

University of York
Department of Health Sciences
Applied Biostatistics

Suggested Answer to the Sample Assessment

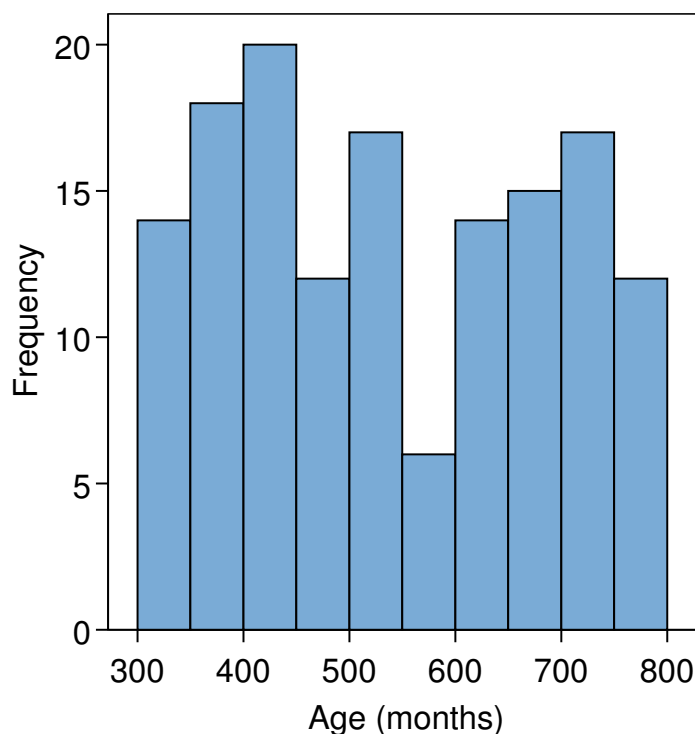
This suggested answer was not used in the marking of the assessment but has been produced especially for current students. Text in square brackets [] are my explanatory comments, not part of the suggested answer.

Applied Biostatistics Assignment 2005/2006

1. Produce charts (or use some other method) to show the distribution of age and years since diagnosis. Briefly describe what those charts show you about the distribution of the variables.

These are both quantitative variables. The best way to show the distribution of a quantitative variable is a histogram. We could also use a box and whisker plot, or a Normal quantile plot.

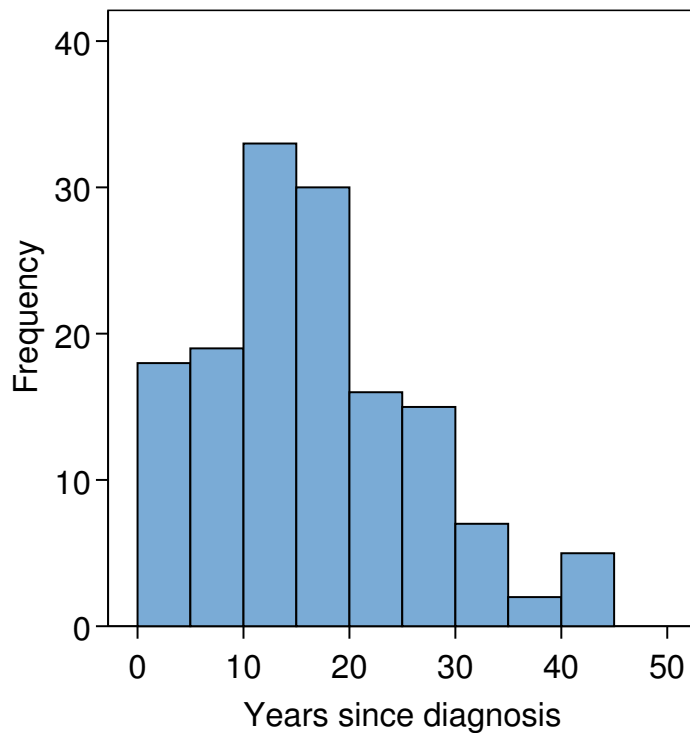
The histogram of age is as follows:



[Note that I have edited the graph to give bigger fonts and reduced the size of the graph so that the font is of similar size to my text. I have altered the interval size to 50 from the default width of 33.333, and changed the scale on the Age axis. I have also used Variables View to give age a label with units.]

There are no obvious outliers. This appears to be a bimodal distribution, with one mode around 400 months and another around 700 months. This suggests that there are two distinct populations of patients.

The histogram of years since diagnosis is as follows:



This distribution is positively skew, the tail on the right being much longer than the tail on the left. There are no obvious outliers. Some of these patients have had psoriasis for a very long time.

2. The people should have been randomly allocated to each of the two groups (treatment and control). Carry out a test, to compare the mean age of the people in each group. What is the result? What do you conclude?

Carry out a test to compare the proportions of males and females in each group. What is the result? What do you conclude?

Age is a quantitative variable, so a comparison of means is indicated. There were 67 subjects in the treated group and 78 in the control group. [This information was obtained by the Frequencies command.] As there are more than 50 subjects in each group, we can use the large sample z method to compare means. This means we can ignore the awkward shape of the age distribution. The results are:

Mean and standard deviation for each group

Group	Number	Mean	Standard deviation	Standard error of the mean
Treated	67	532	147	17
Control	78	547	146	18

[I did this using the two sample t test command. This is the SPSS output:

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
Age (months)	0	78	547.53	146.123	16.545
	1	67	532.43	146.902	17.947

I have given the mean and standard deviation to three significant figures, which seemed enough. We do not need all the figures SPSS produces.]

The mean age of control subjects exceeded the mean age of treated subjects by 15 months, 95% confidence interval –33 to +63 months, $P = 0.5$.

There is no evidence that in the population from which these subjects come there is any difference in mean age between treated and control subjects. We can conclude that there is nothing to suggested that subjects were not allocated randomly to treatment group.

[This was the output:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age (months)	Equal variances assumed	.240	.625	.619	143	.537	15.093	24.400	-33.138	63.324
	Equal variances not assumed			.618	139.494	.537	15.093	24.410	-33.168	63.354

The large sample z method corresponds to the 'equal variance not assumed' row. We do not need any assumptions about variance, so we ignore the Levene test.]

Male or female is a dichotomous, qualitative variable, so a chi-squared test or Fisher's exact test would be appropriate.

The two by two cross-tabulation of sex by group is as follows:

	Treated	Control	Total
Female	37 (47.4%)	35 (52.2%)	72 (49.7%)
Male	41 (52.6%)	32 (47.8%)	73 (50.3%)
Total	78 (100.0%)	67 (100.0%)	145 (100.0%)

The percentage of females in each group was similar and the difference was not significant (chi-squared = 0.33, 1 degree of freedom, $P = 0.5$). Hence there is no evidence that in this population sex differed between the treatment groups and the data are consistent with the groups being allocated to treatment group randomly.

[This was the output:

sex * group Crosstabulation

			group		Total
			0	1	
Sex	0	Count	37	35	72
		Expected Count	38.7	33.3	72.0
		% within group	47.4%	52.2%	49.7%
	1	Count	41	32	73
		Expected Count	39.3	33.7	73.0
		% within group	52.6%	47.8%	50.3%
	Total	Count	78	67	145
		Expected Count	78.0	67.0	145.0
		% within group	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.333(b)	1	.564		
Continuity Correction(a)	.168	1	.682		
Likelihood Ratio	.333	1	.564		
Fisher's Exact Test				.619	.341
Linear-by-Linear Association	.330	1	.565		
N of Valid Cases	145				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 33.27.

Note that I have not used the SPSS table, but produced my own from the SPSS output and that I have lined up the table neatly in columns. The expected frequencies are all large enough for the chi-squared test.]

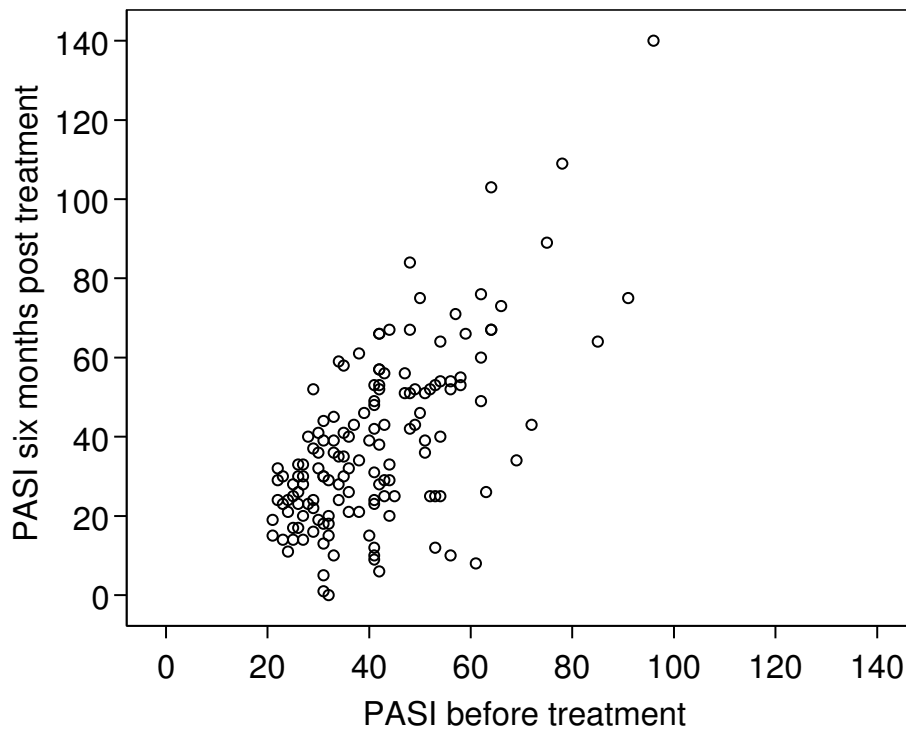
We can conclude that there is no evidence from either the age or the sex distribution to suggest that the patients were not allocated to the two groups randomly.

3. Was the treatment successful at reducing the PASI?

PASI is a quantitative variable, so we want to compare means. We want to compare the mean PASI six months after treatment. The PASI after treatment is likely to be related to the baseline PASI and we should take this into account in the analysis. There are two ways to do this. We could calculate the change in PASI and use that, or we could do a multiple regression and use baseline PASI as a predicting variable or covariate.

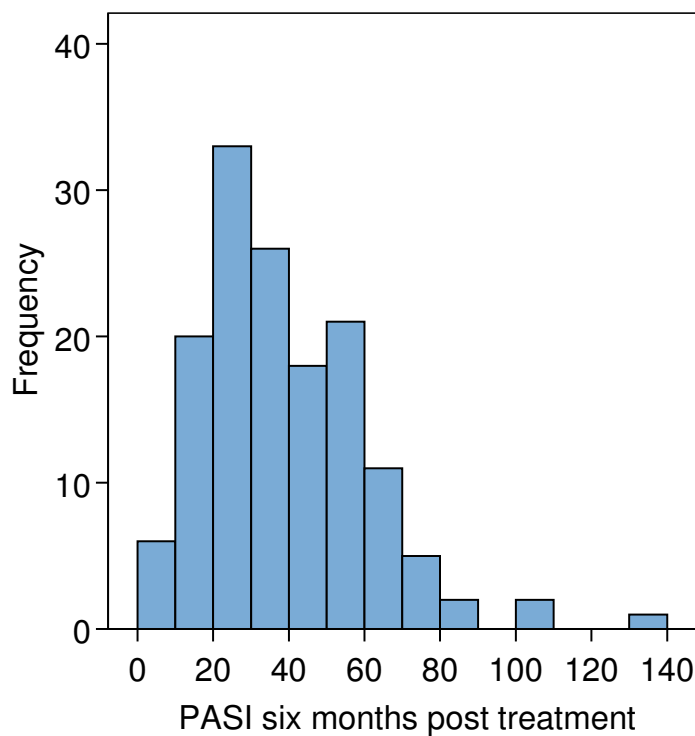
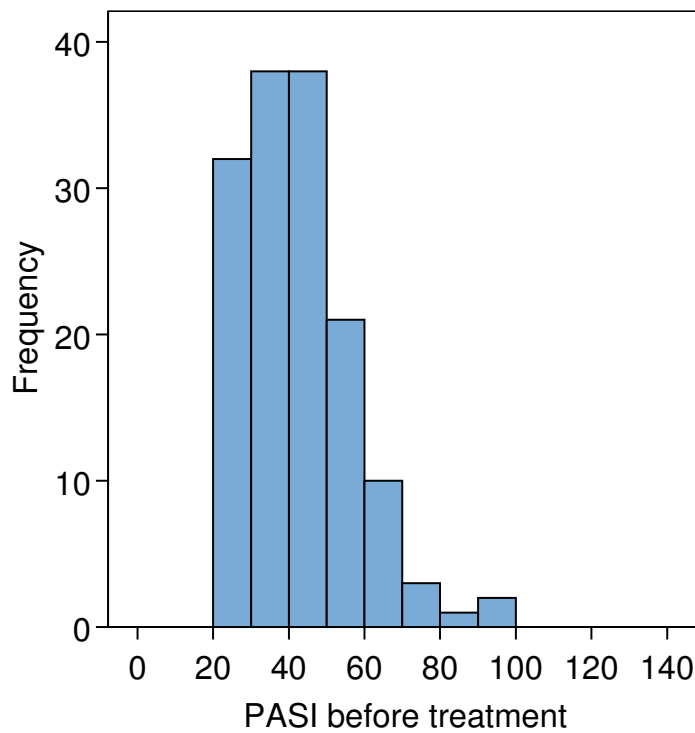
[I shall do both, for illustration. People lost marks for not using the baseline measurement, but simply comparing the post-treatment PASI between the two groups.]

Plotting the data before and after treatment will show whether we need to take the baseline PASI into account. Baseline and post-treatment PASI scores are clearly related, as the scatter diagram shows:



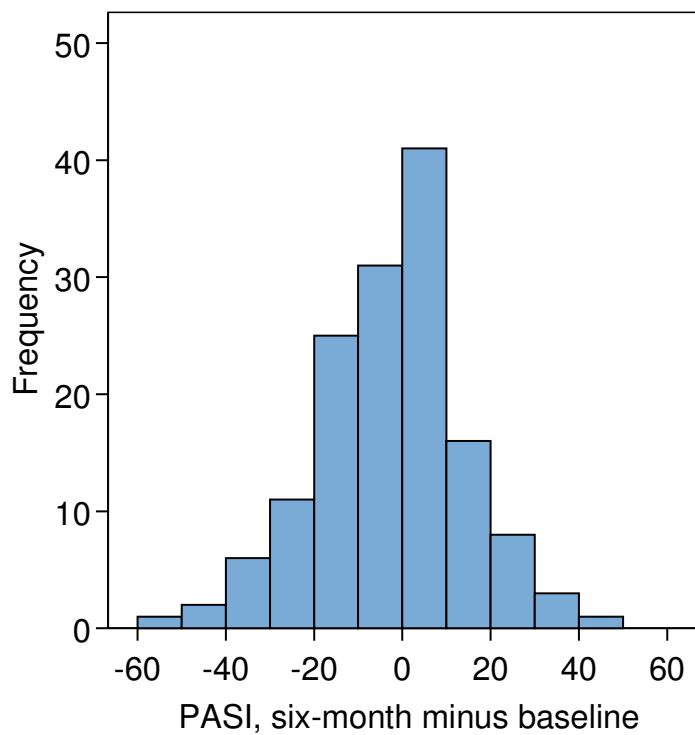
[As usual, I have edited the graph to make the text bigger, etc. I have made the two scales the same, because this is the same variable measured twice. It would have been nice to use a different symbol for each group and to draw a line of equality, where the baseline and six-month PASI would be equal, but I am not an SPSS user and this defeated me.]

Inspection of the graph suggests that both PASI variables are positively skew, as histograms confirm:



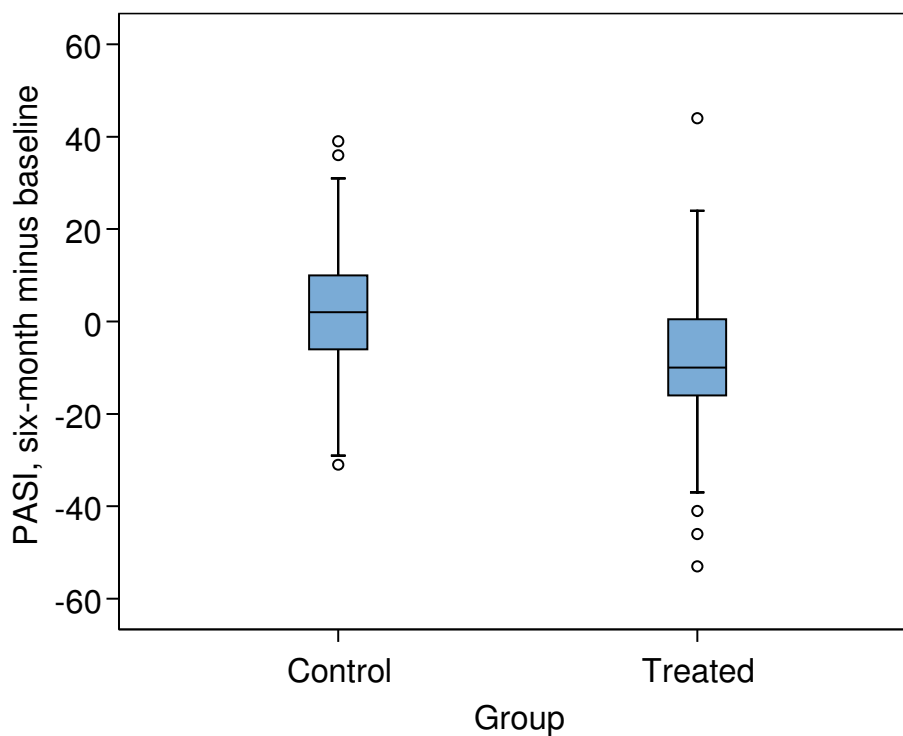
Although the histogram of PASI at six months suggests that there may be an outlier, inspection of the scatter diagram shows that this subject had the highest PASI on both occasions and so there is no reason to reject this measurement.

The distribution of the differences, six-month PASI minus baseline, was as follows:



[I used Transform, Compute to calculate the difference, then labelled it in Variable View. As usual, I edited the histogram.]

The distribution of the differences appears approximately Normal, with no obvious outliers. We can compare the two groups using a box plot:



This suggests that the differences are more negative for the treated group. We can compare these differences using either a two sample t test or a z test, because the data appear approximately Normal with similar variances in the two groups and the samples are larger than 50. I shall use the two sample t test.

The mean and standard deviation of the change in PASI for each group were:

Group	Number	Mean	Standard deviation	Standard error of the mean
Treated	67	-8.70	17.1	2.1
Control	78	1.62	14.9	1.7

[I did this using the two sample t test command. This is the SPSS output:

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
PASI, six-month minus baseline	Control	78	1.6282	14.89714	1.68677
	Treated	67	-8.7015	17.06939	2.08536

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
PASI, six-month minus baseline	Equal variances assumed	.971	.326	3.891	143	.000	10.32970	2.65457	5.08244	15.57696
	Equal variances not assumed			3.851	132.134	.000	10.32970	2.68215	5.02420	15.63520

]

The difference was highly significant ($P < 0.001$) and the mean difference in PASI in the treated group was less than the mean in the control group by 10.3 units, 95% confidence interval 5.0 units to 15.6 units. Hence there is strong evidence that the treatment reduced mean PASI, by between 5.1 and 15.8 units.

We can also use multiple regression. The effect of being in the treated group was -10.3, 95% confidence interval -15.6 to -5.1, $P < 0.001$.

[I used linear regression with PASI at six months as the dependent variable and group and PASI at baseline as the independent variables. The output was:

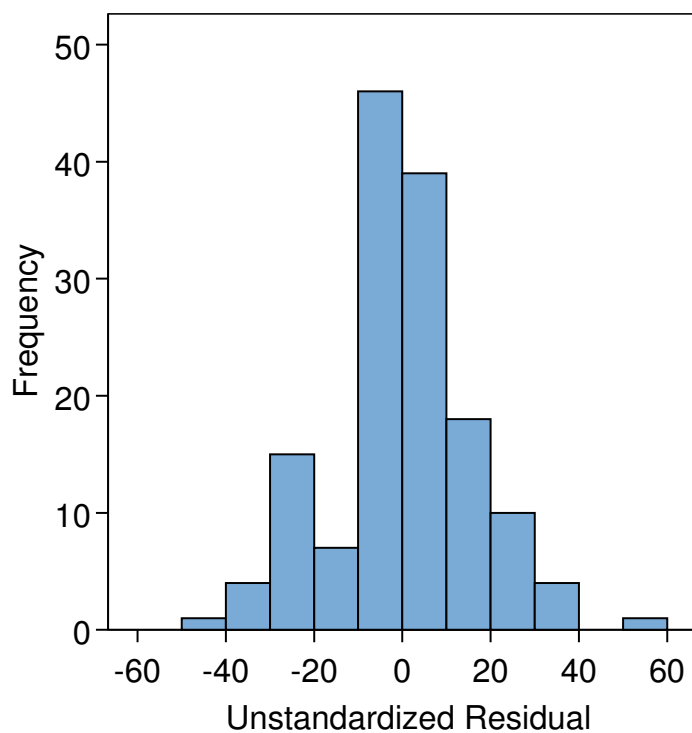
Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.759	4.157		.423	.673	-6.458	9.976
	Group	-10.327	2.665	-.235	-3.874	.000	-15.596	-5.058
	PASI before treatment	.997	.091	.660	10.896	.000	.816	1.178

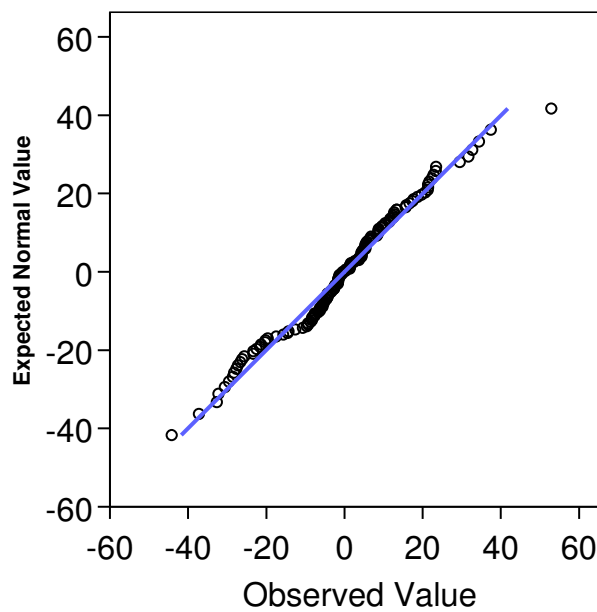
a. Dependent Variable: PASI six months post treatment

]

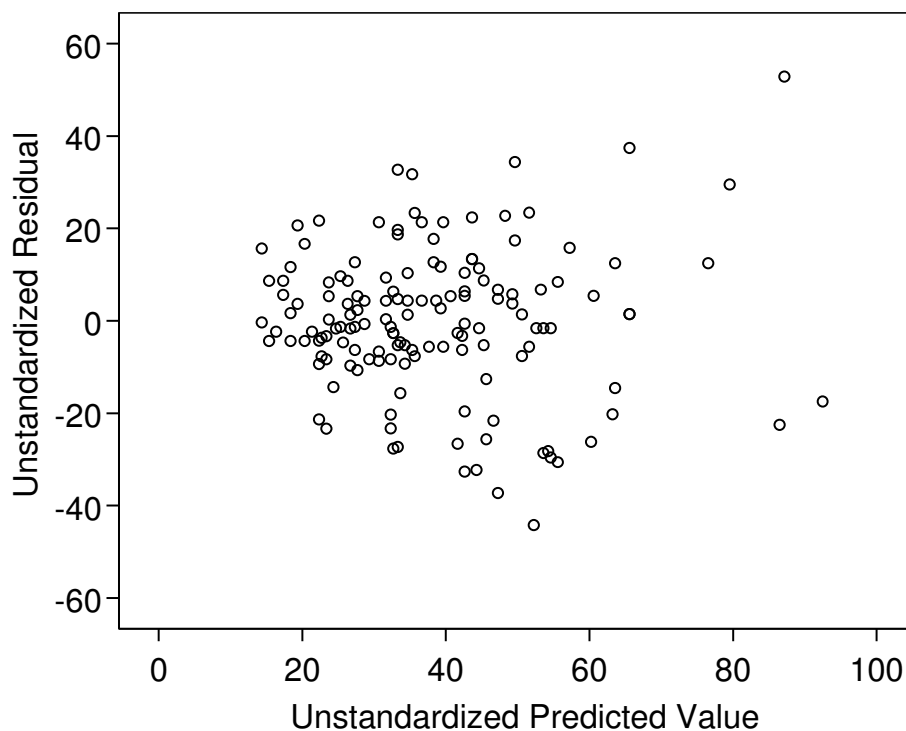
We should check the assumptions of the regression by residual plots. The histogram and Normal plot of the residuals were as follows:



Normal Q-Q Plot of Unstandardized Residual



The histogram appears to fit a Normal distribution and the Normal plot appears fairly straight. Next we look at the plot of residuals against the predicted value, to assess the uniformity of the variance.



There may be some slight increase in variability at large predicted values, but not much. The regression analysis appears to be valid.

We can conclude that treatment reduced mean PASI by between 5 and 15 units.

5. Did a higher proportion of people in the treatment group feel that they had improved than those in the control group?

Feelings of improvement is a dichotomous, qualitative variable so a cross-tabulation and chi-squared or Fisher's exact test is indicated.

The tabulation of perceived improvement by treatment group is as follows:

	Treated	Control	Total
Improved	49 (73.1%)	41 (52.6%)	90 (62.1%)
Not improved	18 (26.9%)	37 (47.4%)	55 (37.9%)
Total	78 (100.0%)	67 (100.0%)	145 (100.0%)

The difference was statistically significant, chi-squared = 6.48, 1 degree of freedom, $P = 0.01$. There were no cells with expected frequencies less than five, so the chi-squared test is valid for this table.

We can conclude that there is good evidence that patients in the treated group were more likely to think that they had improved than were patients in the control group.

[The output was:

improved * Group Crosstabulation

			Group		Total
			Control	Treated	
improved	0	Count	37	18	55
		Expected Count	29.6	25.4	55.0
		% within Group	47.4%	26.9%	37.9%
	1	Count	41	49	90
		Expected Count	48.4	41.6	90.0
		% within Group	52.6%	73.1%	62.1%
Total		Count	78	67	145
		Expected Count	78.0	67.0	145.0
		% within Group	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.478(b)	1	.011		
Continuity Correction(a)	5.633	1	.018		
Likelihood Ratio	6.578	1	.010		
Fisher's Exact Test				.016	.008
Linear-by-Linear Association	6.433	1	.011		
N of Valid Cases	145				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 25.41.

]

The relative risk of reporting improvement for patients in the control group was 1.39 with 95% confidence interval 1.08 to 1.80.

[To do this I had to reverse the order of the groups, to make the order treated then control. I created a variable grouprev by 1 – group. This was then 0 for treated and 1 for control. I then had to switch the row and column variables:

grouprev * improved Crosstabulation

			Improved		Total
			0	1	
grouprev	.00	Count	18	49	67
		Expected Count	25.4	41.6	67.0
		% within improved	32.7%	54.4%	46.2%
	1.00	Count	37	41	78
		Expected Count	29.6	48.4	78.0
		% within improved	67.3%	45.6%	53.8%
Total		Count	55	90	145
		Expected Count	55.0	90.0	145.0
		% within improved	100.0%	100.0%	100.0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for grouprev (.00 / 1.00)	.407	.202	.819
For cohort improved = 0	.566	.358	.896
For cohort improved = 1	1.391	1.077	1.797
N of Valid Cases	145		

The relative risk we want is for improved = 1. Note that 1.797 rounds up to 1.80. It is a good idea to check that $73.1\% / 52.6\% = 1.39$, which it does.]

Hence we can conclude that the treatment lead to increase in the proportion of patients reporting improvement, by a factor estimated to be between 1.08 and 1.80.

Our final conclusions are that there is no reason to think that this trial was not correctly randomised and that patients receiving the new treatment had an improvement in mean PASI score, compared to the control group, which is estimated to be between 5 and 15 units. They were also more likely to report an improvement, by a factor estimated to be between 0.05 and 1.80.

Martin Bland,

11 December 2006