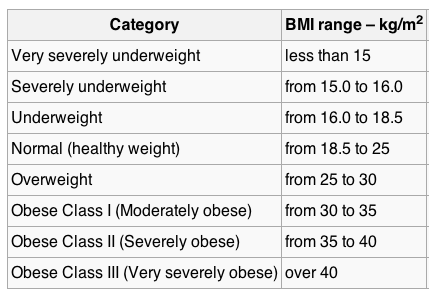
# Instructions

1. Read through "ais-data-description.rtf" and learn about the data file
2. Open "ais.csv" in SPSS
3. Prepare the data file: (a) Add variable labels; (b) Recode sex and sport into numeric variables with value labels rather than and then tidy up the value labels so that they represent full names (e.g., f into "female", "gym" into "gymnastics"); (c) Create a reduced version of the variable sports that that contains only (Swimming, Track > 400 metres, and basketball)
4. Get an overview of the data (a) Examine frequencies and graphs of all the variables to understand them; (b) cross tabulate sport and gender; (c) examine correlations between numeric variables.
5. (a) Calculate your own version of BMI and check that it agrees with that which is already http://en.wikipedia.org/wiki/Body\_mass\_index; (b) create a categorical variable where 1 = very severely underweight; 2 = severely underweight; (c) Produce a table of sport by BMI category
6. (a) Run a t-test with gender as IV; height as DV (see analyze - compare means - independent); (b) Run ANOVA with gender as IV, height as DV (Analyze - GLM - Univariate; gender as fixed factor; height as DV); (c) Run a regression with gender as IV; height as DV (Analyze - Regression- Linear). (d) Compare and contrast the output. (e) examine normality of residuals and homogeneity of variance.
7. (a) Run an ANOVA with IV=sport; DV=BMI obtain post-hoc tests. (b) Interpret the results. (c) examine normality of residuals and homogeneity of variance
8. Run an ANOVA with IV=reduced\_sport (i.e., swim, T400, and basketball) and gender; DV=height. Produce a plot. Provide an overall interpretation of main effects and the interaction effect. Follow-up where appropriate with analysis of simple effects, post hoc tests, etc.
9. **MANOVA:** (a) Review the correlations between all the five blood variables (i.e., rcc, wcc, hc, hg, ferr); (b) Run a MANOVA (Analyze - GLM - Multivariate) with gender as IV (i.e., fixed factor) and the five blood variables as DV and interpret the multivariate effects (c) interpret the univariate effects from the MANOVA output; (d) run a discriminant function analysis with gender as DV and blood variables as IVs (Analyze - Classify - discriminant) and interpret the results (d) compare and contrast the interpretation of the group differences from the univariate versus the DFA approach (e) create a single blood variable by converting each of the five blood variables into z-scores and then summing them up (check that they are all positively correlated to ensure that none of them need to be reversed), then run a univariate ANOVA with this composite variable as a DV; interpret the results (f) compare and contrast creating your own composite with univariate ANOVA with the multivariate MANOVA output.

# Answers

## 1. Read through "ais-data-description.rtf" and learn about the data file

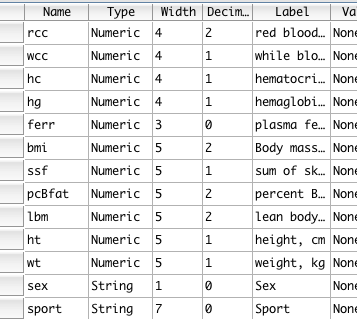
For example, get a sense of what each of the variables means and the units that they are on. If it is a categorical variable, understand the categories.

## 2. Open "ais.csv" in SPSS

This is a comma-separated value file as indicated by the ".csv" extension. Use "File - Read Text Data" to import the data. Make sure to indicate that variable names are included in the top of the file.

## 3. Prepare the data file: (a) Add variable labels;

Just add the labels from the data description:



## (b) Recode sex and sport into numeric variables with value labels rather than and then tidy up the value labels so that they represent full names (e.g., f into "female", "gym" into "gymnastics");

Note that when importing from a text file into SPSS, you can often end up with two issues regarding categorical variables First, the value labels are imported as string variables, whereas SPSS typically prefers numbers with value labels. Second, the actual labels may not be exactly what you want.

To perform the above steps, we could do the following:

Use "transform - automatic recode" on sport and gender to create two new numeric variables (e.g., "sport\_numeric", and "gender\_numeric") with value labels that correspond to the original text in the variable.

Then edit the value labels in SPSS so that they now have the preferred labels.

## (c) Create a reduced version of the variable sports that that contains only (Swimming, Track > 400 metres, and basketball)

This is mainly done to create a categorical variable with a limited number of levels for subsequent analysis. Also each of these categories has a reasonable number of males and females which makes the subsequent factorial ANOVA with gender somewhat interesting.

One way to do this is to use "Transform - recode into different variable". Use "sport "as the input variable and give the output variable a name like "sport\_reduced", and click "change". Under "old and new values..." enter old value and new value for each of the three sports you want to retains. Then, click "all other values" and "system missing".

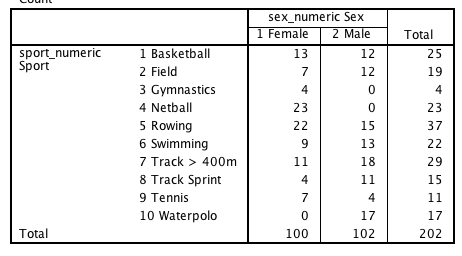
## 4. Get an overview of the data (a) Examine frequencies and graphs of all the variables to understand them;

"analyze - descriptives - frequencies" with the histogram option is a pretty good starting point. In particular this should give you a feel for the distributions of the numeric variables: i.e., many of them are roughly normally distributed, and a few have some positive skew. Note that this an important part of getting a sense of your data.

## (b) cross tabulate sport and gender;

When getting a sense of how important categorical data relate in your dataset it is useful to create a cross-tabulation of frequencies.

In SPSS, go to analyze - descriptives - crosstabs and put one variable in rows and the other in columns. Often it's easier if the smaller number of categories is in the columns.

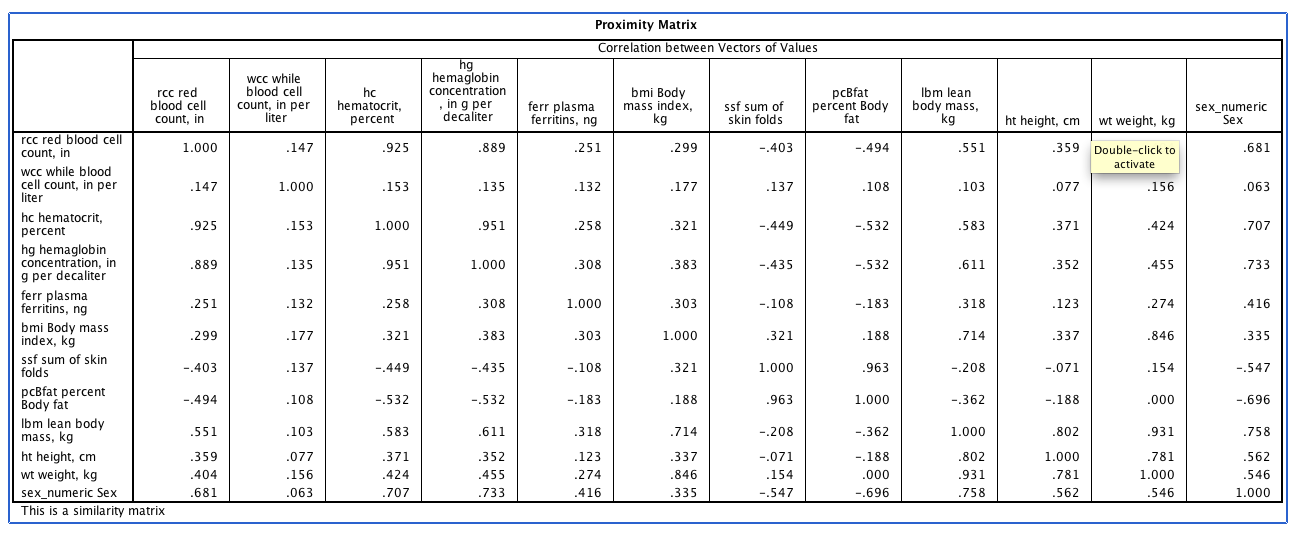


From this we can see that for some sports we have just one gender.

## (c) examine correlations between numeric variables.

You can use analyze - correlate -bivariate for this.

An alternative way that just focuses on the correlations is: analyze - correlate distances (between variables; similarities measure = pearson correlation); note that it can be useful to also include binary variables in such correlation matrices, but not unordered categorical variables with 3 or more levels.



Remember that the important thing to focus on a this point is the pattern of the results. We know with 200 cases, that the correlations will be reasonably accurate representations of population correlations (give or take .1 or .2).

So for example, we might note a few things:

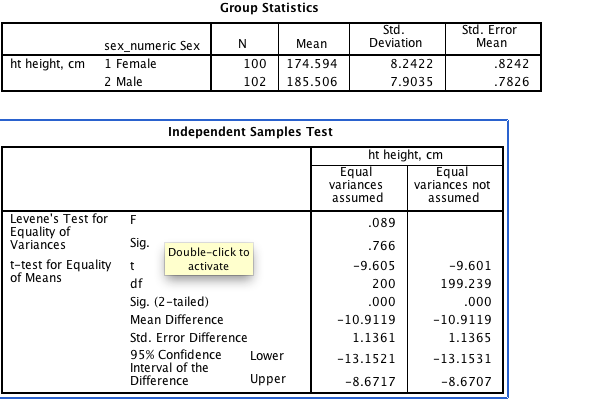
* Some of the blood variables are highly intercorrelated (e.g., red blood cell count, hemagloben concentration, and hematocrit percent).
* Lean body mass and BMI are highly correlated (r=.714). In contrast BMI only weakly correlates with body fat percentage. BMI is often used as a proxy for body fat percentage. Thus this presumably reflects the fact that athletes can often have large quantities of muscle. There are also some fairly high correlations between height, weight, gender, and some of the blood measures.5. (a) Calculate your own version of BMI and check that it agrees with that which is already http://en.wikipedia.org/wiki/Body\_mass\_index; (b) create a categorical variable where 1 = very severely underweight; 2 = severely underweight; (c) Produce a table of sport by BMI category

## 6. (a) Run a t-test with gender as IV; height as DV (see analyze - compare means - independent);

These analyses are intended to show the relationship between t-tests, ANOVA, and regression.

Go to analyze - compare means indpendent samples t-test

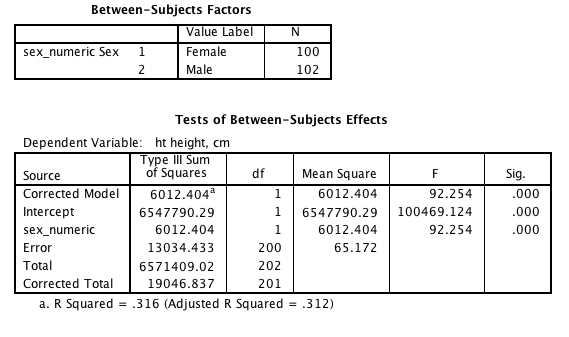
Test variable is "height"; group is "sex\_numeric" (or whatever you called your numeric sex variable).



I.e., mean are 10.9cm taller on average. The difference is statistically significant.

## (b) Run ANOVA with gender as IV, height as DV

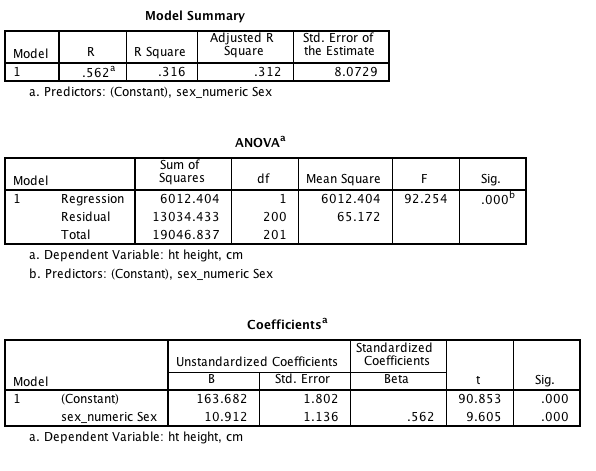
Analyze - GLM - Univariate; gender as fixed factor; height as DV



Here's the basic output, although we could have asked for additional output such as descriptive statistics and so on.

## (c) Run a regression with gender as IV; height as DV

Analyze - Regression- Linear; sex\_numeric as predictor, height as DV



## (d) Compare and contrast the output.

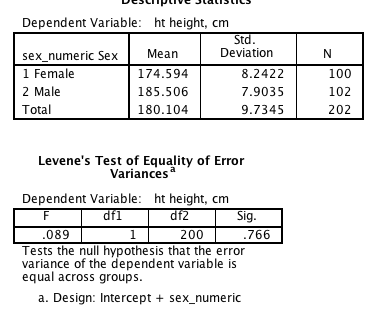
* Although it's a little difficult to see, the p-values are the same across all analyses.
* The t-value of 9.605 can be seen in both the t-test and the regression.
* 200 error or residual df can be seen in all three analyses
* As a side point, the square root of the f-value with 1 numerator degrees of freedom is the t-value (i.e., sqrt(92.254) = 9.605
* The regression coefficient for sex of 10.912 is the same as the difference between group means
* The r-squared is supplied in both the ANOVA and the regression output and is the same.

The main point is that at a deeper level, ANOVA, t-tests, and regression share an underlying linear model.

## (e) examine normality of residuals and homogeneity of variance.

There are many ways of getting this information depending on which SPSS procedure you use.

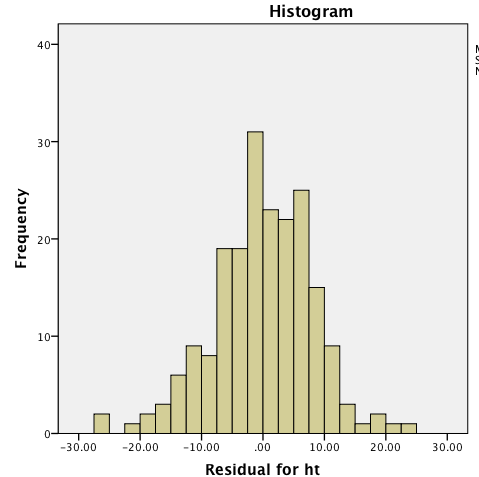
Using "GLM - Univariate", options - descriptive statistics and homogeneity tests will give information needed for assessing homogeneity of variance.



For example, here we see that the standard deviations (i.e., the square root of the sample error variances) are pretty similar. And Levene's test is not significant; thus, no significant difference in group variances.

To examine the distribution of residuals you can click on the "save" button in either the regression of GLM dialog boxes. Select "residuals - unstandardized". Residuals is data minus prediction.

This will add a new variable to your data file, which you can then ask for a histogram on. E.g., analyze - descriptives - frequencies - request histogram.



The plot suggests that the residuals are fairly normally distributed perhaps with a couple of outliers.

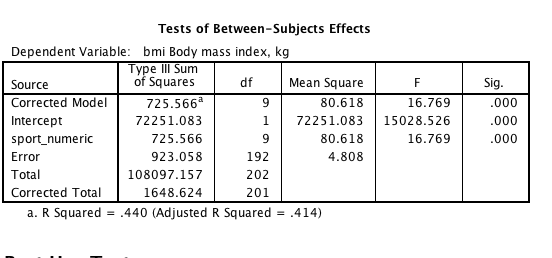
As a side point, all these techniques are relatively robust to modest violations of these assumptions. Examination of assumptions is typically interesting just from the perspective of understanding your data.

## 7. (a) Run an ANOVA with IV=sport; DV=BMI obtain post-hoc tests.

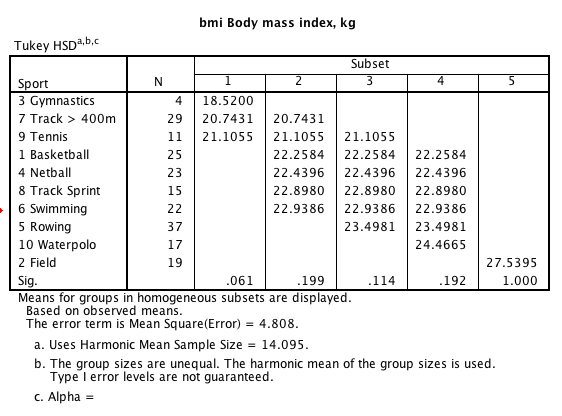
You could use Analyze - GLM - Univariate (fixed factor is sport, and DV is BMI); options: descriptive statistics (for means and SDS) and homogeneity tests for homogeneity of variance) ; post hoc (choose a test; i'll go with Tukey's HSD)

## (b) Interpret the results.

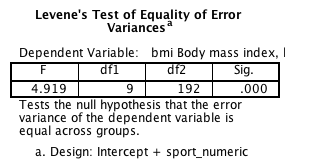
There's clearly an overall effect of sport F (9, 192) = 16.769, p < .001.



The following table gives a good sense of pairwise differences. Numbers that appear in the same column do not have significantly different means. So for example, field athletes had BMI larger than all other athletes. Gymnasts had significantly the lower BMI than most other sports (except long distance and tennis), etc.

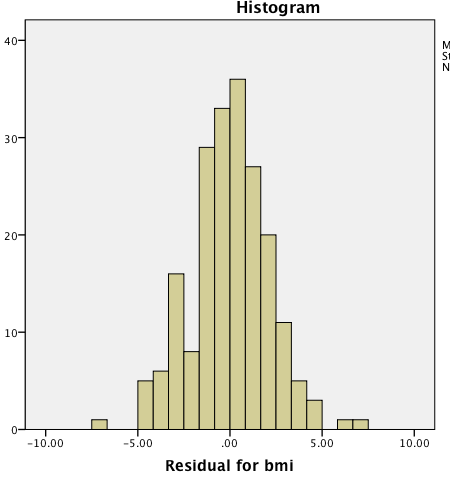


## (c) examine normality of residuals and homogeneity of variance

****

Homogeneity of variance was violated.

You could also save residuals in GLM - univariate and then plot the histogram of the residuals as before. They look fairly normally distributed.

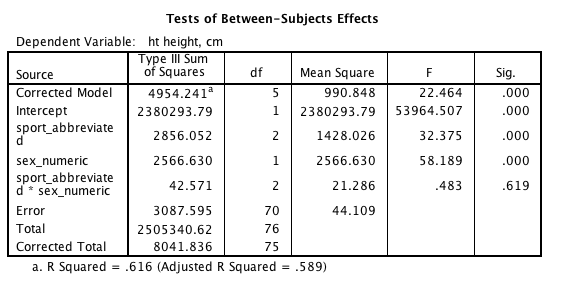


## 8. Run an ANOVA with IV=reduced\_sport (i.e., swim, T400, and basketball) and gender; DV=height. Produce a plot. Provide an overall interpretation of main effects and the interaction effect. Follow-up where appropriate with analysis of simple effects, post hoc tests, etc.

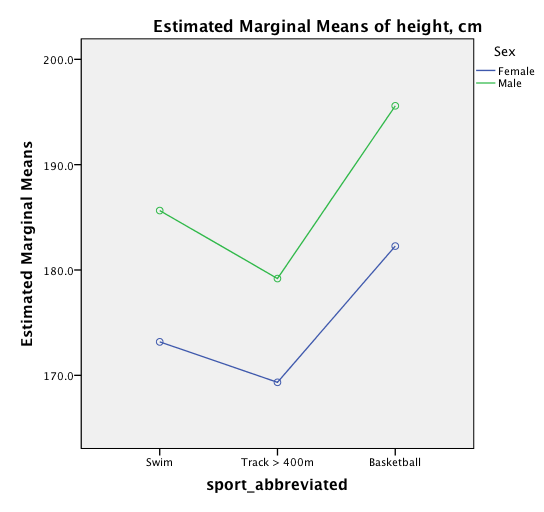
Run Analyze - GLM - Univariate; Fixed factors = sex and "sport abbreviated"; DV is height

Press "Plots..." separate lines = sex; horizontal axis = sport

options: descriptive statistics and homogeneity of variance

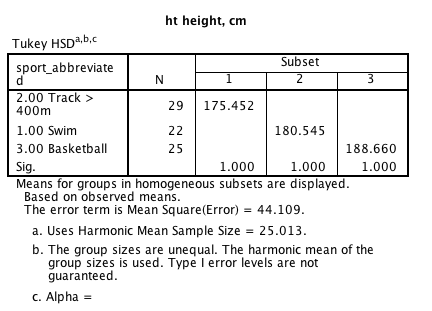


The significance table suggests there is a main effect of sport (p < .001) and sex (p <.001).



The plot elaborates on and explains the main effects. Not surprisingly, males are taller than females. The descriptive statistics suggests that basketballers are taller than swimmers who are in turn smaller than longer distance track runners. That said, we'd typically use post-hoc tests to determine whether each of these pairwise comparisons is significant. The graph also shows that roughly the lines are parallel, indicative of an absence of an interaction effect.

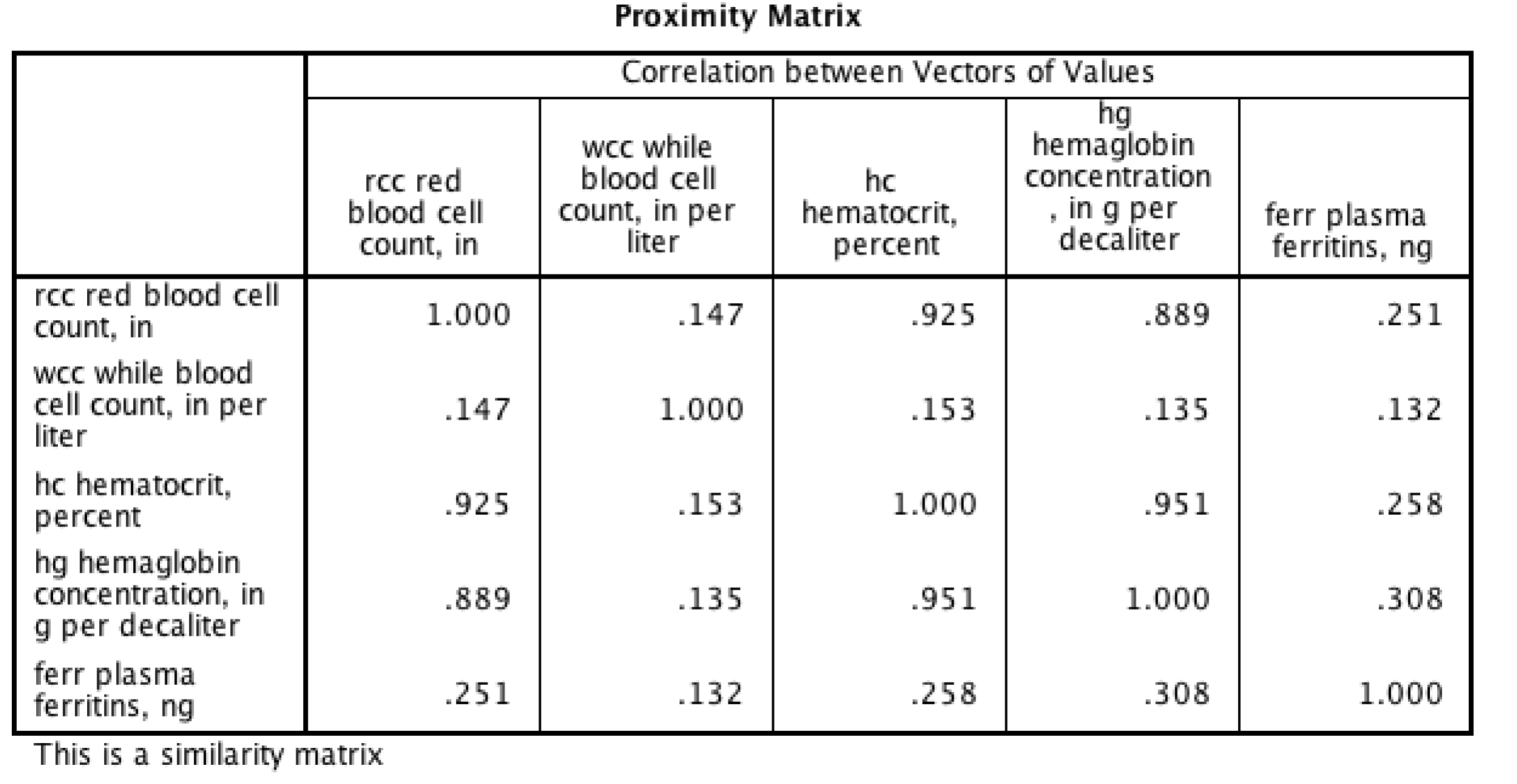
Given there is no interaction effect, there is no need for analysis of simple effects. The main effect for gender is clear because there are only two levels. Thus, the main additional analysis is a post-hoc test on sport. Go back to "analysze - GLM - univariate" and click "post hoc" and add sport and perhaps "Tukey's HSD".



Tukey's HSD suggests that all three sports are significantly different from each other in terms of mean height. (p<.05).

## 9. MANOVA: (a) Review the correlations between all the five blood variables (i.e., rcc, wcc, hc, hg, ferr);

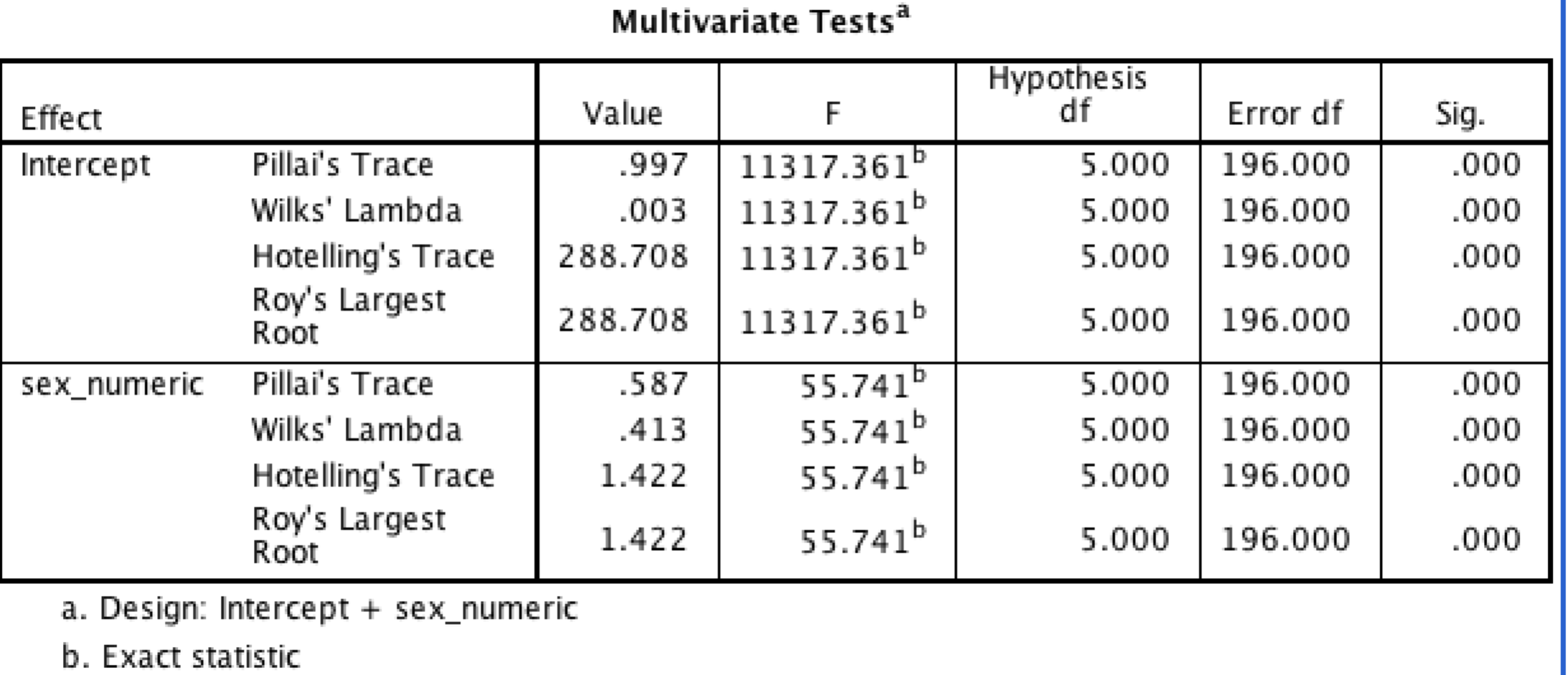
You could do analyze correlate bivariate or use analyze - correlate - distances with additional settings.



The correlations are all positive. hg, hc, and rcc are particularly highly correlated. The other two have smaller intercorrelations with other variables. I don't know enough about the theory of blood related variables to know whether forming a composite makes sense, but the correlations are at least vaguely supportive of such an idea, at least for the purposes of this exercise.

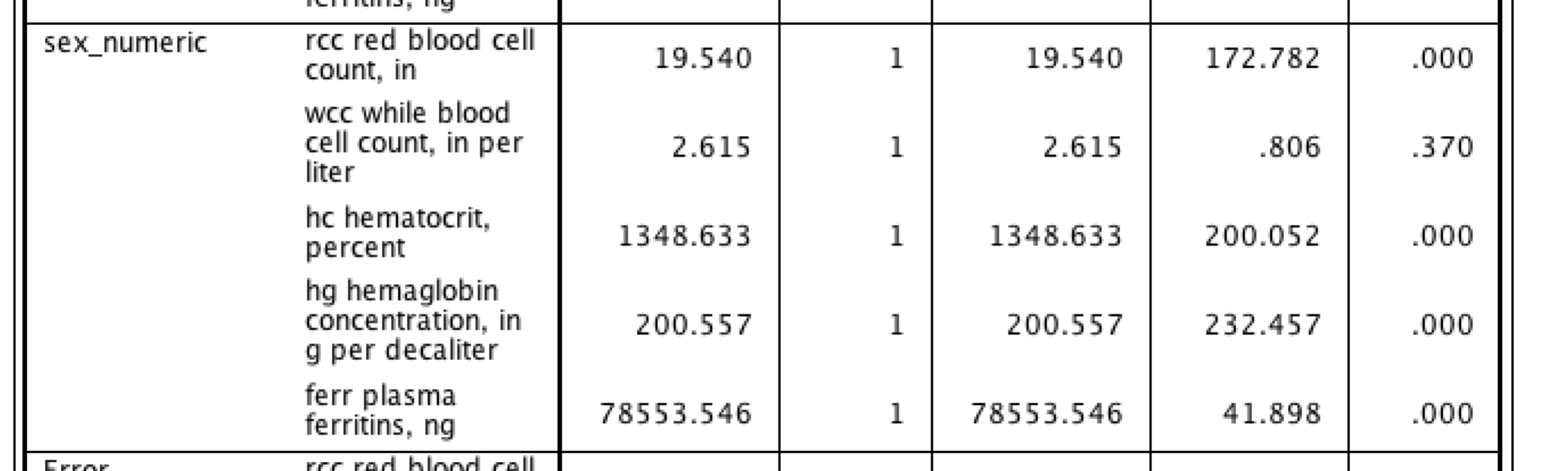
## (b) Run a MANOVA (Analyze - GLM - Multivariate) with gender as IV (i.e., fixed factor) and the five blood variables as DV and interpret the multivariate effects

Go "Analyze - GLM - Multivariate with gender as fixed factor and the five blood variables as DVs. You might also want to ask for descriptive statistics, homogeneity tests and anything else of interest.

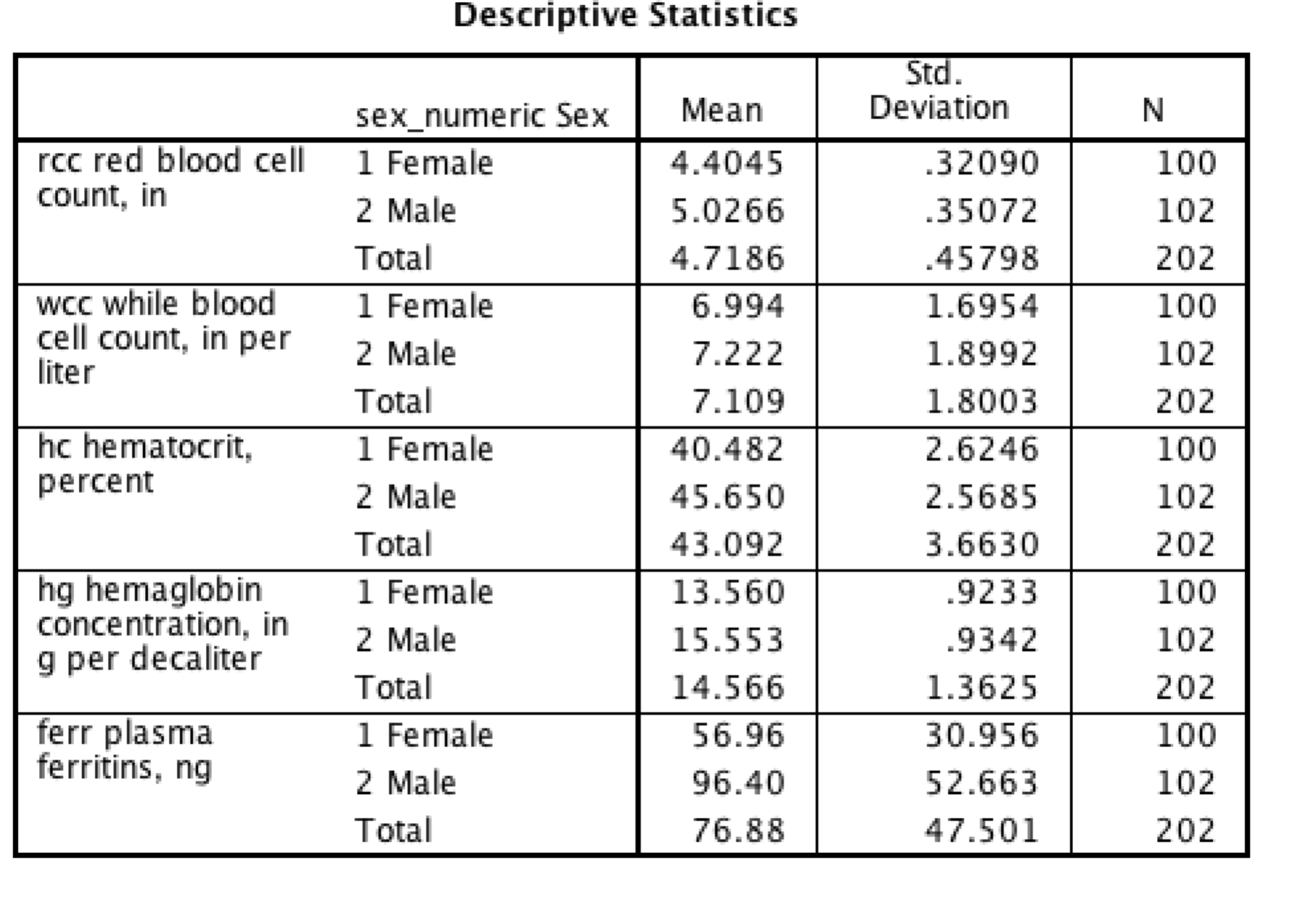


All the ways of computing the multivariate effects are significant (note that with only two levels to the IV, this is always the case). Thus, we know that there are differences between groups in some composite of the DV.

## (c) interpret the univariate effects from the MANOVA output;



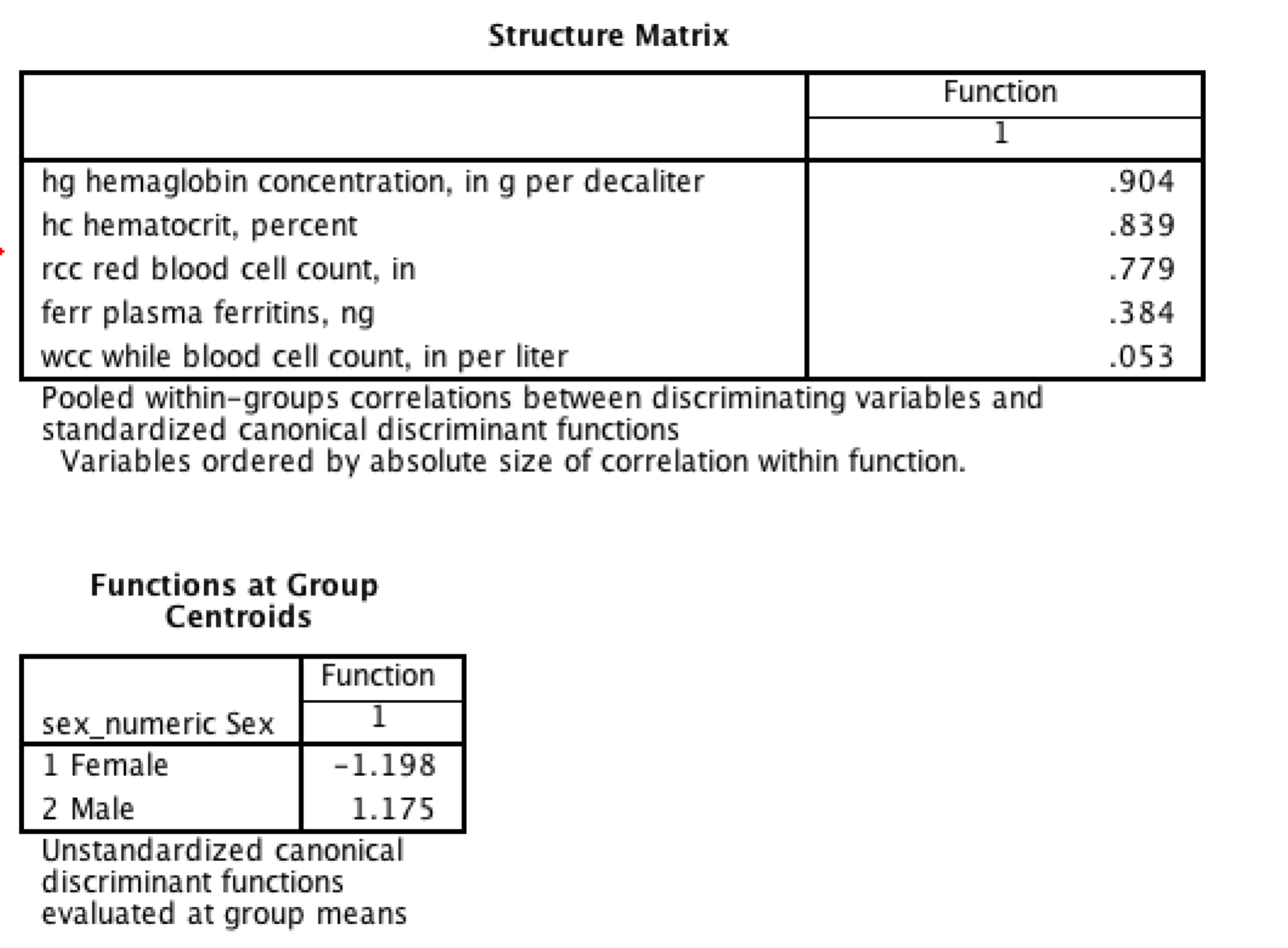
Here is an extract of the relevant components of the significance tests for the univariate tests. It suggests that all but wcc is statistically significant (p<.05)



The following descriptive statistics make this a little clearer. For example, men have significantly higher red blood cell count, hematocric percentage, hemaglobin concentration and plasma ferritins. You could do some quick Cohen's d calculations to interpret these effects (or if you understand the variables intuitively based on their scales, you could use them). For example, Men are on average approximately 2 SD higher on red blood cell count. By examining the pattern of Cohen's d's this should give you a feel for the group differences. It's not directly addressing the question of what is the multivariate composite created by the MANOVA, but it should give you a good feel for the group differences.

## (d) run a discriminant function analysis with gender as DV and blood variables as IVs (Analyze - Classify - discriminant) and interpret the results

Go "Analyze - Classify - Discriminant" and put the DV as sex (coded 1, 2) and the five blood variables as predictors.



The structure matrix shows the correlations between each variable and the discriminant function (which is the same as the multiviate composite implied by the MANOVA). Thus, for example, hg, hc, and rcc are the main variables positively correlating with the discriminant function. ferr is slightly related and wcc is relatively unrelated.

Thus, we now know what it means to be high on the discriminant function: you have high hemaglobin counts, red blood cell counts, etc.

Thus, which group is high and which group is low on this composite?

The functions at group centroids indicates this. Specifically, males are higher and females are lower.

## (e) compare and contrast the interpretation of the group differences from the univariate versus the DFA approach

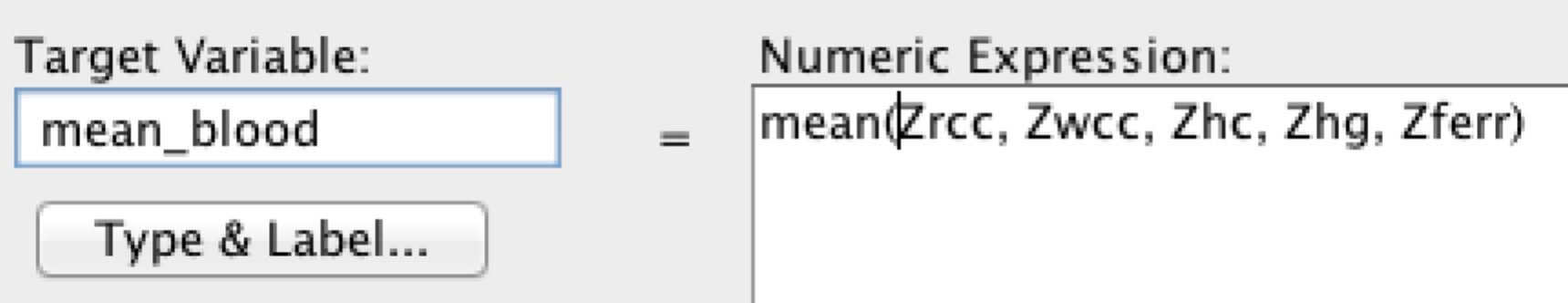
This analysis has thus provided with an understanding of the discriminant function. If we compare to the analysis of univariate effects, and the previous correlation analysis, there is a degree of consistency.

## (f) create a single blood variable by converting each of the five blood variables into z-scores and then summing them up (check that they are all positively correlated to ensure that none of them need to be reversed), then run a univariate ANOVA with this composite variable as a DV; interpret the results

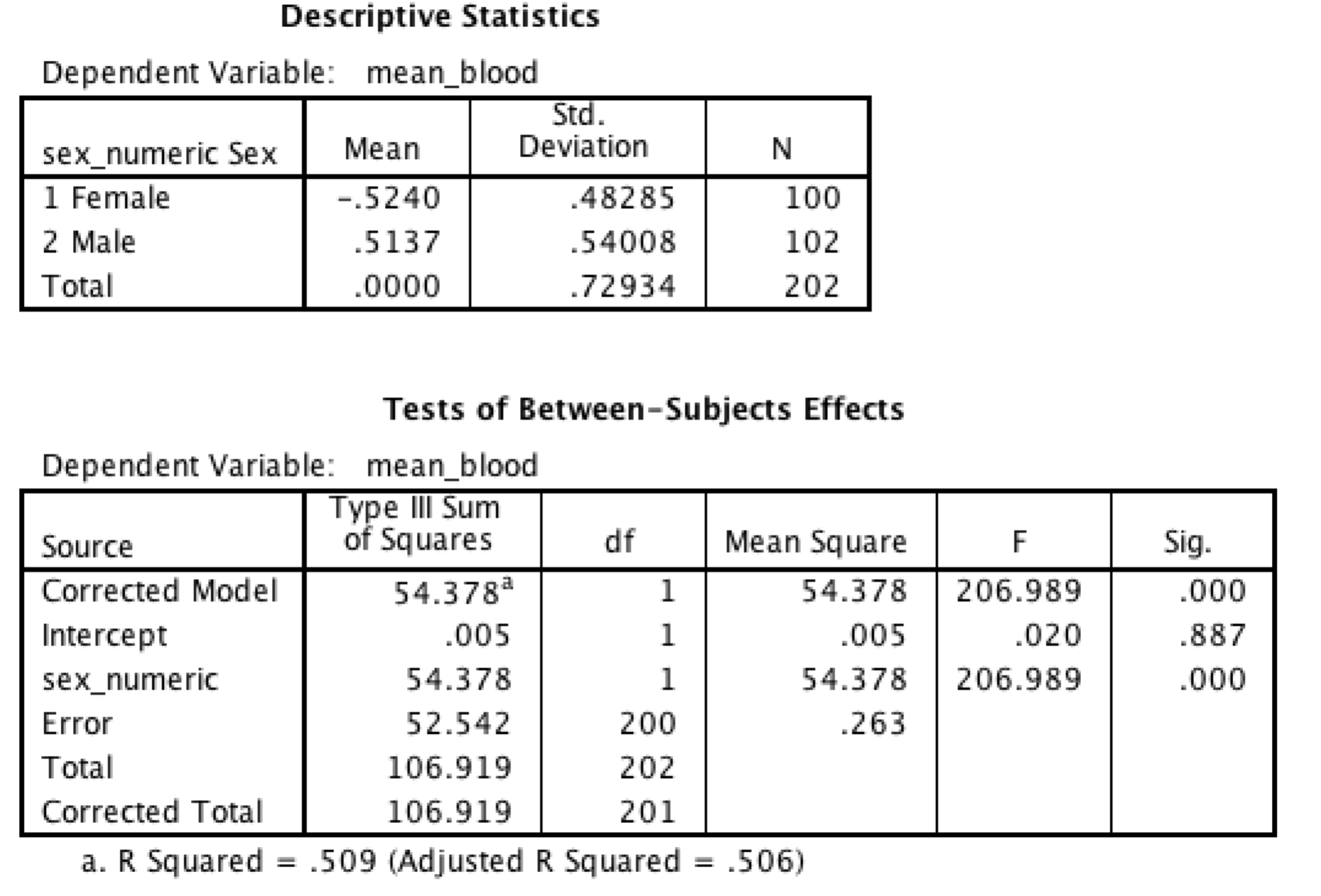
Go to Analyze - descriptives - descriptives and put in the blood variables and click "save as standardized variable".

This should create five new z-score versions of the blood variables.

Then use "transform - compute" to create a function like:



Then run "GLM - univariate" with the new blood variable as DV and gender as fixed factor. Perhaps add descriptive statistics and homogeneity test options.



This analysis shows the significant difference. It reiterates the difference of males being around two standard deviations higher than females.

## (g) compare and contrast creating your own composite with univariate ANOVA with the multivariate MANOVA output.

* When you create your own composite based on z-scores, each dependent variable is weighted equally. We know this because we have specified the composite. In contrast, in MANOVA, the composite is created implicitly so as to maximise group differences. Thus, creating our own composite is simpler to interpret. However, we would have to rely on theoretical judgement or the intercorrelations of the dependent variable to justify the creation of the composite.
* Creating our own composite means that the results are more comparable across studies, because the form of the composite is constant across studies.
* In the present example, the multivariate composite highlighted by the DFA was different to that when we created our own composite. In particular, some variables were weighted more and some less.