The Pretest! Pretest! Pretest! Assignment (Example 2)

May 19, 2003

1 Statement of Purpose and Description of Pretest Procedure

When one designs a Math 10 exam one hopes to measure whether "a student's ability to wield the statistical ideas covered in the course". However, how well a students takes a particular exam may in fact not reflect such a goal. Here I describe a trick used in our math 10 exam to help determine whether our exam really did measure "a student's ability to wield the statistical ideas covered in the course".

1.1 Quantifying My Hopes

In order to test whether a person posses a certain skill it often useful to develop multiple and distinct ways of a assessing that skill. If two such statistics are really measuring the same sort of thing, then they should be correlated in some way to each other. One might hope this will be a linear correlation and hence that the strength of such a linear correlation could be used to give some indication of how well these different measurements are capturing the same underlying quality.

Our midterm exam is an example of such a measurement system. It was intentionally broken into two rather different parts. Part 1, which was comprised of problems very much so like the exercises from the text; and part 2, which examined the student's grasp of the material via the exploration of a pair of experiments performed in class. As the instructor, I was hoping that both these section were measuring the same sort thing, namely the "a student's ability to wield the statistical ideas covered in the course". The fact that the exam was broken into two parts allows me to test whether these different measurements where measuring the same sort of thing. Hence, I will do what I can do, and test the correlation between the performances on the two exam parts.

To be more precise, I will compute the scatter plot, the line of best fit, strength of the correlation, and then test this correlation in order to estimate the P-value associated to this correlation under the null hypothesis that no such correlation exist. If this hypothesis test comes back as statistically significant I will accept this correlation as a real phenomena.

On an interpretive level, if the strength of the correlation is above .5 and the P-value is less than 0.05, then I will choose to chose to **believe** that the exam, at least in part, truly reflected "the students' ability to wield the statistical ideas covered in the course". Making such an interpretive conclusion from such an experiment is very common, though, and **I need to emphasize**, this hypothesis test itself ONLY tells us only that the such scores are likely to be correlated. I would need to perform a more detailed scientific study to determine whether or not the causal conclusion concerning the "student's ability" has any scientific merit.

1.2 The Data

Part 1 of this exam was graded by my assistant and Part 2 was graded by myself. Hence I have eliminated any "desire for correlation" bias. The grade are as follows.

Part1	*Part2*
54	39
43	15
44	26
34	15
48	44
55	42
44	32
48	32
47	32
55	45
46	31
43	42
35	6
53	39
38	18
35	12
33	12
33	11
35	10
33	22
50	45
38	26
43	36
53	42
45	36
55	44
44	14
50	41
55	43
36	18

I randomly permuted the above grades in order to protect my students' identities. (Each of my students should find a pair of scores corresponding to you.) In the figure , we see the scatter plot of this data. The correlation coefficient is 0.87 while the strength of this correlation is 75.6 percent. Further more $n = 30$, $b_1 = 1.47$, $b_0 = -35.74$ and the s_{b_1} from section 9.3 satisfies $s_{b_1} = .157$. In particular the line of best fit will be dented as $\hat{y} = b_1 x + b_0 = 1.47x - 35.74$.

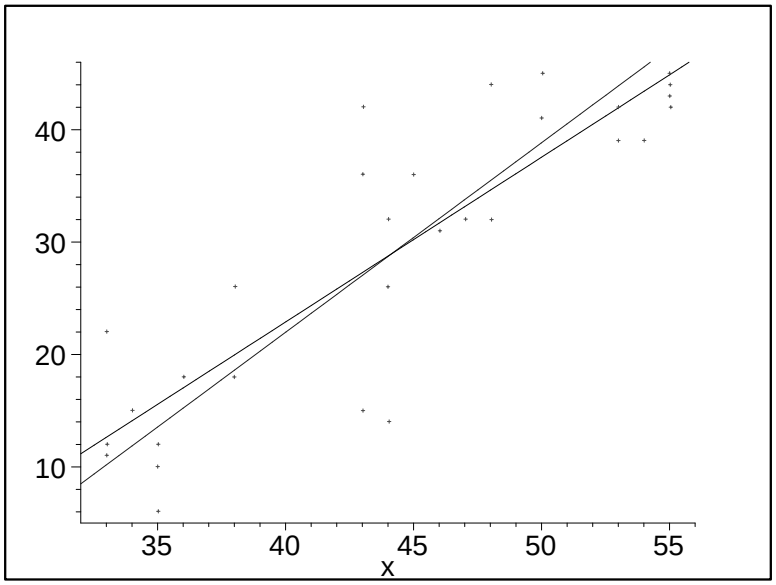

Figure 1: Here we see a scatter plot of our exam data. Namely each point corresponds to a exam. The score associated to part 1 of the exam is the x coordinate of the point, and the score associated to part 2 of the exam is the y-coordinate of the point.

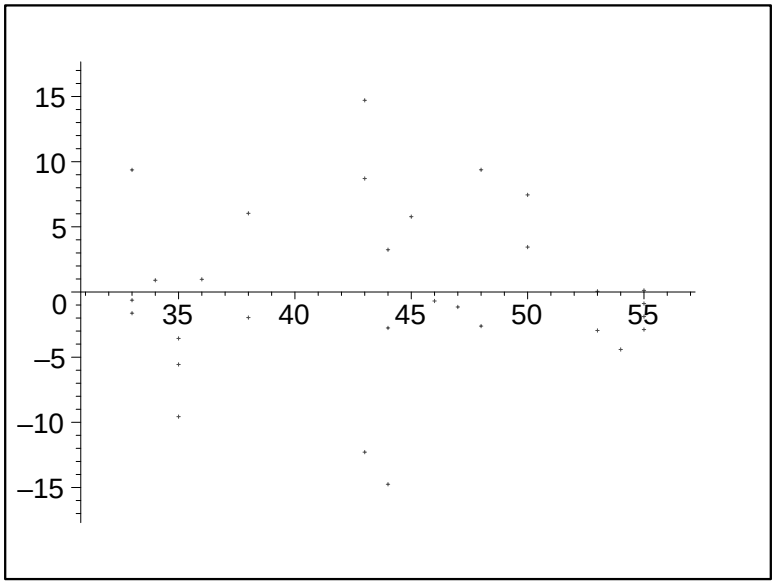

Figure 2: Here is our residual plot. Namely I have graphed the x-coordinate value of our points from figure 1, against the difference between the actual y-coordinate values from figure 1 and the corresponding $\hat{y}$ value. Real $\hat{y}$ is the second part's score as estimated from the line of best fit.

1.3 The Hypothesis Test

The data appears nicely linearly correlated. If I want to test this correlation I can test the positivity of slope β_1 by using a t-test with $n-2=28$ degrees of freedom as described in section 9.3. To be more precise, I test the alternative hypothesis that $\beta_1 > 0$ against the null hypothesis that $\beta_1 = 0$ wit ha levl of significance of $\alpha = 0.01$. In order to honestly perform such a test, I need to verify that my data satisfies the properties of the linear regression model. For a small data set one nice way to asses this is to look at the plot of the residuals, see figure 2. Here I will enumerate the the properties of the linear regression model as was done in section 9.2.

3. Homoscedasticity. Looking at the residuals in figure 2 we see that this assumption appears reasonable. Namely, the variance is not constant but there is not a very dramatic change. If anything, it appears that the variance for higher x-coordinate scores shrinks a little. This is almost certainly do to the fact the exam failed to separated scores at the high end. We can see this issue in the histogram of the scores in figure 3.

2. Linearity. The data looks like the relationship is linear. Looking at the

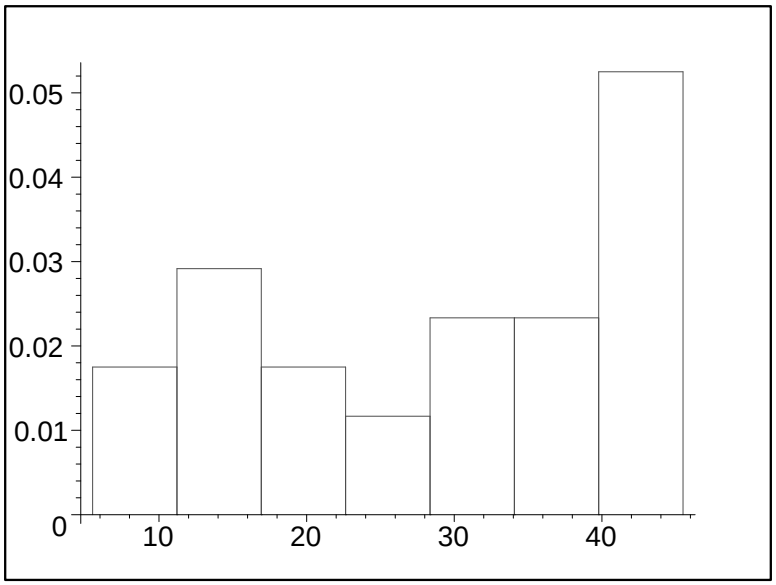

Figure 3: Here we see a histogram of part 2 of the exam. The general shape of this picture is quite typical for a poorly written exam, namely it is ”smashed on the left”. Typically this means that the exam was not hard enough to separate out the students at the high end. Unlike most such curves, this one also has a certain bimodality to it. Upon consulting with my Math 10 class we determined that this bimodality was likely to due to the fact that the two question in part 2 of the exam were of very different difficulty levels. Most people had a good idea how to handle the first of the two questions, while some positive fraction of the class were unable to even get started on the second question. Hence the bimodality.

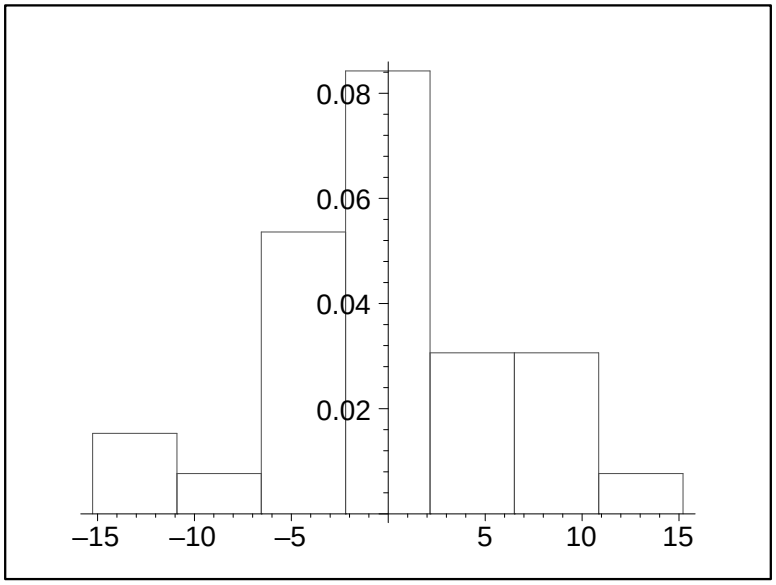

Figure 4: Here is histogram of the residuals. Looking at our residual data we anticipate the decrease in the variance among higher scores explains why this histogram is a bit tight around zero. Certainly it is reasonably "bell shaped".

residuals linearity looks very reasonable indeed, namely the graph does appear to hover around the line $x = 0$ as needed.

4. Independence should be fine here, since the honor principle assure me that who a student's neighbor was when they took the exam will not effect the exams score.

1. Normality. "Experience" tells me to expect Bell shaped curves in general when one fixes the score on one exam and compares another exam score. Since the data nearly satisfies part 3 above, we can test this by simply seeing if the histogram of the residuals has a bell type shape. Looking at this graph in figure 4 we see this assumption looks pretty good. Once again the distribution is perhaps a bit tight around 0, due to the fact that the variance decreased among the high score, see figure 3.

Since we have basically satisfied 1-4 of the regression model we may utilize the hypothesis test on b_1, described above In other words under the null hypothesis,

$$\frac{b_1}{s_{b_1}} = 9.311$$

is basically a t distribution as $30 - 2 = 28$ degrees of freedom. In particular, the P value is nearly zero and such correlation is virtually certain to exist and we easily

pass our hypothesis test. In fact, the 99 percent confidence interval confidence interval for the β_1 I based on this data is

$$[b_1 - t_{\alpha/2} s_{b_1}, b_1 + t_{\alpha/2} s_{b_1}] = [1.47 - (2.76)(.157), 1.47 + (2.76)(.157)] = [1.04, 1.90].$$

2 Conclusions

Looking at my data it looks like I failed to make a really good exam with respect to the linear correlation model's needs. The bulk of the problem is the phenomena explored in figure 3, namely, that the exam failed to separate at the high end of the score spectrum. In the process, the variance of my residuals at the high end was much smaller. This is a violation of homoscedaticity. We saw that this also affected our exploration of the Normality assumption. It seems that these problems could be avoided by attempting to make a more challenging exam, though I'd need to be careful not to make it to hard (in order to avoid a clumping of the values at the low end). Clearly such an attempt can be made in preparing the final.

More importantly, my correlation measurements did not necessarily quantify whether the exam really measured "a student's ability to wield the statistical ideas covered in the course". In particular, since both parts of the exam were taken at the same time it does not take into account the possibility that a given student is just having a bad day. To arrive at a more meaningful sense of whether this goal has been accomplished, I will need to compare performances between more distinct measurement of "a student's ability". It would be nice to compare measurement made a different times and in more transparently distinct ways. For example, the correlation between mini-project grades and the final exam scores would be much more revealing, since the mini-projects and final exams are both attempting to measure "a student's ability to wield the statistical ideas covered in the course" but in VERY different ways. Hence, a more suitable second pretest will be to look at the correlation between the Final and the Mini-project, rather than the correlation between two "parts" of the final. On an interpretive level, if the strength of the such a correlation is above .5 and the P-value is less than 0.05, then I will happily chose to **believe** that the assessment strategy used to determine my students' grades was, at least in part, truly reflecting "my students' ability to wield the statistical ideas covered in the course".