Statistical Literacy

Problem Set 3

Rémi Viné
Due October 16th, 2023

1. **Compute indicators.** Using table 1, that is a sample of workers, compute and briefly comment on

Frequency	Years of higher education	Cumulative Frequency
15	0	15
2	1	17
5	2	22
64	3	86
37	4	123
68	5	191
23	6	214
2	7	216
8	8	224
15	9	239

Table 1: Workers in a given place (sample)

(a) The mode

(b) The median

(c) The average

(d) Compare the median and the average

(e) The range

(f) The IQR

(g) The variance and the standard deviation. Compare the latter with the IQR.

(h) Construct the Boxplot of the variable "Years of higher education". Based on this boxplot, would you suspect outliers?

2. **Distributions.** Using the 4 boxplots (visually):

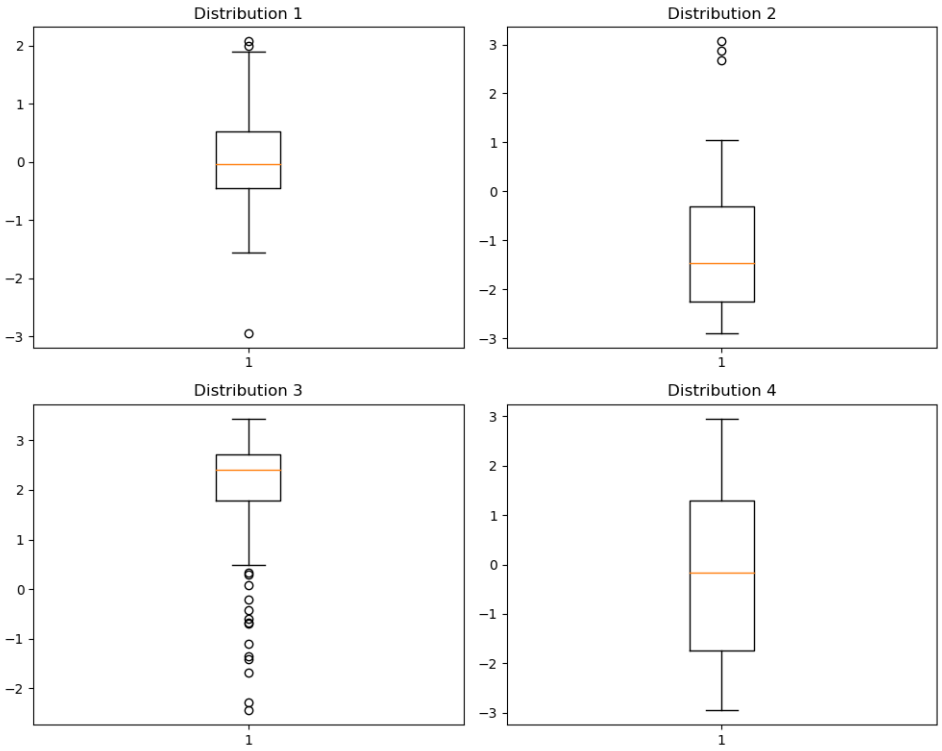
   (a) What distribution has the largest IQR?

   (b) What distribution's average is likely to change the most with and without outliers (so that in the second scenario suspected outliers would be dropped from the data)?

   (c) What distribution(s) would be suspected to be symmetric?

   (d) What distribution(s) would be suspected to be left-skewed (negative skewness)?

   (e) What distribution(s) would be suspected to be right-skewed (positive skewness)?



3. **More on distributions.** The boxplot below helps better understand the distributions of life expectancy depending on population groups[1]:

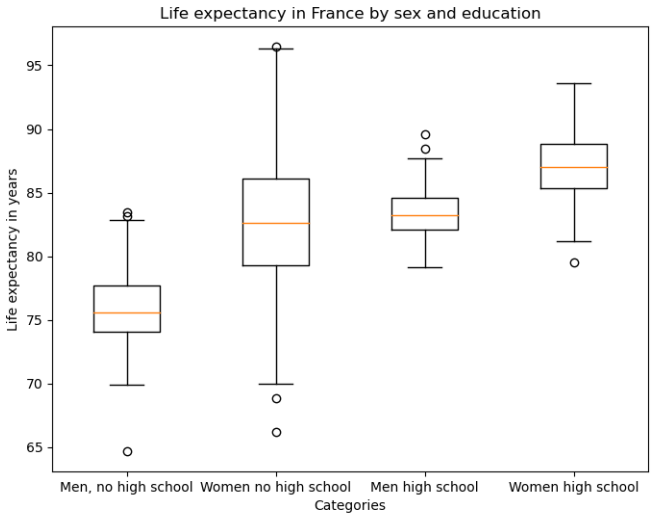
   (a) In the data presented, to what category belongs the person with the lowest life expectancy?

   (b) To what category belongs the person with the highest life expectancy?

   (c) Comparing men with high school diploma and women without, would you conclude on a clear difference of the median?

   (d) What would you conclude on the education difference in life expectancy?

   (e) What would you conclude on the sex of the person with respect to the median life expectancy?



[1] Fictitious variance, but real centrality points; for more information on inequality on life expectancy see for example Didier Fassin (in French)

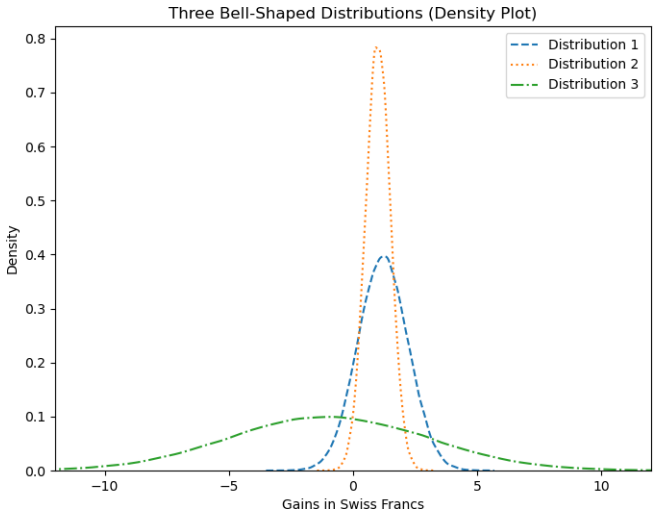
4. **Dispersion and centrality.**

(a) Having the choice between two chance games (see table 2) , what game would you choose and why?

Gains in CHF	Game 1	Game 2
Average gain	100	120
Standard deviation of the gain	20	60

Table 2: Dispersion and centrality

(b) Let's look at the graph below that shows gains from a chance game. There are three scenarii, three games with different outcomes depicted by the three different distributions. To game 1 corresponds distribution 1, to game 2 corresponds distribution 2, to game 3 corresponds distribution 3.



i. What game has the largest variance?

ii. What game has the largest average?

iii. In what case would you choose distribution 3 playing the game?

iv. In what case would you chose distribution 2 playing the game?

v. *Difficult.* Assume distribution 3 has an average of -1 and a standard deviation of 4 and that distribution 2 has an average of 1 and a standard deviation of 1, by how much is it more likely to gain more than 3 Swiss Francs playing game 3 than playing game 2? *(Hint: these distribution are bell-shaped so that one rule seen in class applies.)*

5. **Multiple choice.**

(a) In the given data set : $\{1, 2, 2, 2, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9, 9, 10, 11\}$ what centrality indicator changes the most if the observation valued "6" were instead valued "10000"?

A The average

B The median

C The mode

(b) And what dispersion indicator changes the most in this case?

A The interquartile range

B The standard deviation

(c) Will there be any changes in the possible asymmetry?

A No, the asymmetry is not affected

B Yes, more left-skewness is expected

C Yes, more right-skewness is expected