Correlation

January 9, 2021

Contents

Happy birthday to Ruiting Deng and Mia Joy Roa!

Correlations

People tend to be attracted to those with similar heights. Let's see if in our survey, there is such a relationship between the mothers and fathers of students in this class. To keep the sample size small, let's just focus on students that chose Green as their favorite color. Here's a table the heights (in inches) of the mothers and fathers of the 24 students:

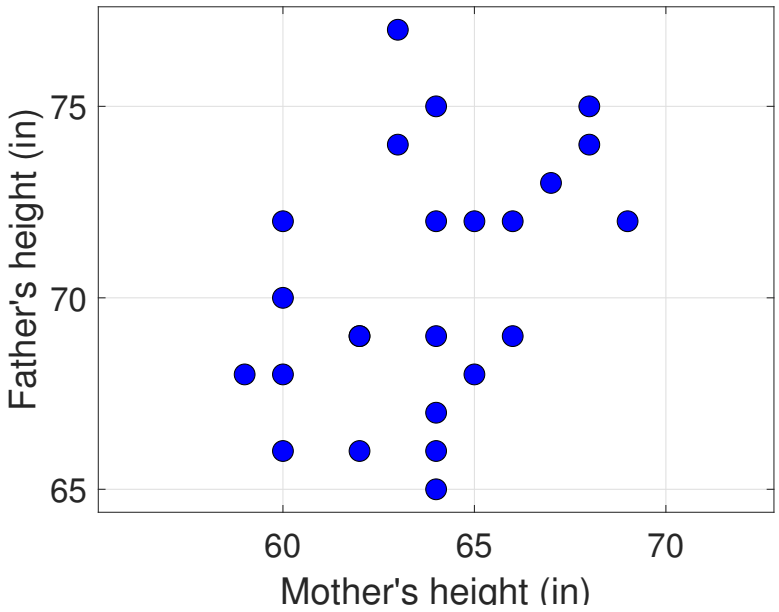
Parent's height (in)	
Mothers	Fathers
60	72
63	74
63	77
64	75
64	72
65	72
68	75
68	74
67	73
66	72
69	72
66	69
65	68
64	69
64	67
64	66
64	65
62	66
60	66
62	69
62	69
60	68
59	68
60	70

You can download the Excel file containing the data for this tutorial here: parentheight-Green.csv

Can you see a relation in this data? See if the tallest mothers are paired with taller than average fathers, and/or vice versa.

The Scattterplot

A natural way to graph paired data like this is with a scatterplot, where the two variables are plotted on the x and y-axes. We'll arbitrarily choose Mothers heights for the x-axis and Fathers heights for the y-axis:

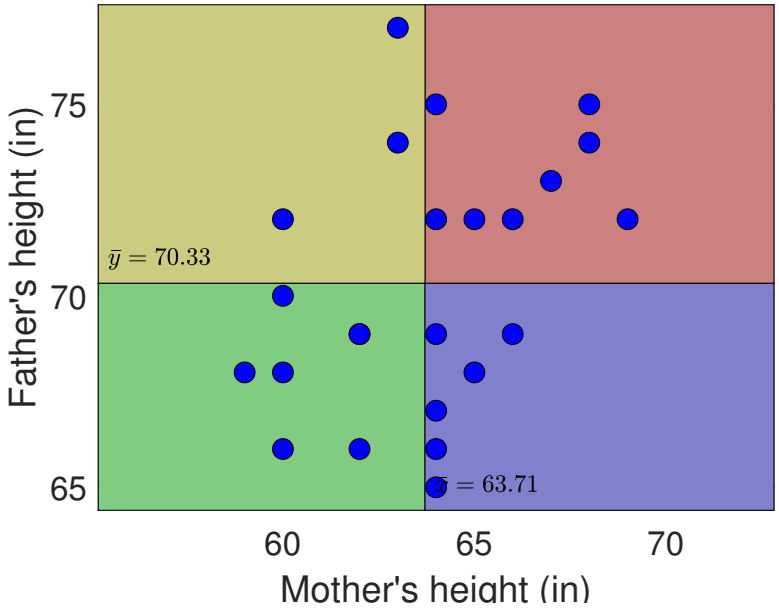


The Pearson correlation

You can see in the scatterplot that there is a trend for taller fathers to pair with taller mothers. To quantify this trend, we'll introduce a statistic called the **correlation**. There different kinds of correlations, but by far the most common is the 'Pearson Product Moment Correlation', which is a number, ranging from -1 to 1, that reflects how well points on a scatter plot are fit by a straight line. This long name is usually shortened to 'Pearson correlation', or even just 'correlation', and is represented by the letter r.

The intuition behind the Pearson correlation is that points with above average values on the x-axis should be paired with above-average values on the y-axis, and points below average on the x-axis should be paired with points below average on the y-axis.

This can be visualized by dividing the scatterplot into four quadrants, where quadrants are split left and right by the mean of the x-values, and split up and down by the mean of y-values. The mean of the 24 Mothers's heights is 63.71 inches, and the mean of the Fathers's heights is 70.33 inches. Here's the same scatterplot, but with the four quadrants drawn in.



You should see that there are more points falling in the red and green (1st and 3rd) quadrants than in the blue and yellow quadrants. This means that there is a **positive** correlation between the heights of Mothers and Fathers.

Values of the Pearson correlation range between -1 and 1. A correlation of 1 means that the points fall exactly on a line with a positive slope. A correlation of -1 fall exactly on a line with negative slope.

You should convince yourself that a scatterplot for points with a correlation of 1 will only have points in the green and red quadrants. Similarly, a correlation of -1 will only have points in the yellow and blue quadrants.

How do we quantify this distribution of scores in the four quadrants? The trick behind the Pearson correlation coeficient is to translate x and y scores into z-scores. Remember, z-scores are calculated by subtracting the mean and dividing by the standard deviation. For the Pearson correlation, we use the population standard deviation (divide by n, and not n-1).

Here's a new table with two new columns representing the z-scores, and a last column that is the product of each pair of z-scores. I've colored the table cells in red, green, blue and yellow depending on which quadrant the pair falls into (the colors are grouped because I sorted the scores so this would happen beforehand).

Mothers	Fathers	z_x	z_y	$(z_x)(z_y)$
60	72	-1.37	0.51	-0.7
63	74	-0.26	1.13	-0.29
63	77	-0.26	2.04	-0.53
64	75	0.11	1.43	0.15
64	72	0.11	0.51	0.05
65	72	0.48	0.51	0.24
68	75	1.58	1.43	2.26
68	74	1.58	1.13	1.78
67	73	1.21	0.82	0.99
66	72	0.84	0.51	0.43
69	72	1.95	0.51	1
66	69	0.84	-0.41	-0.34
65	68	0.48	-0.71	-0.34
64	69	0.11	-0.41	-0.04
64	67	0.11	-1.02	-0.11
64	66	0.11	-1.33	-0.14
64	65	0.11	-1.63	-0.17
62	66	-0.63	-1.33	0.84
60	66	-1.37	-1.33	1.81
62	69	-0.63	-0.41	0.26
62	69	-0.63	-0.41	0.26
60	68	-1.37	-0.71	0.98
59	68	-1.73	-0.71	1.24
60	70	-1.37	-0.1	0.14

What do the scores in the red and green quadrants have in common? If you look at the table you'll see that the product of their z-scores for the red and green scores are all positive. That's because for points in the red quadrant both z-scores are positive, and in the green quadrant both are negative - so their product is also positive. Points in the yellow and blue quadrants will have one positive and one negative z-score, so their product will be negative.

The Pearson correlation is the mean of the product of z-scores. For this example the mean of the products is:

$$r = \frac{\sum (z_x)(z_y)}{n} = \frac{-0.7 - 0.29 - ... + 0.14}{24} = 0.41$$

This mean of product of z-scores should make intuitive sense. If the mean is positive, then the positive products outweigh the negative product of z-scores. That is, the points in the red and green quadrants outweigh the points in the yellow and blue. If the correlation is negative, then the points in the yellow and blue quadrarants outweigh those in the red and green quadrants.

The computational raw-score formula

If you do some algebra and plug in the formula for the z-score, you can show that there are various ways to calculate the Pearson correlation coeficient. The most computationally efficient way is the 'computational raw-score formula':

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum x)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum y)^2}{n}\right)}}$$

If you don't have a stats calculator or computer around (or happen to be taking an exam), you can calculate the components of this formula by setting up columns and adding them up:

x	y	xy	x^2	y^2
60	72	4320	3600	5184
63	74	4662	3969	5476
63	77	4851	3969	5929
64	75	4800	4096	5625
64	72	4608	4096	5184
65	72	4680	4225	5184
68	75	5100	4624	5625
68	74	5032	4624	5476
67	73	4891	4489	5329
66	72	4752	4356	5184
69	72	4968	4761	5184
66	69	4554	4356	4761
65	68	4420	4225	4624
64	69	4416	4096	4761
64	67	4288	4096	4489
64	66	4224	4096	4356
64	65	4160	4096	4225
62	66	4092	3844	4356
60	66	3960	3600	4356
62	69	4278	3844	4761
62	69	4278	3844	4761
60	68	4080	3600	4624
59	68	4012	3481	4624
60	70	4200	3600	4900

$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$
1529	1688	107626	97587	118978

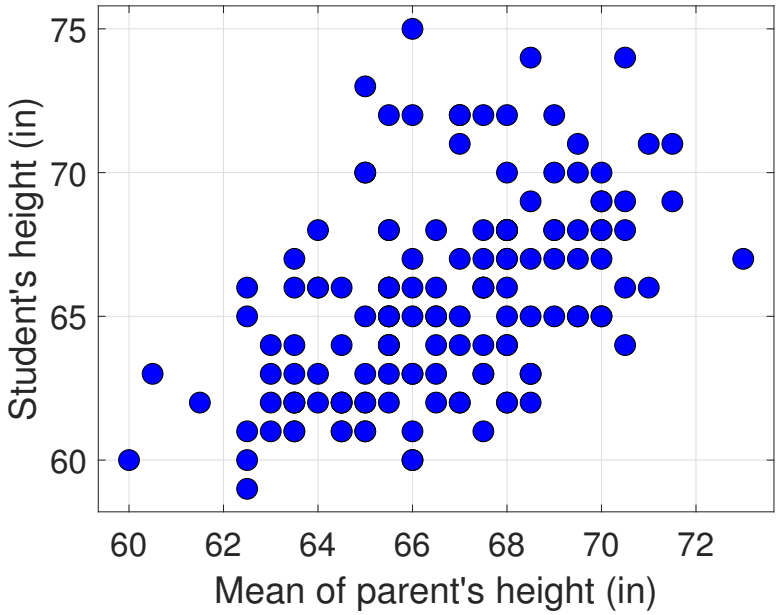
Plugging these sums into the formula gives:

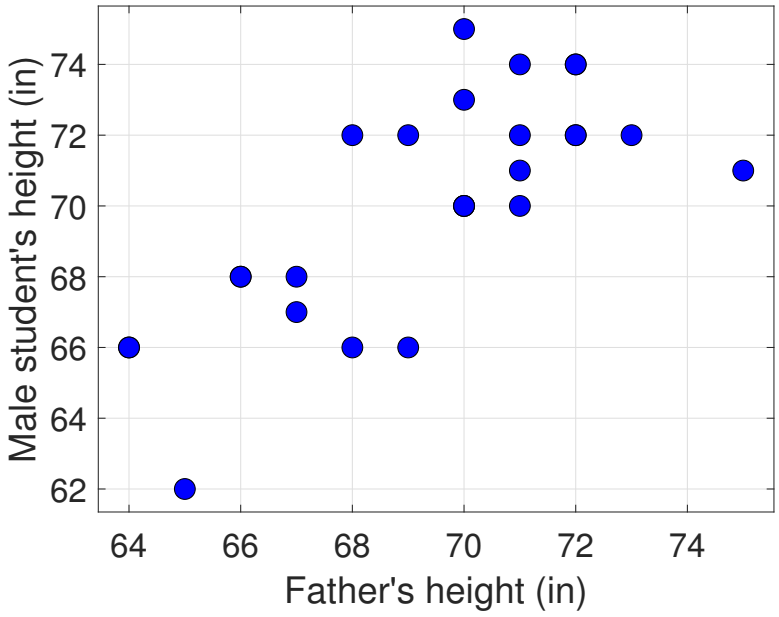
$$r = \frac{107626 - \frac{(1529)(1688)}{24}}{\sqrt{\left(97587 - \frac{(1529)^2}{24}\right)\left(118978 - \frac{(1688)^2}{24}\right)}} = 0.41$$

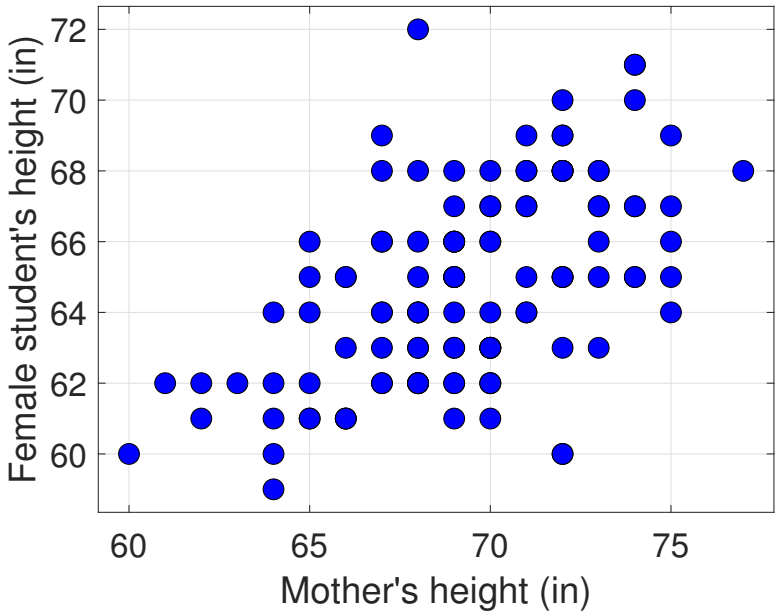
Later on this quarter we will define this formally by asking the question of what is the probablity of getting a correlation this high from a sample of this size under the null hypothesis that the correlation in the population is zero. For now though, do you think it's likely that you'd get a correlation this high by chance?

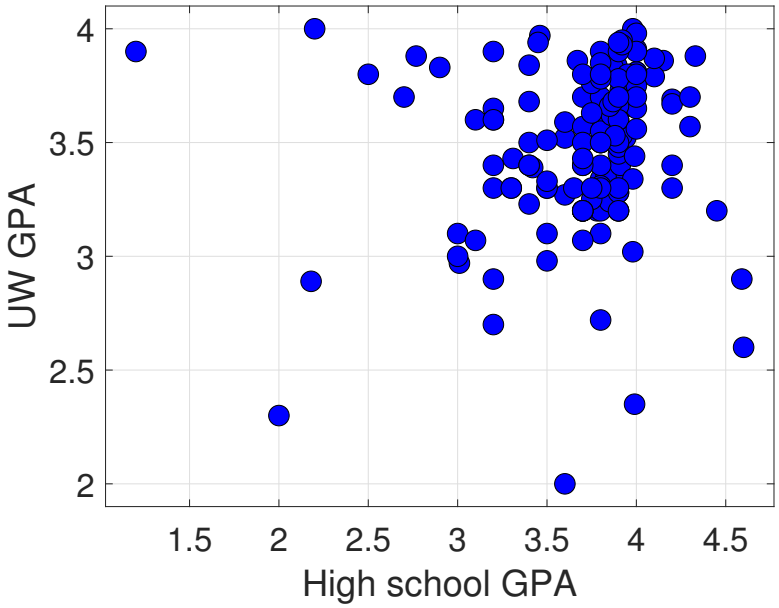
Survey data

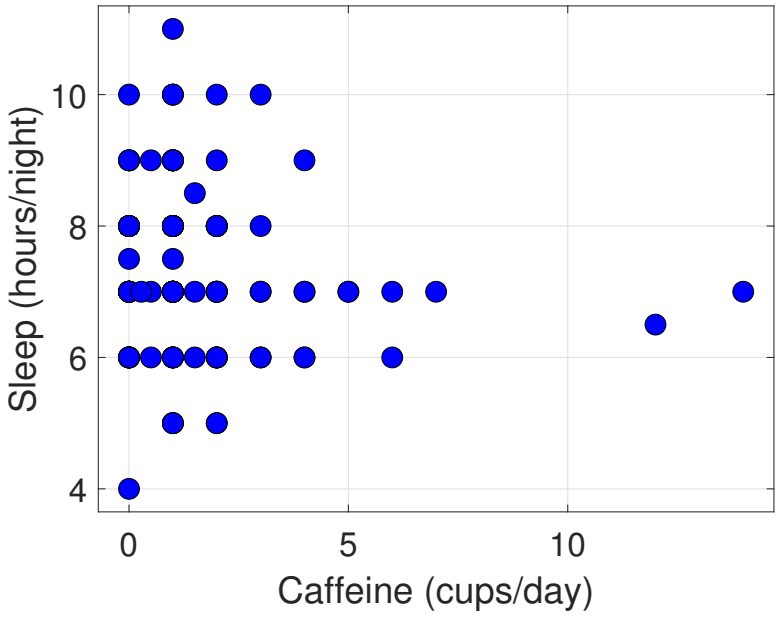
Let's fish through some of your survey data and look at some scatterplots. See if you can guess the correlations by eyeballing the plots. It takes practice but you can get pretty good in just a little time. The answers are at the end of this tutorial.

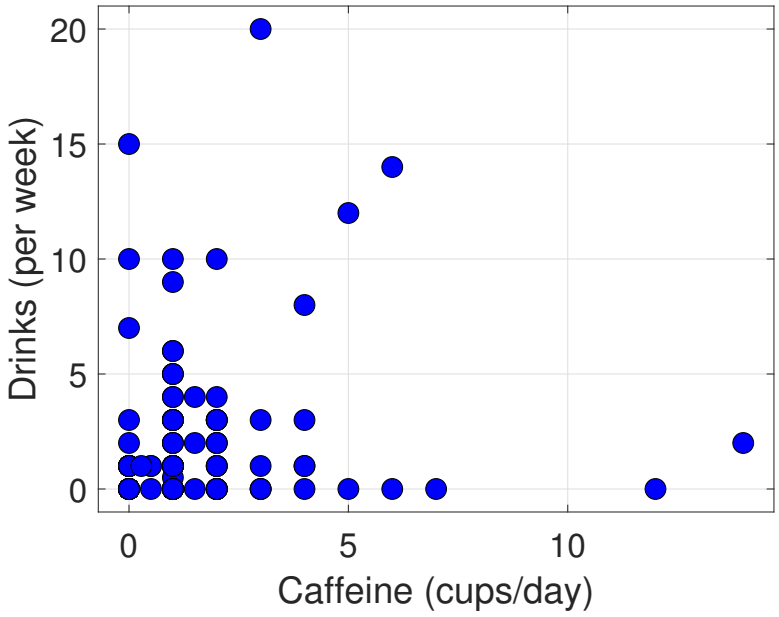


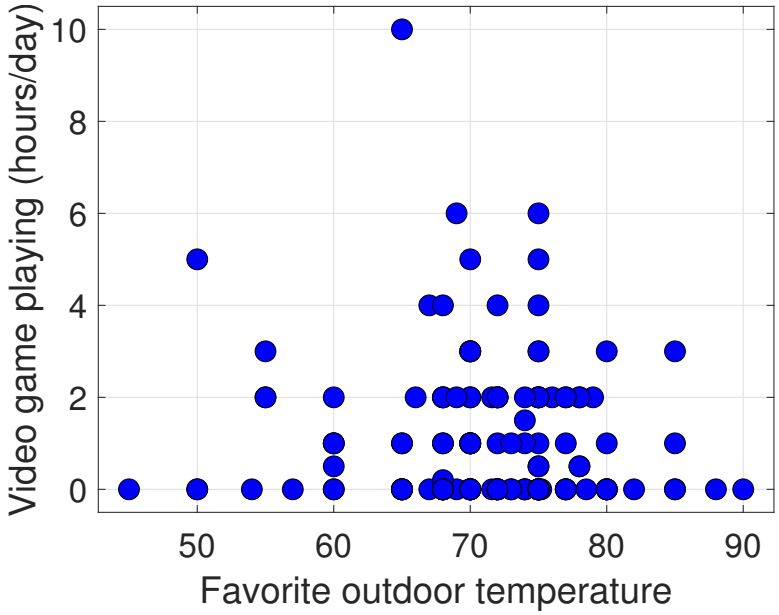

Male student's height (in)
74
72
70
68
66
64
62
64
66
68
70
72
74
Father's height (in)


Female student's height (in)
72
70
68
66
64
62
60
60
65
70
75
Mother's height (in)


UW GPA
4
3.5
3
2.5
2
1.5
2
2.5
3
3.5
4
4.5
High school GPA


Sleep (hours/night)
10
8
6
4
0
5
10
Caffeine (cups/day)





Fun facts about r

The Pearson correlation has some useful properties. Remember, z-scores don't change if you shift and/or scale your original scores. This means that we can change our units from inches to centimeters for either or both axes, and the Pearson correlation won't change.

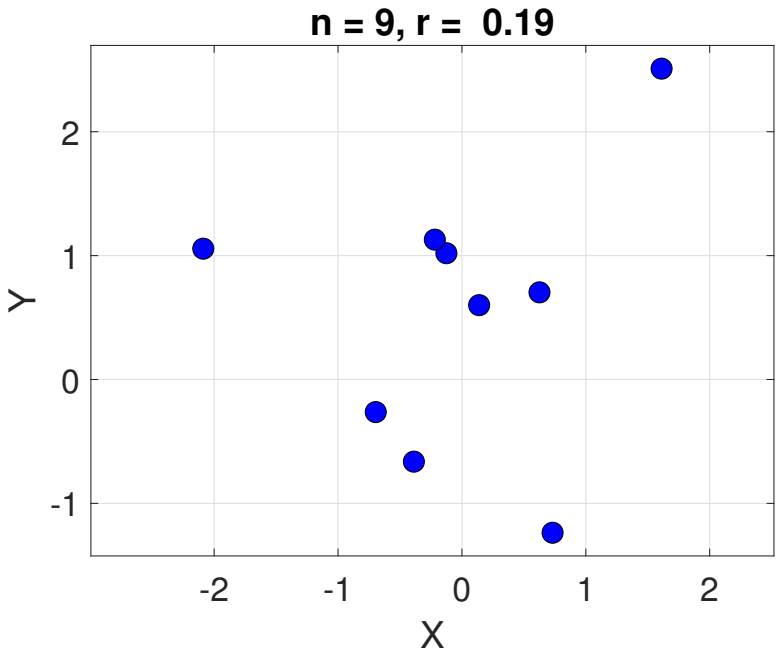
Another fact is that the Pearson correlation will always fall between -1 and 1. Thus, the Pearson correlation has no units - it's just a number that reflects the strength of the relation

between two variables. Later on when we get into regression we'll show that the value of r^2 has a specific meaning with respect to how well a straight line fits the data in the scatterplot.

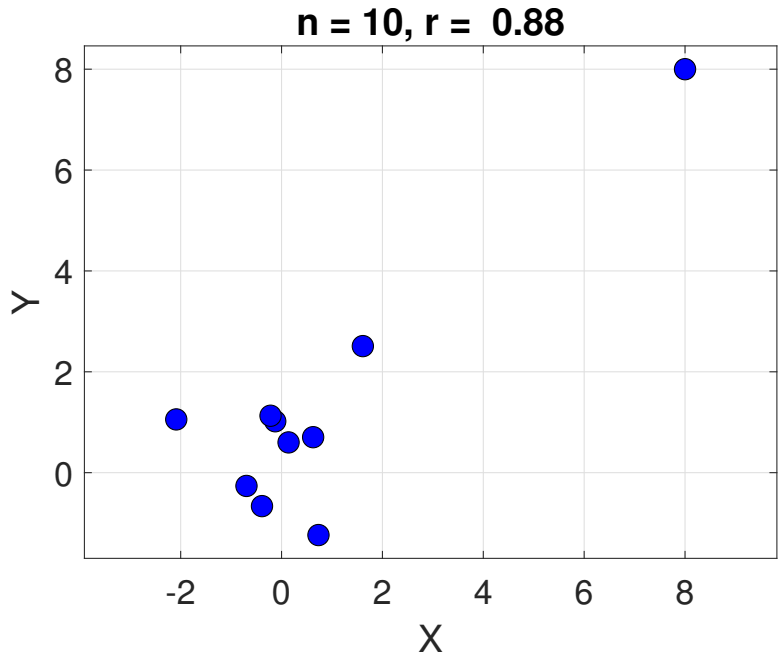
Sensitivity to outliers

If you look back at the z-score forumla, you'll see that the Pearson correlation is affected by not only which quadrant each point falls in, but also by how far the points are away from the means of x and y. If the product of a points's z-scores is large (either in positively or negatively), then that point will have a stronger influence on r. This should remind you of the mean: both the mean and Pearson correlation are sensitive to outliers.

Consider this scatterplot of points showing very little correlation between x and y:



Now we'll add one extra point off in the corner:



See how the correlation jumped from 0.19 to 0.88? That one outlying point has a serious effect on the Pearson correlation.

Spearman rank-order correlation

Remember how when we were dealing with skewed distributions, we showed that the median can be a better measure of central tendency than the mean? Well, another version of the correlation called the **Spearman rank-order correlation** is, like the median, less sensitive to outliers. Also, like the median, it is more appropriate when dealing with distributions that are not normal.

As name suggests, the Spearman rank-order correlation is caculated by first ranking the values for x and y. The lowest score gets a rank of 1, the second lowest 2 and so on. If there's a tie, we take the average of the ranks for the tie. So if there's a tie for the first three scores, each of those gets a rank of the mean of 1, 2 and 3, which is 2.

Here's a table of the x and y values in the scatterplot above, with the last (10th) pair being the outlier at position (8,8).

x	y	$rank_x$	$rank_y$
0.73	-1.24	8	1
-0.13	1.02	5	6
-0.39	-0.66	3	2
-0.7	-0.26	2	3
1.61	2.51	9	9
-0.22	1.13	4	8
-2.09	1.06	1	7
0.63	0.7	7	5
0.14	0.6	6	4
8	8	10	10

Note that the ranks for both x and y for this last point is 10, since they are the highest numbers in their lists.

Assume for the moment that there are no ties. With no ties, the **Spearman rank-order correlation is simply the Pearson correlation of the ranks**.

If we calculate the correlation of the ranks of the first 9 pairs of scores - the scores without the outlier - we get a value of r = 0.07. Including the outlier increases the correlation of the ranks to 0.32. This is a substantial increase in the Spearman correlation, but not as large as for the Pearson correlation, which increased from 0.19 to 0.88. Thus, the Pearson correlation is affected much more strongly by the outlier than the Spearman correlation.

Formula for calculating the Spearman correlation

Instead of calculating the Pearson correlation of the ranks, there's a simpler way to calculate the Spearman correlation. The trick is to first calculate the difference in ranks for each score: $D = rank_x - rank_y$, the squares of these differences, and finally the sum of these squares. Here's a table of these differences, D and their squares, D^2:

$rank_x$	$rank_y$	D	D^2
8	1	7	49
5	6	-1	1
3	2	1	1
2	3	-1	1
9	9	0	0
4	8	-4	16
1	7	-6	36
7	5	2	4
6	4	2	4
10	10	0	0

The sum of these squared differences is $\sum D^2 = 49 + 1 + ... + 0 = 112$

The Spearman correlation, defined as r_s, is then calculated by:

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2-1)}$$

Which for our data is:

$$r_s = 1 - \frac{6(112)}{10(10^2-1)} = 0.32$$

Keep in mind that the number '6' is just part of the formula, and doesn't depend on the data.

Note that the formula gives the same value that we got earlier by calculating the Pearson correlation of the ranks.

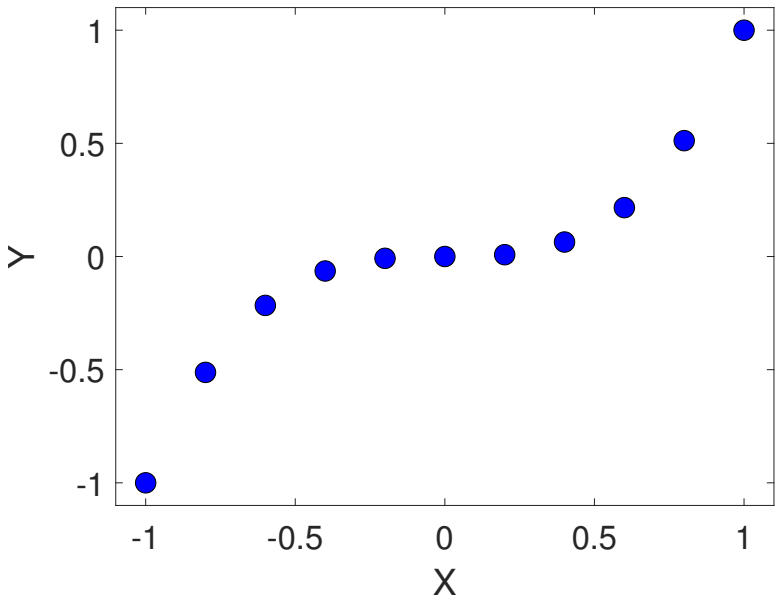
I realize that this formula is not intuitive. It can be derived from the Pearson correlation formula, and some facts about ranks. Yes, it's weird to have a '6' in a formula, but at least the formula is easy to use.

There is a one caveat regarding repeated values, or ties. If there are no ties, then formula above will always give you the same value as the Pearson correlation of the ranks. However, if there are ties, then the values may be different. In the case of ties, you should use the Pearson correlation of the ranks and not the simpler formula.

I'll tell you right now that in this class, if you happen to be asked to calculate the Spearman rank-order correlation, I won't give you examples with ties.

A funny property of the Spearman rank-order correlation

Consider this scatterplot:



The Pearson correlation for these points is r = 0.92. Close to 1, but not a perfect correlation. However, the Spearman rank-order correlation is exactly 1. Why?

Recall that the Pearson correlation is a measure of how well a line fits the data - a straight line, that is. But the Spearman correlation is a correlation of the ranks. The y-values on this plot are monotonically increasing with x. This makes their ranks simply increase by 1 as x goes from left to right. A straight line of y = x fits the ranks perfectly. Here's the table and corresponding calculation of r_s

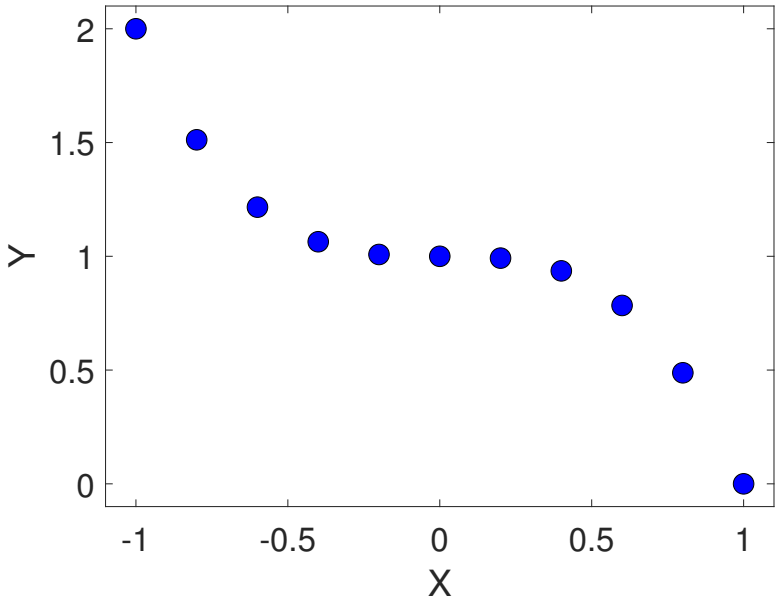
x	y	$rank_X$	$rank_Y$	D	D^2
-1	-1	1	1	0	0
-0.8	-0.512	2	2	0	0
-0.6	-0.216	3	3	0	0
-0.4	-0.064	4	4	0	0
-0.2	-0.008	5	5	0	0
0	0	6	6	0	0
0.2	0.008	7	7	0	0
0.4	0.064	8	8	0	0
0.6	0.216	9	9	0	0
0.8	0.512	10	10	0	0
1	1	11	11	0	0

The difference in ranks is always zero. The calculation of s_p is therefore:

$$r_s = 1 - \frac{6(0)}{11(11^2-1)} = 1.00$$

Which is why the Spearman correlation for monotonically increasing data is equal to 1.

What do you think the Spearman correlation is for this data:


2
1.5
1
0.5
0
Y
-1
-0.5
0
0.5
1
X

Correlations from survey scatterplots

Here are the corrrelations from the scatterplots from the survey:

The correlation between Mean of parent's height (in) and Student's height (in) is 0.49

The correlation between Father's height (in) and Male student's height (in) is 0.74

The correlation between Mother's height (in) and Female student's height (in) is 0.55

The correlation between High school GPA and UW GPA is 0.10

The correlation between Caffeine (cups/day) and Sleep (hours/night) is -0.14

The correlation between Caffeine (cups/day) and Drinks (per week) is 0.16

The correlation between Favorite outdoor temperature and Video game playing (hours/day) is -0.09

Correlations and Scatterplots in R

Simple scatterplots can be generated with R's 'plot' function, and correlations can be calculated with the 'cor' function. The following script generates some of the plots and values seen in this tutorial.

The R commands shown below can be found here: ScatterplotsAndCorrelations.R

```
# ScatterplotsAndCorrelations.R
#
# First, we'll load in the survey data
survey <-read.csv("http://www.courses.washington.edu/psy315/datasets/Psych315W21survey.csv")
# Scatterplots can be made with R's 'plot' function
#
# Here's how to make a scatterplot of Mother's vs. Father's heights
# for students that chose 'Green' as their favorite color.

# We first need to find the subset of the data for which
# survey$color is equal to "Green".  In R, we use '==' to check if
# something is true or not.  For example, to see which students chose
# "Green" as their favorite color, we can use:
students.green <- survey$color=="Green"
head(students.green)
[1] FALSE FALSE FALSE  TRUE FALSE  TRUE

# This list is the same length as survey$color, but is
# 'TRUE' where survey$color is "Green" and 'FALSE' otherwise.
```

```
# We can then refer to a subset of our data by using 'students.green'
# as an 'index'.  For example, the Mother's heights for the students
# that prefer "Green" are:
survey$mheight[students.green]
 [1] 60 60 64 63 69 66 60 62 60 64 64 64 62 64 66 63 64 62 68 65 65 67 68 59

# To make the scatterplot, we can use:
plot(survey$mheight[students.green],survey$pheight[students.green])
# The default labels and symbols are kind of ugly.  We can
# customize our plot by setting parameters like 'xlab' and 'ylab'
# for x and y labels, 'col' for the color and 'pch' to 16
# which is a filled circle, and 'asp' to 1, which makes the
# aspect ratio for the x and y axes to be the same, and 'cex'
# which for some reason sets the symbol size.
plot(survey$mheight[students.green],survey$pheight[students.green],
     xlab = "Mother's Height",
     ylab = "Father's Height",
     pch = 19,
     col = "blue",
     asp = 1,
     cex = 2)
# Search the web for more hints
# For different symbols (or setting for 'pch') see, for example:
# http://www.endmemo.com/program/R/pchsymbols.php
# Computing correlations in R uses the 'cor' function.
#
# To calculate the correlation between x and y, just use 'cor(x,y)'
#
# The correlation in this scatterplot is:
cor(survey$mheight[students.green],survey$pheight[students.green],use = "complete.obs")
[1] 0.4061526

# Next we'll calculate the correlations from the survey that are
# given at the end of the correlation tutorial

# To calculate the correlation between the mean of the parent's height
# and the student's height, we first calculate the mean of the father's
# height (survey$pheight) and the mother's height (survey$mheight). We'll
# call it parent.height:
parent.height <- (survey$pheight+survey$mheight)/2;
# Then we use 'cor' to calculate the correlation between x and
# the student's heights (survey$height):
cor(parent.height,survey$height,use = "complete.obs")
[1] 0.4933317

# Some of our survey data has missing values which you'll see as NaN's
# or 'not a number'. The third argument 'use = "complete.obs"' deals
# with this by telling 'cor' to only include data that has both
```

```
# (complate) observations.

# Next we'll calculate the correlation between the Father's height
# and the male student's height. This will require us to include
# a subset of the data (Male students) in our analysis.
#
# This can be done by referring to the subset of the data for which
# survey$gender is equal to "Male":
male.students <- survey$gender == "Male"
head(male.students)
[1]  TRUE FALSE FALSE FALSE FALSE  TRUE

# The heights of the male students that like Green is:
male.heights = survey$height[male.students]
# The heights of the fathers of the male students is:
male.heights.father = survey$pheight[male.students]
# and the correlation between the male student's heights and their
# fathers is:
cor(male.heights,male.heights.father,use = "complete.obs")
[1] 0.7397582

# Verify that the following computes the correlation between the
# female student's heights and their mother's heights:
female.students = survey$gender == "Female"
cor(survey$height[female.students],survey$mheight[female.students],use = "complete.obs")
[1] 0.4923266

# Here are the other correlations from the tutorial:
cor(survey$GPA_HS,survey$GPA_UW,use = "complete.obs")
[1] 0.1015848
cor(survey$caffeine,survey$sleep,use = "complete.obs")
[1] -0.1361281
cor(survey$caffeine,survey$drink,use = "complete.obs")
[1] 0.1598318
cor(survey$temperature,survey$games_hours,use = "complete.obs")
[1] -0.08867498

# You can use 'cor' to calculate a whole matrix of pairwise correlations
# This is done by creating a 'data frame' containing the variables
# of interest.  For example, let's look at the pairwise correlations
# between birth year, drinks/week, sleep, and caffeine
X = data.frame(survey$year,survey$drink,survey$sleep,survey$caffeine)
cor(X,use = "complete.obs")
                survey.year survey.drink survey.sleep survey.caffeine
survey.year      1.00000000   0.05181923   0.13244169       0.1119065
survey.drink     0.05181923   1.00000000   0.07421125       0.1726680
survey.sleep     0.13244169   0.07421125   1.00000000      -0.1331541
survey.caffeine  0.11190650   0.17266805  -0.13315414       1.0000000
```

```
# Notice that the diagonal values area all '1.000000'. That's
# because the correlation of a variable with itself is 1.0

# 'cor' will also calculate the Spearman correlation if you set
# the parameter 'method' to "spearman" instead of the default "pearson":
cor(survey$mheight,survey$pheight,
    use = "complete.obs",
    method = "spearman")
[1] 0.3153675
```