Instructions

Upload your own solutions (handwritten or typed) to the corresponding Canvas Assignment on or before the due date: 26/11/2021.

Grading scheme

a	b	c	d	e	f	g	h	Total
2	4	4	4	2	1	2	1	20

Exercise 1

About the data. A small tech startup company wants to advertise their new product during the Python tutorials of a YouTube creator. The startup would like the creator to show a 30 second video at the beginning of her next 10 videos. The creator offers them two possible deals.
Offer 1: The startup pays a flat rate of 100€ per video,
Offer 2: The startup pays 1€ for every thousand views per video.

The startup has a total budget of 1500€. They can't afford to go over their budget, and ideally would like to minimize their spending. In order to make a good choice, the startup collects the number of views of the last $n = 30$ videos released by the creator.

Number of views (rounded to the nearest thousand)									
45 000	57 000	61 000	55 000	30 000	50 000	33 000	89 000	53 000	102 000
75 000	77 000	43 000	27 000	67 000	30 000	52 000	78 000	111 000	179 000
87 000	13 000	47 000	32 000	68 000	26 000	85 000	51 000	5 000	55 000

a. Use a simple calculator to compute the sample mean $\overline{x}$ and the sample standard deviation s. Round your answers to the nearest hundred.

Hint: The mean, standard deviation, and percentiles have the same units as the measurements, so you can first get rid of the three zeroes at the end of each measurement, do your computations, and then add the zeroes back. Just make sure your final answers make sense!

b. Compute the sample median. Comparing the sample mean and median, what does the difference tell you about the data's distribution?

c. Compute the 1^{st} and 3^{rd} quartiles, and determine the length of the inter-quartile range.

d. Determine outliers using the 1.5 × IQR-rule. Is the number of outliers large? Are there any extreme observations according to the 3 × IQR-rule?

e. Compute the interval $(\overline{x} - 3s, \overline{x} + 3s)$.

f. Given what you know about YouTube, do you think the outlier(s) and/or extreme observation(s) (if any) are due to a measurement error or are they representative of the nature of the data?

Let $Y_1, Y_2, \ldots, Y_{10}$ be random variables denoting the number of views that the next 10 videos might get, and let $Y = \sum_i Y_i$. For simplicity, let's assume that the data is normally distributed, with mean 60 and standard deviation 30 (both measured in thousands of views), that is $Y_i \sim$

$N(60,\ 900)$. In that case we have $Y \sim N(600,\ 9000)$. If the startup chooses Offer 2, and $Y \geq 1500$, the company will be out of budget.

g. Under the above normality assumption, what is the probability that $Y \geq 1500$? (You might need to use the internet instead of a statistical table).

h. With all the information at hand, which offer would you choose? Explain your reasons.

Exercise 2

This is a personal reflection, you **do not** need to submit anything for this part.

Before you begin, make sure you have finished the mandatory SPSS assignment (do not submit both exercises together, there are two separate Canvas assignments!).

Based only on the histogram, descriptive statistics, and QQ-plot, do you think the data looks normal?

What if you were to drop the outlier in the data set? Would the data look convincingly normal without this outlier?

Given what you know about the internet, what do you think is the probability of a video going viral? (eg. one in a million? one in a billion?) Does it match the probability you computed in part **g.**?

Taking everything into account, do you think it is reasonable to make the normality assumption described at the end of Part 1?

Follow up. The YouTube creator releases 10 more videos. Out of these, 9 have between 50 thousand and 100 thousand views. In the last video, the creator's cat jumps on her shoulder at the exact moment she is making a Python pun. The video goes viral and receives 10 million views in one week. Needless to say, the startup, which had gone for Offer 2, is now in financial trouble.

The normal assumption guaranteed that it was virtually impossible to receive so many views on a single video, yet we know that in real life videos do go viral. What went wrong?

Do's and Do not's for choosing a model

- 👍 Always exercise caution and be critical (but remember we never have absolute certainty)
- 👍 Make use of the tools you have: plots, descriptive statistics, normality tests ...
- 👍 Take into account knowledge of the domain: is there a reason why the data would be symmetric around the mean? Are extreme observations as unlikely as the normal model predicts? What models have people before you used in similar situations?
- 👍 Ask someone or look for online help when in doubt

- 👎 Don't assume that everything is automatically normal!
- 👎 Don't discard outliers unless you are absolutely sure there was a mistake