Lecture 7, Chapter 7 summary

Scatterplots, Association, and Correlation

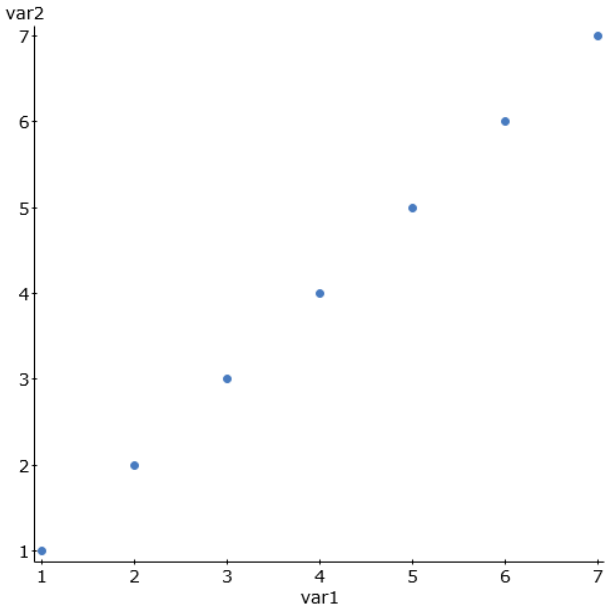
Topic:

- Association between two quantitative variables
- Use scatterplots to see the type of association
- It does not matter which variable goes into the x-axis (horizontal) or the y-axis (vertical line).
    - Sometimes, one variable is the outcome measure. The outcome variable, also known as the response variable, gets plotted in the y-axis, against another variable, the explanatory variable, in the x-axis. The explanatory variable explains the outcome.
- When we talk about association, we talk about:
    - Strength: Weak, moderate, strong
    - Form: straight line (linear), curve
    - Direction: upward trend (positive), downward trend (negative)
- We denote correlation with r
- $-1 \leq r \leq +1$

Let's see some examples of correlation between two quantitative variables:

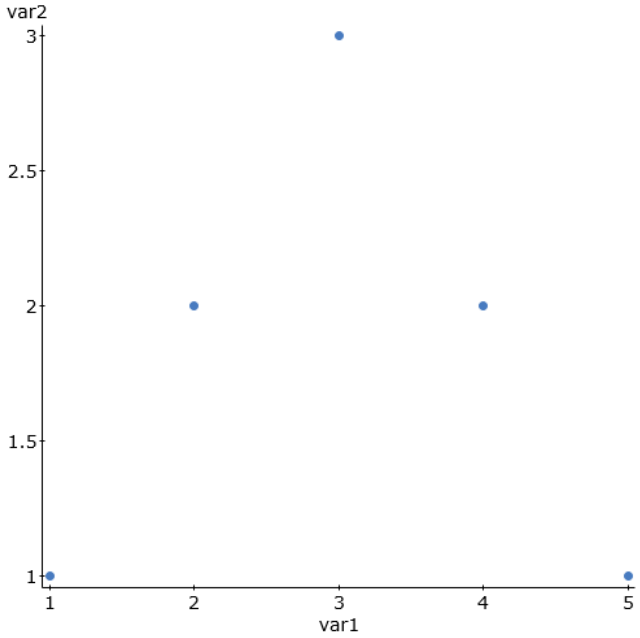
○ Perfect linear positive association: r = +1

Variable 1	Variable 2
1	1
2	2
3	3
4	4
5	5
6	6
7	7



○ No linear relation:

Variable 1	Variable 2
1	1
2	2
3	3
4	2



Important to know:

- When two quantitative variables are correlated, it does not mean that one variable causes the other.
- When we look at the relationship between the two variables, there are hidden (lurking) variables that we did not take into account. These are confounding variables. Hence, we avoid saying one variable causes another. See page 181, figure 7.9
- Correlation is sensitive to outliers. It is a good idea to investigate the outliers and see how the data is behaved with or with the outliers. Do not just remove the outliers, but investigate.

Here is the example we tried to explore towards the end of our class, on Tuesday. The example is exercise #18 on page 192.

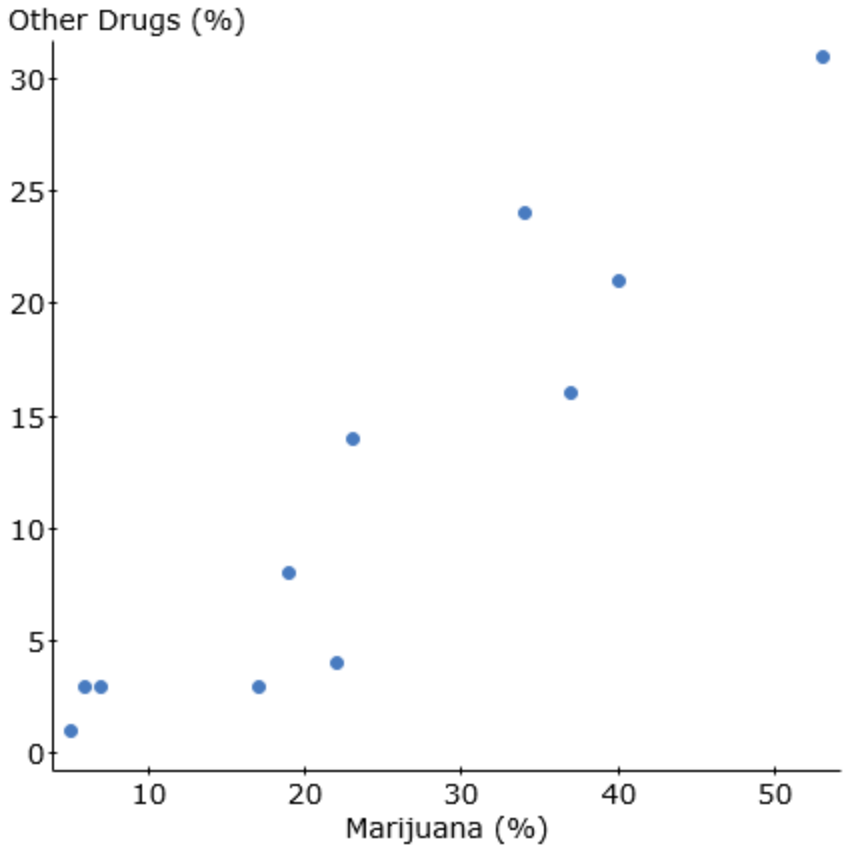
A survey was conducted in the United States and 10 countries of Western Europe determined the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the table.

Country	Marijuana (%)	Other Drugs (%)
CzechRep	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No.Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
USA	34	24

Summary statistics:

Column	Mean	Std. dev.
Marijuana (%)	23.909091	15.552842
Other Drugs (%)	11.636364	10.239851

- **The mean percent of teenagers who use marijuana is 23.91 (rounded to two decimal points), with standard deviation 15.55%**
- **The mean percent of teenagers who use other drugs is 11.64 (rounded to two decimal points), with standard deviation 10.24%**
- **Let's look at the pair of relationship (marijuana, other drugs)**



Correlation between Marijuana (%) and Other Drugs (%) is:
0.93410002

There is a strong positive linear association with % of teenagers who use marijuana and other drugs.

The formula for correlation, r is: $r = \dfrac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}}{S_x S_y}$

- the numerator $\dfrac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$ is the covariance of x and y (denoted as Sxy).
- Here are the steps to figure out r:
- $(x-\bar{x})$ is Marijuana%-mean(Marijuna%)
- $(y-\bar{y})$ is Other Drugs% - mean(Other Drugs%)
- Product is $(x-\bar{x})(y-\bar{y})$
- The sum of the product from the summary statistics: 1487.64
- The sum of product over n-1 is: 1487.64/10 =148.764

From the summary statistics:

- The Sx (standard deviation of %marijuana) is 15.55
- The Sy (standard deviation of % other drug) is 10.24

So, the correlation is: r = (148.764)/(15.55*10.24) = 0.934

Country	Marijuana (%)	Other Drugs (%)	Marijuana%-mean(Marijuna%)	Other Drugs% - mean(Other Drugs%)	product
CzechRep	22	4	-1.91	-7.64	14.58
Denmark	17	3	-6.91	-8.64	59.67
England	40	21	16.09	9.36	150.67
Finland	5	1	-18.91	-10.64	201.12
Ireland	37	16	13.09	4.36	57.12
Italy	19	8	-4.91	-3.64	17.85
No.Ireland	23	14	-0.91	2.36	-2.15
Norway	6	3	-17.91	-8.64	154.67
Portugal	7	3	-16.91	-8.64	146.03
Scotland	53	31	29.09	19.36	563.31
USA	34	24	10.09	12.36	124.76

Summary Statistics

Column	n	Sum	Mean	Std. dev.
Marijuana (%)	11	263	23.909091	15.552842
Other Drugs (%)	11	128	11.636364	10.239851
product	11	1487.6364	135.23967	157.30572

The following are the steps for how to open the file, exercise 18, on page 192 in StatCrunch.

- **If you purchased your text book new and have the eBook version (where you registered your access code on www.mystatlab.com using the courseID aslemand94682), then you would simply log in and open statcrunch and go to dataset for the textbook, under chapter 7, click on drug abuse.**
- **If you have a used book, you can purchase an access code from www.mystatlab.com to get StatCrunch. It is about $25 for 12 months access.**
  - **The following are the steps on how to open the file:**
  - **You either have a CD at the back of the book where you can get the data files or you may e-mail me to send them to you!**
  - **Ok, now back to how to open the file in statcrunch.**

Data >Load>From file>On my computer

- **Browse your file from the location in which you saved the file**
- **It is a good idea to first open the excel file before opening it in StatCrunch and to see if columns have headings(names) – a good data file should have variable names. Our example does ☺**
- **Check off use first line as column name (I think by default it is checked for us).**
- **Leave everything else as is. Scroll down the page to click upload file**

Graph>scatterplot> X-column: marijuana%
Y-column: other drug use %
Click compute

Stat>summary stats> correlation > select columns (hold the shift key to select two variables): marijuana%, other drug use %