Some Statistical Traps

Nikodem Lewandowski

University of Gdańsk

All the tables have the same mean and SD

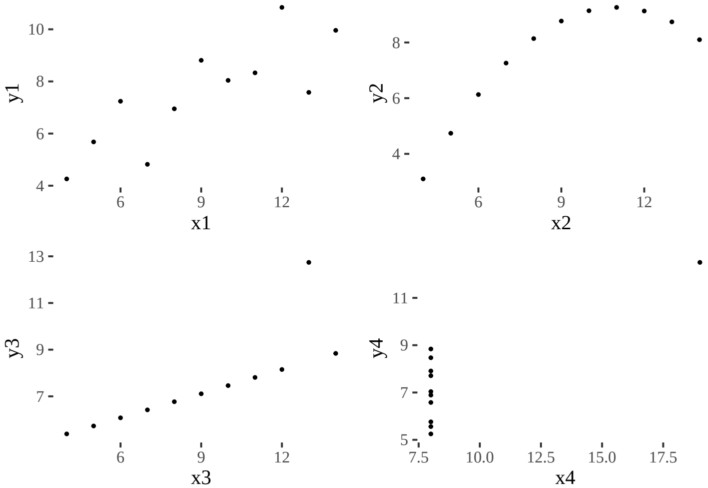
	val1	val2	val3	val4	val5	val6	val7	val8	val9	val10	val11	Mean	SD
x	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00	9.000000	3.316625
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68	7.500909	2.031568

	val1	val2	val3	val4	val5	val6	val7	val8	val9	val10	val11	Mean	SD
x	10.00	8.00	13.00	9.00	11.00	14.0	6.00	4.0	12.00	7.00	5.00	9.000000	3.316625
y	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74	7.500909	2.031657

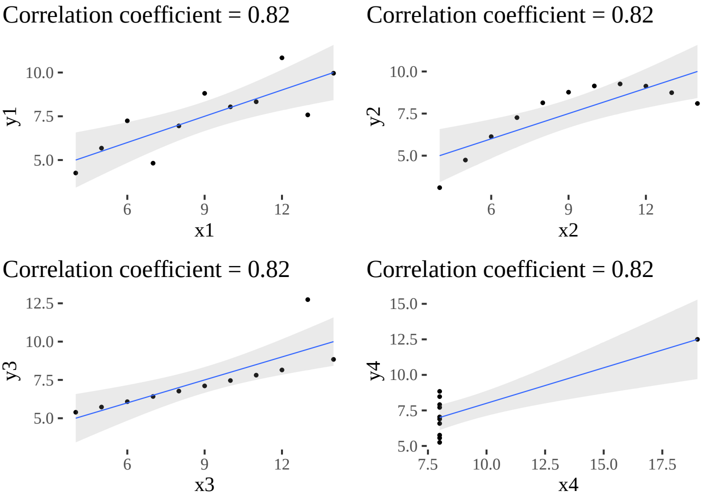
	val1	val2	val3	val4	val5	val6	val7	val8	val9	val10	val11	Mean	SD
x	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00	9.0	3.316625
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73	7.5	2.030424

	val1	val2	val3	val4	val5	val6	val7	val8	val9	val10	val11	Mean	SD
x	8.00	8.00	8.00	8.00	8.00	8.00	8.00	19.0	8.00	8.00	8.00	9.000000	3.316625
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89	7.500909	2.030578

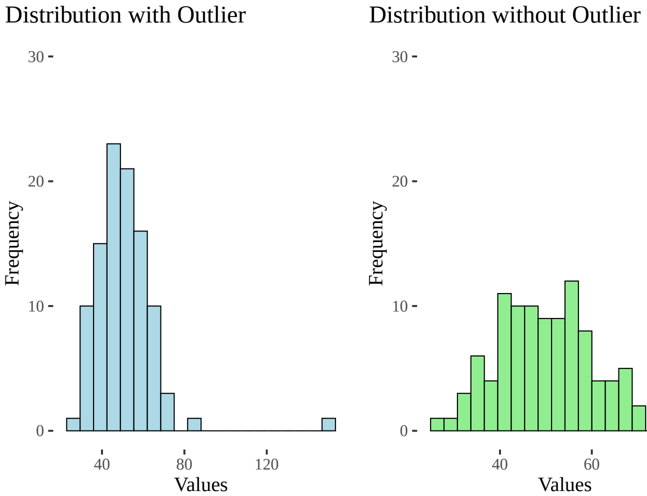
Visualizations of those tables



Linear regression and Pearson's Correlation



Outliers can affect the look of the distribution



Testing probabilities

- You have built a new spam filter for your email
- It was able to filter 99.5% of spam emails
- Is it accurate then?

Testing probabilities

- Despite the impressive 99.5% filtering rate, the accuracy of the spam filter needs closer examination.
- It turns out that only 0.1% of all emails are spam, and it categorized 25% of good emails as spam.
- **Specificity (True Negative Rate):**

$$Specificity = \frac{TN}{TN + FP}$$

$$\frac{0.75}{0.75 + 0.25} = 0.75$$

$$Specificity = 0.75$$

- **Sensitivity (True Positive Rate):**

$$Sensitivity = \frac{TP}{TP + FN}$$

$$\frac{0.995}{0.995 + 0.005} = 0.995$$

$$Sensitivity = 0.995$$

Suplementary materials

Cool YouTube videos to watch:

- How statistics can be misleading - Mark Liddell
- How We're Fooled By Statistics