5. Correlation

Objectives

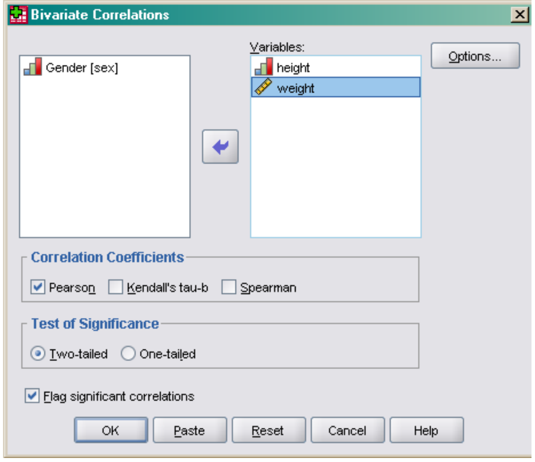
- Calculate correlations
- Calculate correlations for subgroups using split file
- Create scatterplots with lines of best fit for subgroups and multiple correlations

Correlation

The first inferential statistic we will focus on is correlation. As noted in the text, correlation is used to test the degree of association between variables. All of the inferential statistics commands in SPSS are accessed from the Analyze menu. Let's open SPSS and replicate the correlation between height and weight presented in the text.

✓ **Open** *HeightWeight.sav*. Take a moment to review the data file.

✓ Under **Analyze**, select **Correlate/Bivariate**. Bivariate means we are examining the simple association between 2 variables.



Bivariate Correlations

Gender [sex]

Variables: height, weight

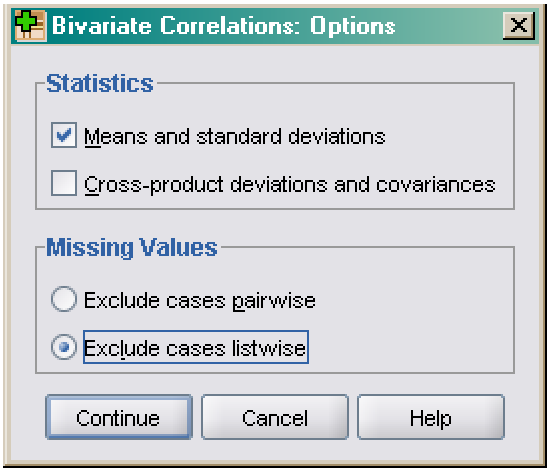
Correlation Coefficients: ☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

✓ In the dialog box, select height and weight for **Variables**. Select **Pearson** for **Correlation Coefficients** since the data are continuous. The default for **Tests of Significance** is **Two-tailed**. You could change it to One-tailed if you have a directional hypothesis. Selecting **Flag significant correlations** means that the significant correlations will be noted in the output by asterisks. This is a nice feature. Then click **Options**.

✔ Now you can see how descriptive statistics are built into other menus. Select **Means and standard deviations** under **Statistics**. Missing Values are important. In large data sets, pieces of data are often missing for some variables.



Bivariate Correlations: Options

Statistics

☑ Means and standard deviations

□ Cross-product deviations and covariances

Missing Values

○ Exclude cases pairwise

● Exclude cases listwise

For example I may run correlations between height, weight, and blood pressure. One subject may be missing blood pressure data. If I check **Exclude cases listwise**, SPSS will not include that person's data in the correlation between height and weight, even though those data are not missing. If I check **Exclude cases pairwise,** SPSS will include that person's data to calculate any correlations that do not involved blood pressure. In this case, the person's data would still be reflected in the correlation between height and weight. You have to decide whether or not you want to exclude cases that are missing any data from all analyses. (Normally it is much safer to go with listwise deletion, even though it will reduce your sample size.) In this case, it doesn't matter because there are no missing data. Click **Continue**. When you return to the previous dialog box, click **Ok**. The output follow.

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
HEIGHT	68.72	3.66	92
WEIGHT	145.15	23.74	92

Correlations

		HEIGHT	WEIGHT
HEIGHT	Pearson Correlation	1.000	.785**
	Sig. (2-tailed)	.	.000
	N	92	92
WEIGHT	Pearson Correlation	.785**	1.000
	Sig. (2-tailed)	.000	.
	N	92	92

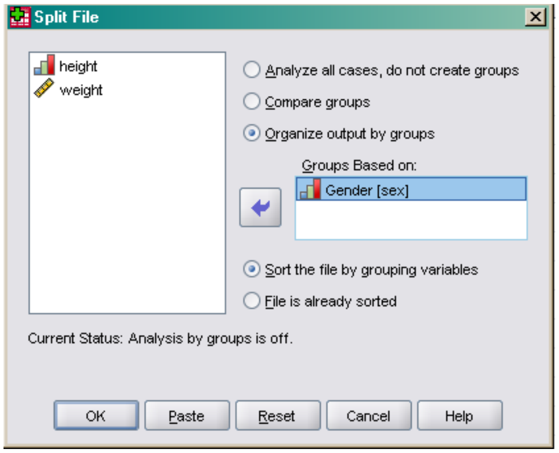
**. Correlation is significant at the 0.01 level

Notice, the correlation coefficient is .785 and is statistically significant, just as reported in the text. In the text, Howell made the point that heterogeneous samples affect correlation coefficients. In this example, we included both males and females. Let's examine the correlation separately for males and females as was done in the text.

Subgroup Correlations

We need to get SPSS to calculate the correlation between height and weight separately for males and females. The easiest way to do this is to split our data file by sex. Let’s try this together.

✓ In the Data Editor window, select **Data/Split file**.



Split File

height
weight

○ Analyze all cases, do not create groups
○ Compare groups
● Organize output by groups

Groups Based on:
Gender [sex]

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

✓ Select **Organize output by groups** and **Groups Based on** Gender. This means that any analyses you specify will be run separately for males and females. Then, click **Ok**.

✓ Notice that the order of the data file has been changed. It is now sorted by Gender, with males at the top of the file.

✓ Now, select **Analyze/Correlation/Bivariate**. The same variables and options you selected last time are still in the dialog box. Take a moment to check to see for yourself. Then, click **Ok**. The output follow broken down by males and females.

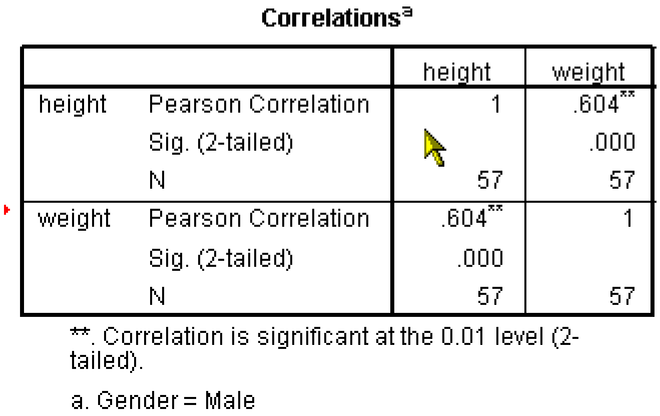
Correlations

SEX = Male

Descriptive Statistics[a]

	Mean	Std. Deviation	N
HEIGHT	70.75	2.58	57
WEIGHT	158.26	18.64	57

a. SEX = Male



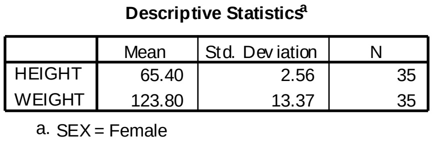
Correlations[a]

		height	weight
height	Pearson Correlation	1	.604**
	Sig. (2-tailed)		.000
	N	57	57
weight	Pearson Correlation	.604**	1
	Sig. (2-tailed)	.000	
	N	57	57

**. Correlation is significant at the 0.01 level (2-tailed).

a. Gender = Male

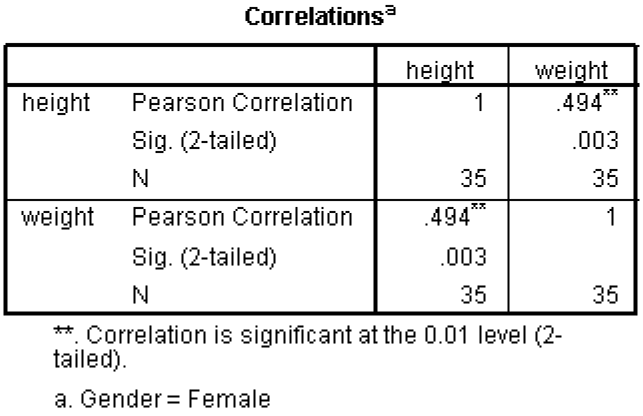
SEX = Female



Descriptive Statistics[a]

	Mean	Std. Deviation	N
HEIGHT	65.40	2.56	35
WEIGHT	123.80	13.37	35

a. SEX = Female



Correlations[a]

		height	weight
height	Pearson Correlation	1	.494**
	Sig. (2-tailed)		.003
	N	35	35
weight	Pearson Correlation	.494**	1
	Sig. (2-tailed)	.003	
	N	35	35

**. Correlation is significant at the 0.01 level (2-tailed).

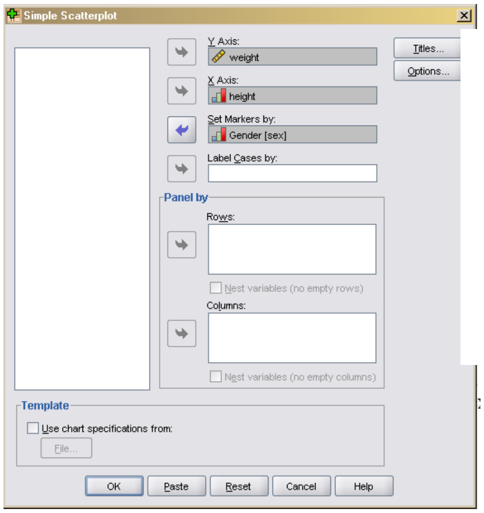
a. Gender = Female

As before, our results replicate those in the text. The correlation between height and weight is stronger for males than females. Now let’s see if we can create a more complicated scatterplot that illustrates the pattern of correlation for males and females on one graph. First, we need to turn off split file.

✓ Select **Data/Split file** from the Data Editor window. Then select **Analyze all cases, do not compare groups** and click **Ok**. Now, we can proceed.

Scatterplots of Data by Subgroups

✓ Select **Graphs/Legacy/Scatter**. Then, select **Simple** and click **Define**.


Simple Scatterplot
Titles...
Options...
Panel by
Rows:
□ Nest variables (no empty rows)
Columns:
□ Nest variables (no empty columns)
Template
□ Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

Y Axis: weight

X Axis: height

Set Markers by: Gender [sex]

Label Cases by:

✓ To be consistent with the graph in the text book, select weight as the **Y Axis** and height as the **X Axis**. Then, select sex for **Set Markers by**. This means SPSS will distinguish the males dots from the female dots on the graph. Then, click **Ok**.

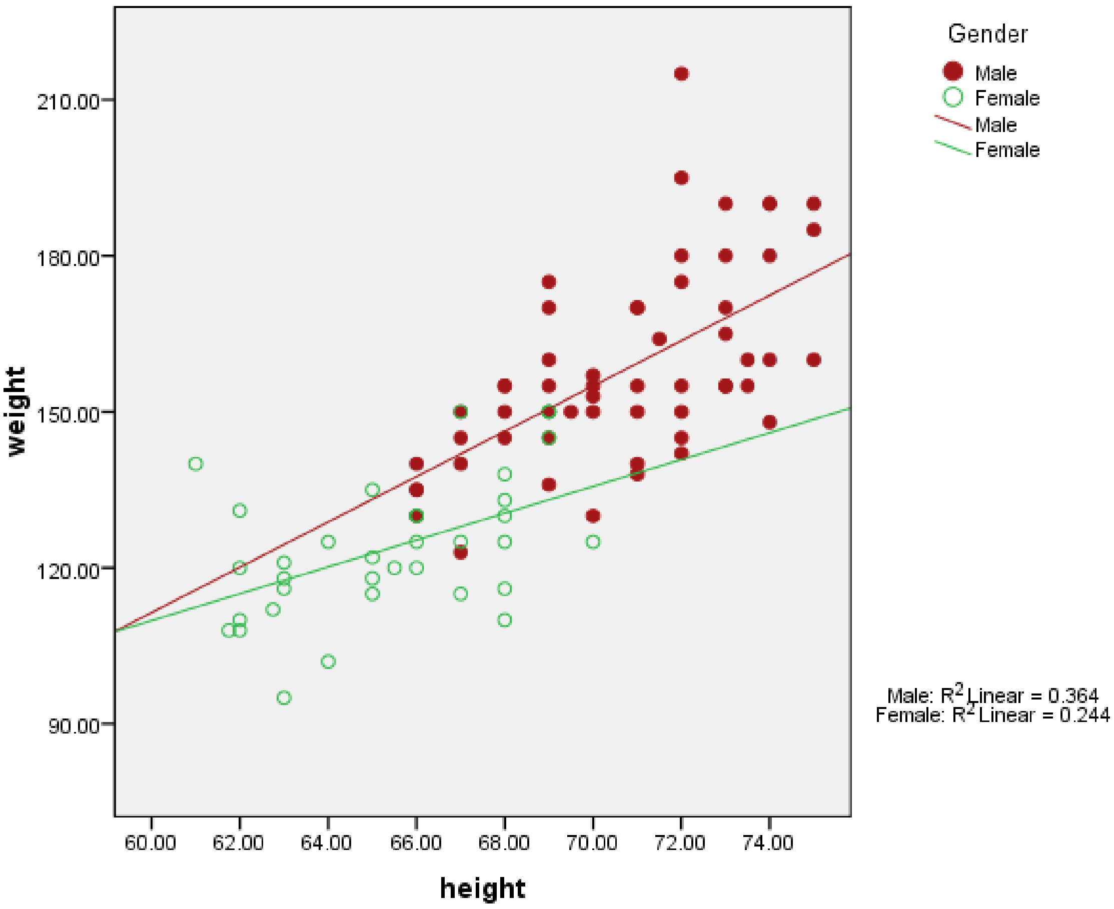
When your graph appears, you will see that the only way males and females are distinct from one another is by color. This distinction may not show up well, so let's edit the graph.

✓ Double click the graph to activate the Chart Editor. Then double click on one of the female dots on the plot. SPSS will highlight them. (I often have trouble with this. If it selects all the points, click again on a female one. That should do it.) Then click the Marker menu.

✓ Select the circle under **Marker Type** and chose a **Fill** color. Then click **Apply**. Then click on the male dots, and select the open circle in **Marker Type** and click **Apply**. Then, close the dialog box. The resulting graph should look just like the one in the textbook.

I would like to alter our graph to include the line of best fit for both groups.

✓ Under **Elements, select Fit Line** at **Subgroups**. Then select **Linear** and click **Continue**. (I had to select something else and then go back to Linear to highlight the **Apply** button.) The resulting graph follows. I think it looks pretty good.



- ✓ Edit the graph to suit your style as you learned in Chapter 3 (e.g., add a title, change the axes titles and legend).

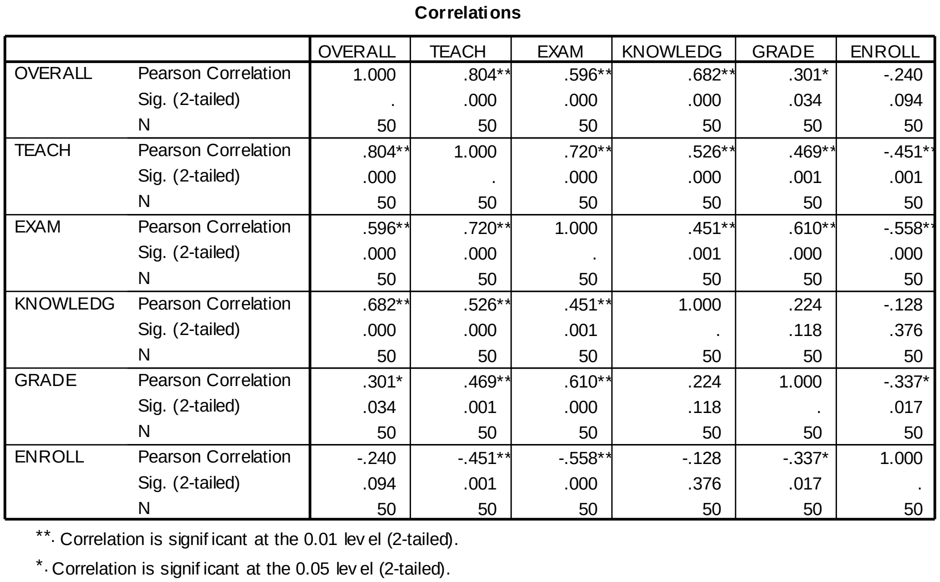
This more complex scatterplot nicely illustrates the difference in the correlation between height and weight for males and females. Let's move on to a more complicated example.

Overlay Scatterplots

Another kind of scatterplot that might be useful is one that displays the association between different independent variables with the same dependant variable. Above, we compared the same correlation for different groups. This time, we want to compare different correlations. Let's use the course evaluation example from the text . It looks like expected grade is more strongly related to ratings of fairness of the exam than ratings of instructor knowledge is related to the exam. I'd like to plot both correlations. I can reasonably plot them on the same graph since all of the questions were rated on the same scale.

- ✓ **Open** *courseevaluation.sav*. You do not need to save *HeightWeight.sav* since you did not change it. So click **No**.

✓ First, let’s make sure the correlations reported in the text are accurate. Click **Analyze/Correlation/Bivariate** and select all of the variables. Click **Ok**. The output follow. Do they agree with the text?



Correlations

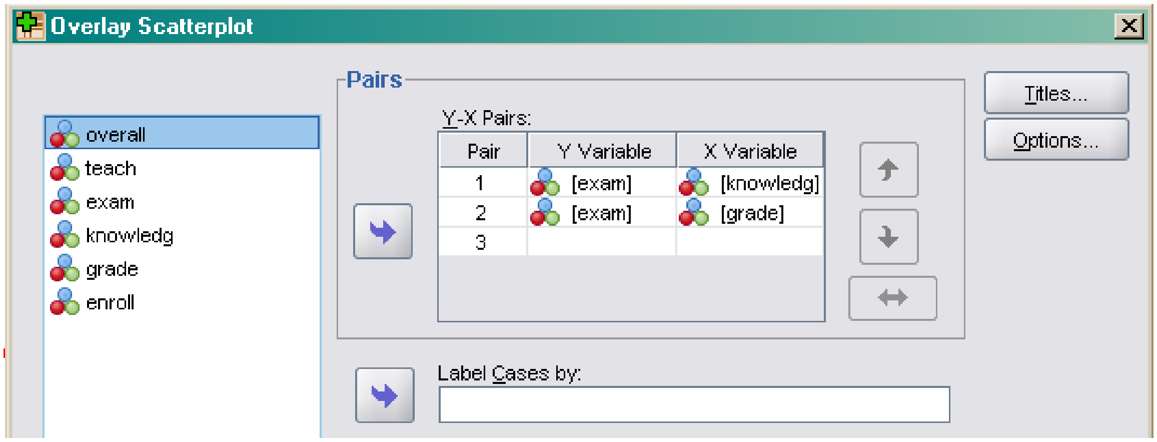
		OVERALL	TEACH	EXAM	KNOWLEDG	GRADE	ENROLL
OVERALL	Pearson Correlation	1.000	.804**	.596**	.682**	.301*	-.240
	Sig. (2-tailed)	.	.000	.000	.000	.034	.094
	N	50	50	50	50	50	50
TEACH	Pearson Correlation	.804**	1.000	.720**	.526**	.469**	-.451**
	Sig. (2-tailed)	.000	.	.000	.000	.001	.001
	N	50	50	50	50	50	50
EXAM	Pearson Correlation	.596**	.720**	1.000	.451**	.610**	-.558**
	Sig. (2-tailed)	.000	.000	.	.001	.000	.000
	N	50	50	50	50	50	50
KNOWLEDG	Pearson Correlation	.682**	.526**	.451**	1.000	.224	-.128
	Sig. (2-tailed)	.000	.000	.001	.	.118	.376
	N	50	50	50	50	50	50
GRADE	Pearson Correlation	.301*	.469**	.610**	.224	1.000	-.337*
	Sig. (2-tailed)	.034	.001	.000	.118	.	.017
	N	50	50	50	50	50	50
ENROLL	Pearson Correlation	-.240	-.451**	-.558**	-.128	-.337*	1.000
	Sig. (2-tailed)	.094	.001	.000	.376	.017	.
	N	50	50	50	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Now, let’s make our scatterplot.

✓ Select **Graphs/Legacy/Scatter**. Then select **Overlay** and click **Define**.

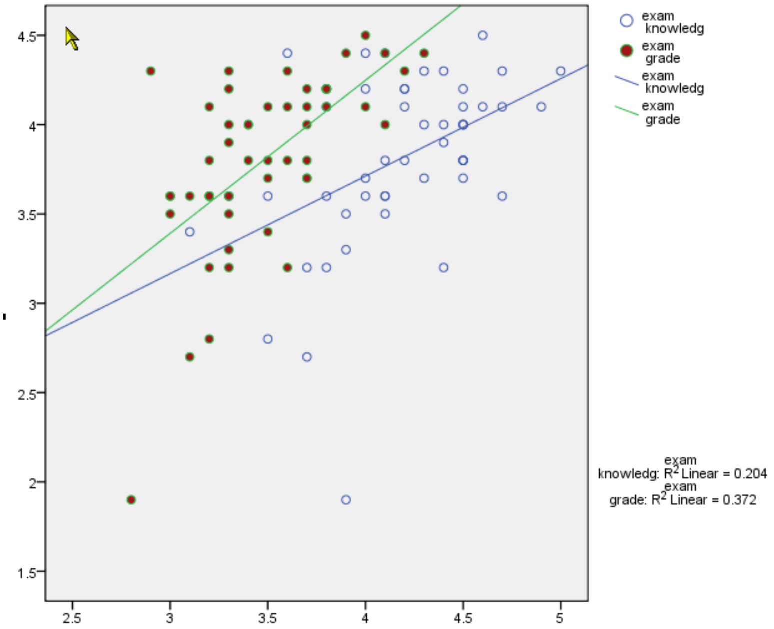


Y-X Pairs:

Pair	Y Variable	X Variable
1	[exam]	[knowledg]
2	[exam]	[grade]
3		

✓ Click on exam and grade and shift them into **Y-X Pairs**. Then click on exam and knowledge and click them into **Y-X pairs**. Since exam is the commonality between both pairs, I’d like it to be on the Y axis. If it is not listed as Y, highlight the pair and click on the two-headed arrow. It will reverse the ordering. Exam should then appear first for both. Then, click **Ok**.

✔ As in the previous example, the dots are distinguished by color. Double click the graph and use the **Marker** icon to make them more distinct as you learned above. Also use the **Elements** menu to **Fit line at total.** It will draw a line for each set of data.



Note that the axes are not labeled. You could label the Y Axis Grade. But you could not label the X axis because it represents two different variables-exam and knowledge. That is why the legend is necessary. (If you figure out how to label that axis, please let me know. It should be so easy.)

As you can see, the association between expected grade and fairness of the exam is stronger than the correlation between instructor's knowledge and the fairness of the exam.

Now, you should have the tools necessary to calculate Person Correlations and to create various scatterplots that compliment those correlations. Complete the following exercises to help you internalize these steps.

Exercises

Exercises 1 through 3 are based on *appendixd.sav*.

1. Calculate the correlations between Add symptoms, IQ, GPA, and English grade twice, once using a one-tailed test and once using a two-tailed test. Does this make a difference? Typically, when would this make a difference.

2. Calculate the same correlations separately for those who did and did not drop out, using a two-tailed test. Are they similar or different?

3. Create a scatterplot illustrating the correlation between IQ score and GPA for those who did and did not drop out. Be sure to include the line of best fit for each group.

4. Open *courseevaluation.sav*. Create a scatterplot for fairness of exams and teacher skills and exam and instructor knowledge on one graph. Be sure to include the lines of best fit. Describe your graph.