Statistical Literacy

Problem Set 5

Rémi Viné
Due October 30th, 2023

1. **Random variables.**

   (a) Using the plot below (figure 1), what statement **could** be correct?

      i. $P(Observations\ in\ shaded\ area) = 1.43$
      ii. $P(Observations\ in\ shaded\ area) = 0.55$
      iii. $P(Observations\ in\ shaded\ area) = -0.4$
      iv. $P(Observations\ in\ shaded\ area) = 4$

   (b) Using the distribution plot below figure (2), what is the share of the observations outside the shaded area?

      i. $P(Observations\ outside\ shaded\ area) = 0.55$
      ii. $P(Observations\ outside\ shaded\ area) = 0.45$
      iii. $P(Observations\ outside\ shaded\ area) = 0$
      iv. $P(Observations\ outside\ shaded\ area) = 0.5$

   (c) Using the distribution plot below figure (2), can you conclude on the share of the observations outside of the shaded interval and on the *right* of this interval?

      i. Yes, it is 0.45
      ii. Yes, it is 0.225
      iii. Yes, it is 0.1
      iv. No, we need more information to conclude

   (d) In figure (2), if we are told that the right part of the non-shaded area accounts for a third of the non-shaded area, what is the share of all observations that are in the left part of the non-shaded area?

      i. Yes, it is 0.45
      ii. Yes, it is 0.67
      iii. Yes, it is 0.3
      iv. No, we need more information to conclude

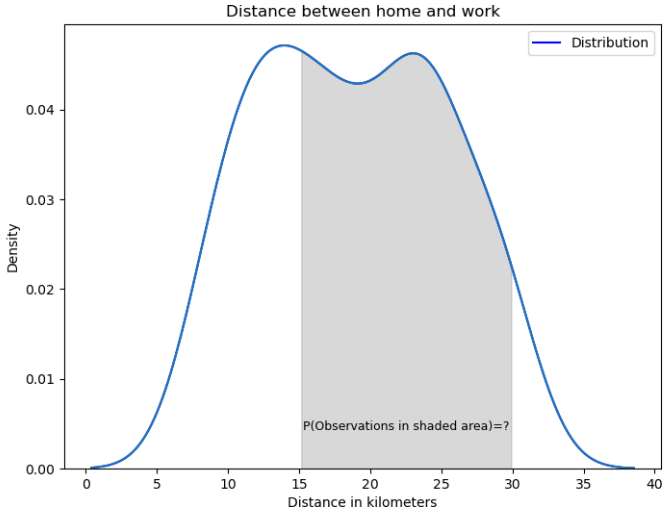


Figure 1: First distribution

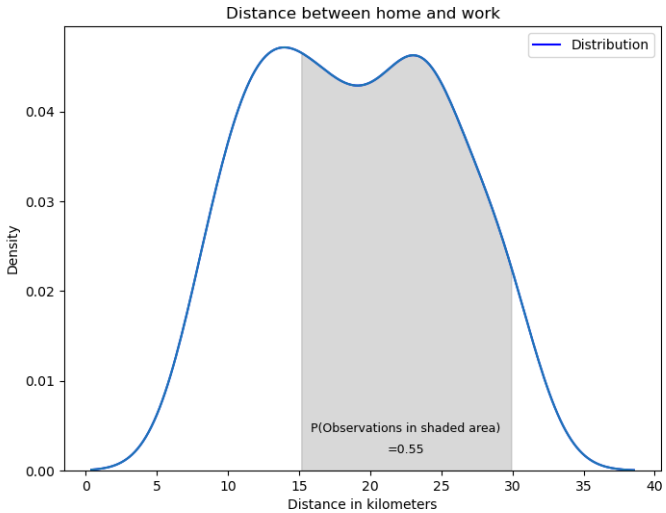
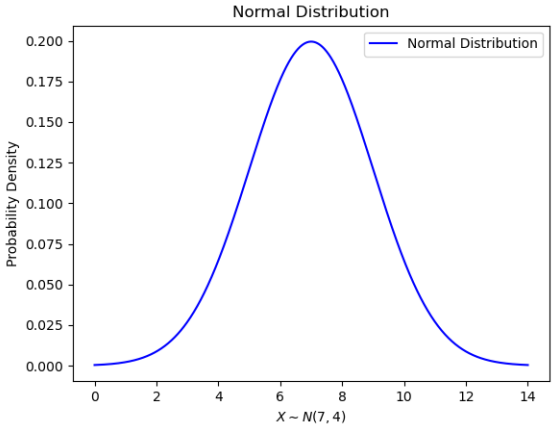


Figure 2: Second distribution

2. **Normal distribution.** *Hint: recall the empirical rule.*

Assume the (log) distance (so expressed in log of kilometers) between the hometown and Geneva of IHEID students is Normally distributed: $X \sim N(7; 4)$.

Note here that the use of Log does not change the calculations, it simply changes the interpretation: for example $x = 6$ does not mean 6 kilometers but $log(6)$ kilometers. Without the use of log, the distances would probably have been right skewed while it is now assumed to be Normal. For questions (a) to (e), simply do the calculations using the Empirical Rule approximation.



(a) What is the share of students living less than 3 (log km) from Geneva?

(b) What is the share of students living less than 1 (log km) from Geneva?

(c) What is the share of students living more than 7 (log km) from Geneva?

(d) What is the share of students living between 3 and 13 (log km) from Geneva?

(e) What is the share of students living less than 5 (log km) from Geneva or more than 11 (log km) from Geneva?

(f) What is the share of students living less than 4 (log km) from Geneva?

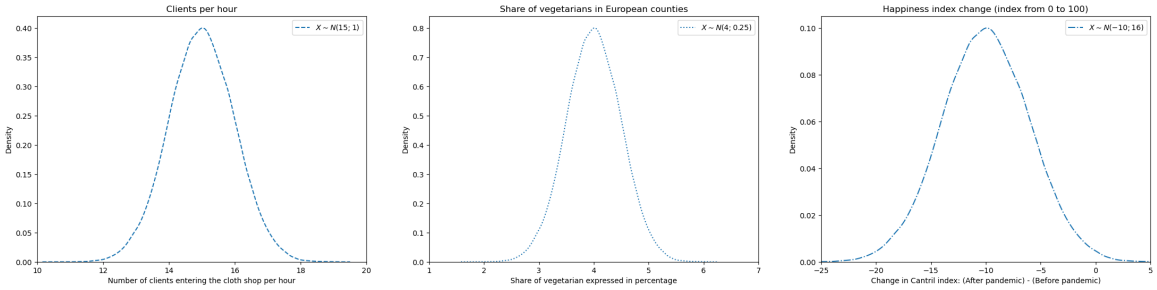
3. **Z table reading.** *If possible (might not always be possible!)*, using the Z table, find

   (a) $P(Z > 1)$
   (b) $P(Z < 1)$
   (c) $P(Z < 2)$
   (d) $P(-1.5 < Z)$
   (e) $P(Z > -1.44)$
   (f) $P(-1.8 < Z < -0.7)$
   (g) $P(Z < -1.9 \cup Z > 1.3)$
   (h) $P(X < 0.3)$
   (i) What is the z-score so that there are 5% of all observations above it?
   (j) What is the z-score so that there are 2.5% of all observations below it?
   (k) Find the two *z-scores* so that $P(z_1 < Z < z_2) = 0.8$.

4. **Standardizing.**

You are given three different variables, each is assumed to follow a Normal distribution. The first variable is the number of clients entering a given shop per hour ($X \sim N(15; 1)$). The second variable is the share of vegetarians in different European districts ($X \sim N(4; 0.25)$). The third variable is the change of the Cantril index **put on a $0-100$ scale**, looking at $(Cantril_{2022} - Cantril_{2019})$, with $(X \sim N(-10; 16))$.[1]

Distributions are depicted below. Pay attention to the scale.



(a) For each distribution, indicate the (i) average, (ii) standard deviation

(b) For each distribution, what is the probability to have an observation below 5 (expressed in the different units - clients, percent, score)?[2]

(c) What if, instead, you are asked to compute the probability that the *z-score* for each distribution is lower than 1?

(d) For each distribution, retrieve the corresponding value of X for $Z = 1$ and interpret it in the context of each distribution.

[1] Note that Cantril's book, *The pattern of human concerns*, 1965, is available at the library of IHEID.

[2] Here, look at the bound closer to the central indicator if the z-score is not directly put on the table and mark an inequality rather than an equality. For example, take $z = 1.64$ rather than 1.65 for $P(Z > z) = 5\%$.