS or Excel will do it for you (Minitab did it for me), but the picture might not be very pretty. More about that later.]

4. If just some of the players, which ones? The questions already raised are equally applicable to "sub-populations", where a sub-population consists of one of the 30 teams. (The number of observations for each of the teams varies from a low of 28 to a high of 38, with most around 35.) Sub-populations can and should be treated just like entire populations.

5. If you care about two of the sub-populations, say Team 10 and Team 27, how are you going to handle the "nesting" problem? (Player is nested within team.) Are you interested in the relationship between height and weight with the data for the two teams "pooled" together or in that relationship within each of the teams? This turns out to be one of the most important considerations in correlational research and is also the one that is often botched in the research literature. The "pooled" correlation can be vastly different from the correlation within each team. [Think about what would happen if you plot weight against height for a group of people that consists of half males and half females.] Therefore, even if you don't care about the within-team correlation, you should plot the data for each team separately and calculate the within-team correlations separately in order to determine whether or not the two sets of data are "poolable".

6. If you're interested in the entire population but rather than study it in full you want to take a sample from the population (the usual case) and generalize from sample to population (the usual objective), how do you do the sampling? Randomly? (Ideally.). What size sample, and why? With or without replacement? If with replacement, what will you do if you sample the same person more than once (unlikely, but possible)?

7. Suppose you choose to draw a simple random sample of size 30 without replacement. How will you do that? Use a table of random numbers found in the back of a statistics textbook? (The players are numbered from 1 to 1034, with the American League players listed first.) Use the random-sampling routine that is included in your favorite statistical "package" ? (All three of Minitab, SPSS, and Excel have them, and they are of varying degrees of user-friendliness.) Or use one of those that are available on the internet, free of charge? (There is a nice one at the random.org website.)

8. Do you expect the random sample of size 30 to consist of one player from each of the 30 teams? If so, you're dreaming. (It's possible but extremely unlikely.) Under what circumstances, if any, would you feel unsatisfied with the sample you draw and decide to draw a different one? (Please don't do that.)

9. Suppose you choose to draw a "stratified" random sample rather than a simple one. What variable (team, position, age) would you stratify on? Why? If age (a continuous variable carried out to two decimal places!), how would you do the stratification? How do you propose to "put the pieces (the sub-samples) back together again" for analysis purposes?

10. Whether you stratify or not, will you be comfortable with the routines that you'll use to plot the data, calculate the Pearson r, and carry out the inference from sample to population? Do you favor significance tests or confidence intervals? (Do you know how the two are related?)

Plotting

Let's consider, in turn, three of the examples alluded to above:

1. Entire population
2. Two sub-populations (Team 10 and Team 27)
3. Simple random sample from entire population

I asked Minitab to plot weight (Y) against height (X) for all 1034 players. Here it is: [N* = 1 indicates that one point is missing; the symbols in the heart of the plot indicate how many points are in the various regions of the X,Y space...the * indicates a single point and the + indicates more than 9]:

```
        -
     300+
        -                                          *
weight  -                             *      *
        -                             *
        -                   *            2   *            *      *
     250+                *     *   2  4  4   5  *  *   *     *
        -                   *  2   4  6  +   6  4  2   3
        -            *   *  3  6   +  +  +   8  7  *      *  *
        -         *  *   5  +  +   +  +  +   +  5  *   *
        -      *  *  3   6  +  +   +  +  +   8  6  4
     200+      *  *  6   +  +  +   +  +  +   +  3  *
        -         2  +   +  +  +   +  +  +   5  *  3
        -  *   *  8  +   +  +  +   +  7  *
        -  *   *  4  8   8  +  6   8  2  4   *
        -      2  2  4   3  3  4   2  *
     150+      *         *  2
        -
        -
          --------+---------+---------+---------+---------+-----height
               69.0      72.0      75.0      78.0      81.0

        N* = 1
```

Notice the very-close-to-elliptical shape of the plot. Does that suggest to you that the relationship is linear but not terribly strong? [It does to me.] That's nice, because Pearson r is a measure of the direction and the magnitude of linear relationship between two variables. The famous British statistician Karl Pearson invented it over 100 years ago. It can take on any values between -1 and +1, where -1 is indicative of a perfect inverse (negative) linear relationship and +1 is indicative of a perfect direct (positive) linear relationship. Notice also how difficult it would be to label each of the data points according to team membership. The plot is already jam-packed with indicators of how many points there are in the various regions.

Take a guess as to the value of the corresponding Pearson r. We'll come back to that in the "Calculating" section (below).

I then asked Minitab to make three weight-vs.-height plots for me, one for Team 10, one for Team 27, and one for the two teams pooled together. Here they are:

First, Team 10 (number of players = 33):

```
weight  -
        -                                                          *
        -
     225+
        -                               2      *
        -                  *      *                   *
        -
        -     *                   *     *      *
     200+                  *      5            *                   *
        -
        -     *     *                   2      2
        -                         *                   *
        -                         *            2
     175+                  *
        -                  *
        -
        -                         *
        -
     150+
           ------+---------+---------+---------+---------+------+height
              70.5      72.0      73.5      75.0      76.5      78.0
```

Second, Team 27 (number of players = 36):

```
weight  -
        -                            *                           *
        -
     245+                        *
        -                        *
        -                        *   *
        -                        *
        -                        2
     210+                        2   *   *
        -           2   *            *   *
        -           *   *
        -               2   *       *
        -               2   3
     175+               *
        -    *   *   *           *
        -    *
        -
        -        *
     140+
           --+---------+---------+---------+---------+---------+-height
          70.0      72.5      75.0      77.5      80.0      82.5
```

Third, both teams pooled together (number of players = 69):

```
weight  -
         -                           *                           *
         -
      245+                       *
         -                       *           *
         -                       *   *
         -                       *
         -           *   *   2   3   *
      210+                       2   *   *
         -   *       3   7   *   2   *   *   *
         -           *   *
         -   *   *       2   3   2   2
         -               4   3   2
      175+           *   *
         -   *   *   2           *
         -   *           *
         -
         -       *
      140+
           --+---------+---------+---------+---------+---------+-height
          70.0      72.5      75.0      77.5      80.0      82.5
```

There are several things to notice regarding those three plots. [I'll bet you've "eye-balled" some of the features already.] The first thing that caught my eye was the difference in the plots for the two teams taken separately. The Team 10 plot, although reasonably linear, is rather "fat". The Team 27 plot is very strange-looking, does not appear to be linear, and has that extreme "outlier" with height over 82.5 inches and weight over 245 pounds. (He's actually 83 inches tall and weighs 260 pounds...a very big guy.) The third plot, for the pooled data, looks more linear but is dominated by that outlier.

Once again, try to guess what the three correlations will be, before carrying out the actual calculations (see the following section). What do you think should be done with the outlier? Delete it? Why or why not? Are the data for the two teams poolable, and having combined them is OK? Why or why not?

Lastly, I asked Minitab to draw a simple random sample of 30 players from the entire population of 1034 players, and to plot their weights against their heights. [The Minitab command is simply "sample 30 C1 C2", where C1 is the column containing a list of all 1034 ID numbers and C2 is where I wanted to put the ID numbers of the 30 sampled players. How would you (did you) do it?]

Here's the plot:

```
         -                             *
 weight  -
         -                                    *
         -
      250+
         -
         -                                    *                   *
         -                      *             *     *
         -                      *                   *
      225+                             *
         -                      *      *                   *
         -         *      *     *             *
         -                *
         -         2            *
      200+  *      *      *            *
         -                      *
         -                             *                          *
         -
         -         2
           --------+---------+---------+---------+---------+-----height
                72.0      73.5      75.0      76.5      78.0
```

What do you think about that plot? Is it "linear enough"? (See the section on testing linearity and normality, below.) Guess what the Pearson r is. If you used your software to draw a simple random sample of the same size, does your plot look like mine?

Calculating

I asked Minitab to calculate all five Pearson r's for the above examples (one for the entire population; three for teams 10 and 27; and one for the random sample). Here are the results:

Entire population: r = .532. I'd call that a moderate, positive relationship.

Team 10: r = .280. Low, positive relationship?

Team 27: r = .723 (including the outlier). Strong, positive relationship, but partially attributable to the outlier. The correlation is .667 without the outlier.

Two teams combined: r = .587 (including the outlier). Moderate, positive, even with the outlier. The correlation is .510 without the outlier. Pooling was questionable, but turned out to be not as bad as I thought it would be.

Random sample: r = .439. Low-to-moderate, positive relationship. It's an under-estimate of the correlation for the entire population. It could just as easily have been an over-estimate. That's what chance is all about

Inferring

As I indicated earlier in this paper, for the entire-population example and for the two-team sub-population example, there is no statistical inference to be made. The correlation is what it is. But for the random-sample example you might want to carry out one or more statistical inferences from the sample data that you know and have, to the population data that you (in real life) would not know and wish you had. Let's see what the various possibilities are.

First, point estimation. If someone put a gun to your head and demanded that you give one number that is your best guess for the correlation between height and weight in the population of 1034 baseball players, what would you say? After plotting my random sample data and calculating the corresponding Pearson r for that sample, I'd say .439, i.e., the sample correlation itself. I would not be very comfortable in so doing, however, for three reasons: (1) my sample of 30 is pretty small; (2) I probably should make a "finite population correction", because the population is not of infinite size and the sample takes a bite (albeit small) out of that population; and (3) I happen to know (and you probably don't) from the mathematical statistics literature that the sample correlation is not necessarily "the best" single estimate of the population correlation. [It all has to do with unbiased estimation, maximum likelihood estimation, Bayesian inference, and other such esoteric matters, so let's not worry about it, since we can see that point estimation is risky business.]

Second, interval estimation. Rather than provide one number that is our best single guess, how about a range of numbers that might "capture" the population correlation? [We could always proclaim that the population correlation is between -1 and +1, but that's the entire range of numbers that any Pearson r can take on, so that would not be very informative.] It turns out that interval estimation, via so-called confidence intervals, is the usually preferred approach to statistical inference. Here's how it works.

You must first decide how confident you would like to be when you make your inference. This is always "researcher's choice", but is conventionally taken to be 95% or 99%, with the former percentage the more common. (The only way you can be 100% confident is to estimate that the correlation is between -1 and +1, but as already indicated that doesn't narrow things down at all.)

Next you must determine the amount of sampling error that is associated with a Pearson r when drawing a simple random sample of a certain size from a population. There are all sorts of formulas for the sampling error, but if you can assume that there is a normal bivariate (elliptical) distribution of X and Y in the population from which the sample has been drawn [more about this later], and you want your computer to do the work for you, all you need to do is tell your software program what your sample correlation and your sample size are and it will give you the confidence interval automatically. My favorite source is Richard

Lowry's VassarStats website. I gave his interval estimation routine 95% confidence, my .439 correlation, and my "n" of 30, and it returned the numbers .094 and .690. I can therefore be approximately 95% confident that the interval from .094 to .690 captures the population correlation. Since I have the full population data (but wouldn't in real life) I see that my interval does capture the population correlation (of .532). But it might not have. All I now know is that in the long run 95% of intervals constructed in this way will do so and 5% will not.

Third, there is hypothesis testing, which (alas) is the most common type of statistical inference. On the basis of theory and/or previous research I could hypothesize (guess) that the correlation between height and weight in the population is some number, call it ρ (the Greek rho), which is the population counterpart to the sample Roman letter r. Suppose I claim that there is absolutely no (zero) linear relationship between the heights and the weights of Major League baseball players, because I theorize that there should be just as many short and heavy players, and tall and light players, as there are short and light and tall and heavy ones. I would therefore like to test the hypothesis that ρ is equal to zero (the so-called "null" hypothesis). Or, suppose that Smith and Jones had carried out a study of the heights and the weights of adults in general and found that the Pearson correlation between height and weight was .650. I might claim that the relationship should be the same for baseball players as it is for adults in general, so I would like to test the hypothesis that ρ is equal to .650. Let's see how both tests would work:

Hypothesis 1: $\rho = 0$, given that $r = .439$ for $n =30$

The test depends upon the probability that I would get a sample r of .439 or more when the population $\rho = 0$. If the probability is high, I can't reject Hypothesis 1; if it's low (say, .05 or thereabouts...the usual conventional choice), I can. Just as in the interval estimation approach (see above) I need to know what the sampling error is in order to determine that probability. Once again there are all sorts of formulas and associated tables for calculating the sampling error and, subsequently, the desired probability, but fortunately we can rely on Richard Lowry and others who have done the necessary work for us. I gave Lowry's software my r and my n, and a probability (p) of .014 was returned. Since that probability is less than .05 I can reject the hypothesis that $\rho = 0$. (I might be wrong, however. If so, I'm said to have made what the statisticians call a Type I Error: rejecting a true hypothesis.)

[Important aside: There is an interesting connection between interval estimation and hypothesis testing that is especially relevant for Pearson r's. If you determine the 95% confidence interval for ρ and that interval does not include 0, then you are justified in rejecting the hypothesis that $\rho = 0$. For the example just completed, the 95% confidence interval around .439 was found to go from .094 to .690, and that interval does not include 0, so once again I can reject Hypothesis 1, indirectly. (I can also reject any other values that are not in that interval.) The .439 is said to be "significant at the 5% level" (5% is the complement of 95% and I got a "p-value" of .05.]

Hypothesis 2: ρ = ..650, given that r = .439 for n =30

The logic here is similar. If the probability is high of getting a sample r of .439 (or anything more discrepant) when the population ρ = .650, I can't reject Hypothesis 2; if the probability is low, I can. But the mathematics gets a little heavier here, so I will appeal to the preceding "Important aside" section and claim that I cannot reject Hypothesis 2 because .650 is within my 95% confidence interval. (I might be wrong again, for the opposite reason, and would be said to have made a Type II Error: Not rejecting a false hypothesis.)

Since I know that ρ is .532...but wouldn't know that in real life (I can't stress that too often), I "should have" rejected both Hypothesis 1 and Hypothesis 2.

Rank correlations and non-parametric inference

In the Inferring section (see above) I said that certain things (e.g., the use of Richard Lowry's routine for determining confidence intervals) follow if you can assume that the bivariate distribution of X and Y in the population is normal. What if you know it isn't or you are unwilling to assume that it is? In that case you can rank-order the X's, rank-order the corresponding Y's, and get the correlation between the ranks rather than the actual measures. There are several kinds of rank correlations, but the most common one is the Spearman rank correlation, call it r_S (pronounced "r-sub-S"), where the S stands for Charles Spearman, the British psychologist who derived it. It turns out that Spearman's r_S is identical to Pearson's r for the ranked data.

Consider as an example the height (X) and weight (Y) data for Team 27. Here are the actual heights, the actual weights, the ranks for the heights, and the ranks for the weights (if two or more people have the same height or the same weight, they are assigned the mean of the ranks for which they are tied):

```
   ID    X      Y    Xrank  Yrank

    1    73    196   12.0   17.0
    2    73    180   12.0    9.0
    3    76    230   31.0   31.5
    4    75    224   24.0   30.0
    5    70    160    1.5    2.0
    6    73    178   12.0    7.0
    7    72    205    6.5   22.5
    8    73    185   12.0   11.5
    9    75    210   24.0   26.0
   10    74    180   17.5    9.0
   11    73    190   12.0   15.0
   12    73    200   12.0   19.5
   13    76    257   31.0   35.0
   14    73    190   12.0   15.0
   15    75    220   24.0   29.0
   16    70    165    1.5    3.0
   17    77    205   34.5   22.5
   18    72    200    6.5   19.5
   19    77    208   34.5   24.0
   20    74    185   17.5   11.5
   21    75    215   24.0   28.0
   22    75    170   24.0    5.0
   23    75    235   24.0   33.0
   24    75    210   24.0   26.0
   25    72    170    6.5    5.0
   26    74    180   17.5    9.0
   27    71    170    3.5    5.0
   28    76    190   31.0   15.0
   29    71    150    3.5    1.0
   30    75    230   24.0   31.5
   31    76    203   31.0   21.0
   32    83    260   36.0   36.0 [the "outlier"]
   33    75    246   24.0   34.0
   34    74    186   17.5   13.0
   35    76    210   31.0   26.0
   36    72    198    6.5   18.0
```

As indicated above, the Pearson r for the actual heights and weights is .723. The Pearson r for the ranked heights and weights (the Spearman r_S) is .707. [Thank you, Minitab, for doing the ranking and the correlating for me.]

Had this been a random sample (which it is not...it is a sub-population) you might have wanted to make a statistical inference from the sample to the population from which the sample had been randomly drawn. The procedures for so doing are similar to the procedures for ordinary Pearson r and are referred to as "non-parametric". The word "non-parametric" derives from the root word "parameter" that always refers to a population. If you cannot assume that there is a normal bivariate distribution of X and Y in the population whose principal parameter is the population correlation, "non-parametric" inference is called for.

Testing for linearity and normality

If you're really compulsive and can't judge the linearity and normality of the relationship between X and Y by visual inspection of the X, Y plot, you might want to carry out formal tests for both. Such tests are available in the literature, but it would take us too far afield to go into them. And if your data "fails" either or both of the tests, there are data transformations that make the relationship "more linear" or "more normal".

Regression

Closely related to the Pearson correlation between two variables is the regression of one of the variables on the other, for predictive purposes. Some people can't say "correlation" without saying "regression". [Some of those same people can't say "reliability" without saying "validity".] In this section I want to point out the connection between correlation and regression, using as an example the data for my simple random sample of 30 players drawn from the entire population of 1034 players.

Suppose that you were interested not only in estimating the direction and the degree of linear relationship between their heights (X) and their weights (Y), but were also interested in using their data to predict weight from height for other players. You would start out just like we already have, namely plotting the data, calculating the Pearson r, and making a statistical inference, in the form of an estimation or a hypothesis test, from the sample r to the population ρ. But then the focus switches to the determination of the line that "best fits" the sample plot. This line is called the Y-on-X regression line, with Y as the dependent variable (the predictand) and X as the independent variable (the predictor). [There is also the X-on-Y regression line, but we're not interested in predicting height from weight.] The reason for the word "regression" will be explained soon.

I gave Minitab the command "regr c4 1 c3", asking for a regression analysis with the data for the dependent variable (Y) in column 4 and with the corresponding data for the one independent variable (X) in column 3. Here is what it (Minitab) gave me:

```
The regression equation is
 weight = - 111 + 4.39 height

 Predictor       Coef         Stdev     t-ratio
 Constant      -111.2         126.5       -0.88
 height         4.393         1.698        2.59

 s = 19.61       R-sq = 19.3%     R-sq(adj) = 16.4%

 Analysis of Variance

 SOURCE       DF           SS           MS
 Regression    1       2576.6       2576.6
 Error        28      10772.1        384.7
 Total        29      13348.7

 Unusual Observations
 Obs.  height    weight       Fit Stdev.Fit  Residual   St.Resid
   6     79.0    190.00    235.87      8.44    -45.87     -2.59R
  27     75.0    270.00    218.30      3.68     51.70      2.68R
```

Let's take that output one piece at a time.

The first and most important finding is the equation of the best-fitting regression line. It is of the form Y' = a + bX, where a is the Y intercept and b is the slope. [You might remember that from high school algebra. Y' is the "fitted Y", not the actual Y.] If you want to predict a player's weight from his height, just plug in his height in inches and you'll get his predicted weight in pounds. For example, consider a player who is 6 feet 2 inches (= 74 inches) tall. His predicted weight is -111 + 4.39 (74) = 214 pounds. Do you know that will be his exact weight? Of course not; it is merely approximately equal to the average of the weights for the six players in the sample who are approximately 74 inches tall. (See the plot, above.) But it is a heck of a lot better than not knowing his height.

The next section of the output provides some clarifying information (e.g., that the intercept is actually -111.2 and the slope is 4.393) and the first collection of inferential statistics. The intercept is not statistically significantly different from zero (trust me that a t-ratio of -0.88 produces a p-value greater than .05), but the slope is (big t, small p; trust me again). The intermediate column (Stdev) is the so-called "standard error" for the intercept and the slope, respectively. When combined with the coefficient itself, it produces the t-ratio, which in turn produces the [unindicated, but less than .05] p-value. Got that?

The third section provides a different kind of standard error (see below)...the s of 19.61; the squared correlation in percentage form (check that .193 is the square of .439); and the "adjusted" squared correlation of 16.4%. The squared correlation needs to be adjusted because you are trying to fit a regression line for two variables and only 30 points. [Think what would happen if you had just two points. Two points determine a straight line, so the line would fit perfectly but unsatisfactorily.] The square root of the adjusted squared correlation, which is equal to .405, might very well provide "the best" point estimate of the population correlation (see above).

The "Analysis of variance" section re-inforces (actually duplicates) the information in the previous two sections and would take us too far afield with unnecessary jargon, so let's ignore that for the time being. [Notice, however, that if you divide the MS for Regression by the MS for Error, you get the square of the t-ratio for the slope. That is not accidental. What is accidental is the similarity of the correlation of .439 to the slope of 4.393. The slope is equal to the correlation multiplied by the ratio of the standard deviation of Y to the standard deviation of X (neither is indicated, but trust me!), which just happens to be about 10.]

The last section is very interesting. Minitab has identified two points that are pretty far off the best-fitting regression line, one above the line (Obs. 27) and one below the line (Obs. 6), if you think they should be deleted. [I don't.]

Going back to predicting weight from height for a player who is 74 inches tall: The predicted weight was 214 pounds, but that prediction is not perfect. How good is it? If you take the 214 pounds and lay off the standard error (the second one) of 19.61 a couple of times on the high side and a couple of times on the low side you get a 95% confidence interval (yes, a different one) that ranges from 214 - 2(19.61) to 214 + 2(19.61), i.e., from 175 pounds to 253 pounds. That doesn't narrow things down very much (the range of weights in the entire population is from 150 pounds to 290 pounds), but the prediction has been based on a very small sample.

Why the term "regression"? The prediction just carried out illustrates the reason. That player's height of 74 inches is below the mean for all 30 players (you can see that from the plot). His predicted weight of 214 pounds is also below the mean weight (you can see that also), but it is closer to the mean weight than his height is to the mean height (how's that for a mouthful?), so his predicted weight is "regressed" toward the mean weight. The matter of "regression to the mean" comes up a lot in research in general. For example, in using a pre-experimental design involving a single group of people (no control group) measured twice, once before and once after the intervention, if that group's performance on the pretest is very low compared to the mean of a larger group of which it is apart, its performance on the posttest will usually be closer to the posttest mean than it was to the pretest mean. [It essentially has nowhere to go but up, and that is likely to be mis-interpreted as a treatment effect, whereas it is almost entirely attributable to the shape of the plot, which is a function of the less-than-perfect correlation between pretest and posttest. Think about it!]

Sample size

For my random-sample example I chose to take a sample of 30 players. Why 30? Why indeed? What size sample should I have I taken? Believe it or not, sample size is arguably the most important consideration in all of inferential statistics (but you wouldn't know it from actual practice, where many researchers decide on sample sizes willy-nilly and often not random sample sizes at that.) Briefly put, the appropriate sample size depends upon how far wrong you're willing to be when you make a statistical inference from a sample to the population from which the sample has been drawn. If you can afford to be wrong by a lot, a small sample will suffice. If you insist on never being wrong you must sample the entire population. The problem becomes one of determining what size sample is tolerably small but not so small that you might not learn very much from it, yet not so large that it might approximate the size of the entire population. [Think of the appropriate sample size as a "Goldilocks" sample size.] So, what should you do?

It depends upon whether you want to use interval estimation or hypothesis testing. For interval estimation there are formulas and tables available for determining the appropriate sample size for a given tolerable width for the confidence interval. For hypothesis testing there are similar formulas and tables for determining the appropriate sample size for tolerable probabilities of making Type I Errors and Type II Errors. You can look them up, as Casey Stengel used to say. [If you don't know who Casey Stengel is, you can look that up also!]

Multiple regression

I haven't said very much about age. The correlation between height and weight for the entire population is .532. Could that correlation be improved if we took into account the players' ages? (It really can't be worse even if age doesn't correlate very highly with weight; its actual correlation with weight for these data is only .158, and its correlation with height is -.074...a negative, though even smaller, correlation.) I went back to the full data and gave Minitab the command "regr c4 2 c3 c5", where the weights are in Column 4, the heights are in Column 3, and the ages are in Column 5. Here is what it returned (in part):

```
The regression equation is
 weight = - 193 + 0.965 age + 4.97 height

 1033 cases used 1 cases contain missing values

 Predictor       Coef         Stdev     t-ratio
 Constant     -192.66         17.89      -10.77
 age           0.9647        0.1249        7.72
 height        4.9746        0.2341       21.25

 s = 17.30       R-sq = 32.2%     R-sq(adj) = 32.1%
```

We can ignore everything but the regression equation (which, for those of you who are mathematically inclined, is the equation of a plane, not a line), the s, and the R-sq, because we have full-population data. Taking the square root of the R-sq of .322 we get an R of .567, which is higher than the r of .532 that we got for height alone, but not much higher. [It turns out that R is the Pearson r correlation between Y and Y'. Nice, huh?] We can also use the new regression equation to predict weight from height and age, with a slightly smaller standard error of 17.30, but let's not. I think you get the idea.

The reason it's called multiple regression is because there is more than one independent variable.

Summary

That's about it (for now). I've tried to point out some important statistical concepts that can be squeezed out of the baseball data. Please let me know (my e-mail address is tknapp5@juno.com) if you think of others. And by all means (another bad pun) please feel free to ask me any questions about this "module" and/or tell me about all of the things I said that are wrong.

It's been fun.

Oh, one more thing: It occurred to me that I never told you how to calculate a Pearson r. In this modern technological age most of us just feed the data into our friendly computer program and ask it to do the calculations. But it's possible that you could find yourself on a desert island some time without your computer and have a desire to calculate a Pearson r. The appendix that follows should help. (And you might learn a few other things in the process.)

APPENDIX [My thanks to Joe Rodgers for the great article he co-authored with Alan Nicewander, entitled "Thirteen ways to look at the correlation coefficient", and published in The American Statistician, 1988, volume 42, number 1, pages 59-66. I have included some of those ways in this appendix.]

Here are several mathematically equivalent formulas for the Pearson r (actually ρ, since these formulas are for population data):

1. $$\rho = \frac{\sum z_X z_Y}{n}$$

This is the best way to "think about" Pearson r. It is the average (mean) product of standardized variable X and standardized variable Y. A standardized variable z , e.g., z_X , is equal to the raw variable minus the mean of the variable, divided by the standard deviation of the variable, i.e., $z_X = (X - M_X)/s_X$. This formula for r also reflects the product-moment feature (the product is of the z's; a moment is a mean). Since X and Y are usually not on the same scale, what we care about is the relative relationship between X and Y, not the absolute relationship. It is not a very computationally efficient way of calculating r, however, since it involves all of those intermediate calculations that can lead to round-off errors.

2. $\rho = 1 - 1/2\ [s^2 \text{ of } (z_Y - z_X)]$

This a variation of the previous formula, involving the difference between "scores" on the standardized variables rather than their product. If there are small differences, the variance of those differences is small and the r is close to +1. If the differences are large (with many even being of opposite sign), the variance is large and the r is close to -1.

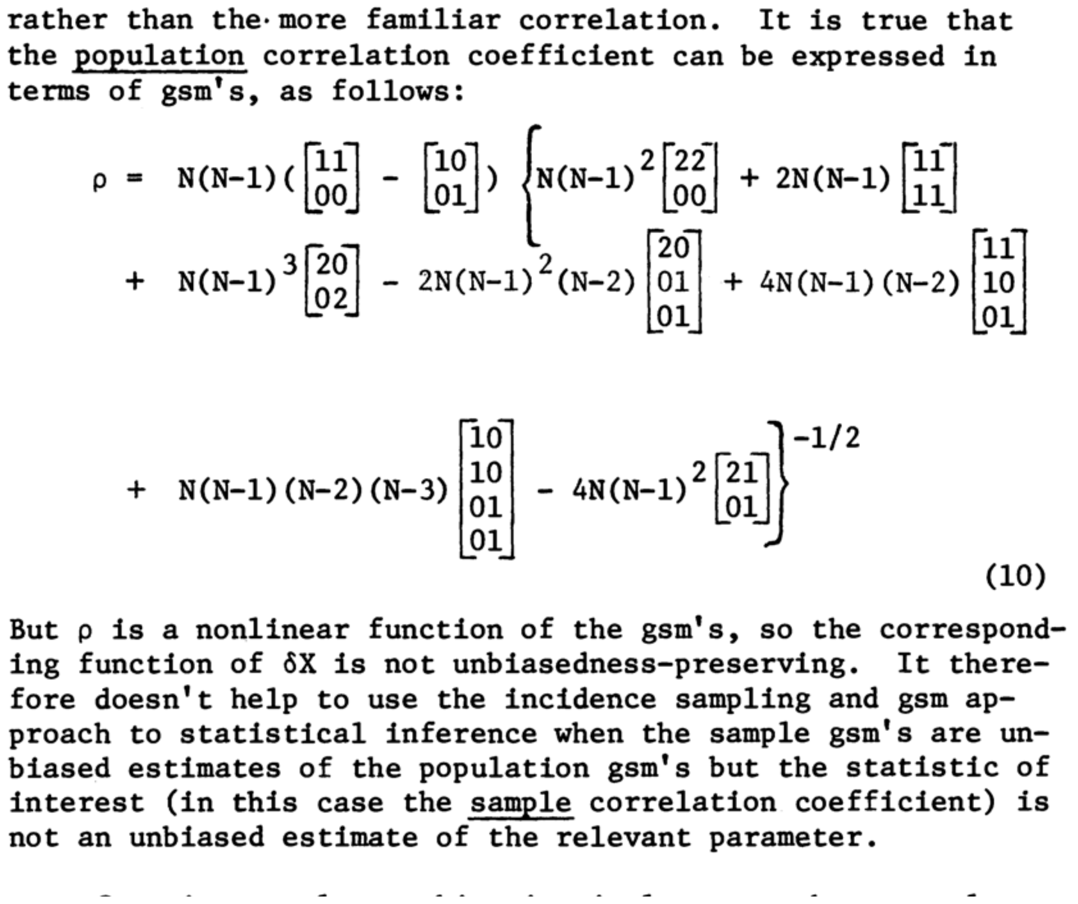
3. $$\rho = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2]\,[n\sum Y^2 - (\sum Y)^2]}}$$

(N.B.: the square root is taken of the product of the bracketed terms in the denominator)

This formula looks much more complicated (and it is, in a way), but involves only the number of observations, the actual X and Y data, and their squares. In "the good old days" before computers, I remember well entering the X's in the left end of the keyboard of a Monroe or Marchant calculator, entering the Y's in the right end, pushing a couple of buttons, and getting $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, and $2\sum XY$ in the output register all in one fell swoop! [That was quite an accomplishment then.]

4. ρ = cosine (θ), where θ is the angle between a vector for the X variable and a vector for the Y variable in the n-dimensional "person space" rather than the two-dimensional "variable space". [If you don't know anything about trigonometry or multi-dimensional space, that will mean absolutely nothing to you.]

5. If you really want a mathematical challenge, try using the formula in the following excerpt from an article I wrote about 25 years ago (in the Journal of Educational Statistics, 1979, volume 4, number 1, pages 41-58). You'll probably have to read a lot of that article in order to figure out what a gsm is, what all of those crazy symbols are, etc.; but as I said, it's a challenge!



rather than the·more familiar correlation. It is true that the population correlation coefficient can be expressed in terms of gsm's, as follows:

$$
\begin{aligned}
\rho = {} & N(N-1)\left(\begin{bmatrix}11\\00\end{bmatrix} - \begin{bmatrix}10\\01\end{bmatrix}\right)\left\{N(N-1)^2\begin{bmatrix}22\\00\end{bmatrix} + 2N(N-1)\begin{bmatrix}11\\11\end{bmatrix}\right. \\
& + N(N-1)^3\begin{bmatrix}20\\02\end{bmatrix} - 2N(N-1)^2(N-2)\begin{bmatrix}20\\01\\01\end{bmatrix} + 4N(N-1)(N-2)\begin{bmatrix}11\\10\\01\end{bmatrix} \\
& \left. + N(N-1)(N-2)(N-3)\begin{bmatrix}10\\10\\01\\01\end{bmatrix} - 4N(N-1)^2\begin{bmatrix}21\\01\end{bmatrix}\right\}^{-1/2}
\end{aligned}
\tag{10}
$$

But ρ is a nonlinear function of the gsm's, so the corresponding function of δX is not unbiasedness-preserving. It therefore doesn't help to use the incidence sampling and gsm approach to statistical inference when the sample gsm's are unbiased estimates of the population gsm's but the statistic of interest (in this case the sample correlation coefficient) is not an unbiased estimate of the relevant parameter.