Statistics for EMCL students
Assignment 4
Solutions

A.

A background explanation for this part of the assignment: As you know, standard deviation is a measure of the spread of the distribution, that is, of how far away from the mean the data points are located. The idea behind it is to average the distances from the average. In other words, the distance of a data point from the mean is called the ***deviation*** of that data point, and the mean of these deviations should describe the "width", or the "spread", of the distribution. Thus, a possible measure of spread would be:

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| .$$

In words: we first calculate the absolute value of the difference of x_i (a data point) and of the mean $\bar{x}$ for each *i*. This difference $\left|x_i - \bar{x}\right|$ is the deviation of data point x_i. Then, we calculate the mean of the deviations (summing them up, and dividing them by *n*, the number of data points). This is exactly the value that you have (should have) calculated in the assignment: first, you introduced a new variable that was the absolute value of the deviation, and then you let SPSS calculate the mean of this new variable.

Now, a trick is that the absolute value can be replaced by the square root of the square:

$$\left|x_i - \bar{x}\right| = \sqrt{\left(x_i - \bar{x}\right)^2} .$$

So the same measure of spread can also be written thus:

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(x_i - \bar{x}\right)^2} .$$

Nonetheless, as mentioned earlier in this course, statisticians decided not to use this formula for mathematical reasons. In order to have all our statistical procedures work, we need to change two details in this formula:

1. replace *n* with *n-1*;

2. reverse the order of the mean computation and of the square root (first sum up the squares and divide it by *n, i.e. b*y *n-1;* and only then take its square root).

Therefore, the standard deviation has this form (compare it to the previous formula):

$$\sigma_{(n-1)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} .$$

The goal of the assignment, beside learning how to introduce new variables by using variable transformations, was exactly to compare how much these two changes influence the measure of spread. Here are the results returned by SPSS (the first table concerns variable *MLU*, whereas the second table concerns the derived variables *absdev*):

Statistics

MLU

N	Valid	20
	Missing	0
Mean		5,80
Std. Deviation		2,484

Statistics

absdev

N	Valid	20
	Missing	0
Mean		2,1600

Hence, the correct answers are:

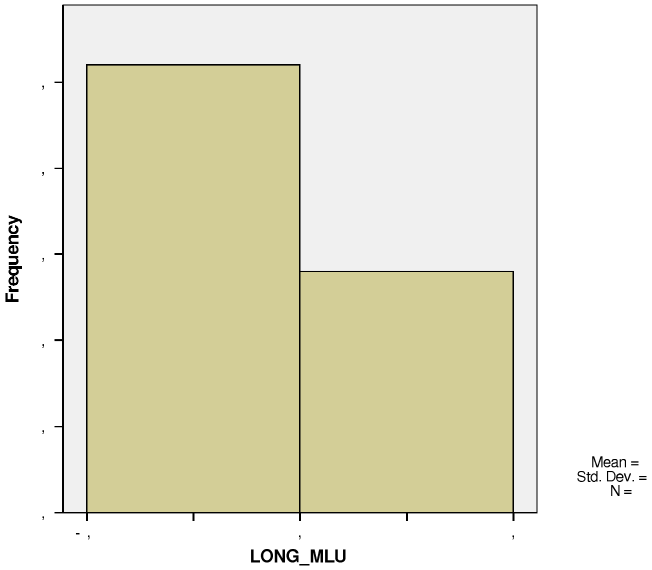
*** 1. Copy the mean of ABSDEV to your report.**

Mean of absdev: 2.16

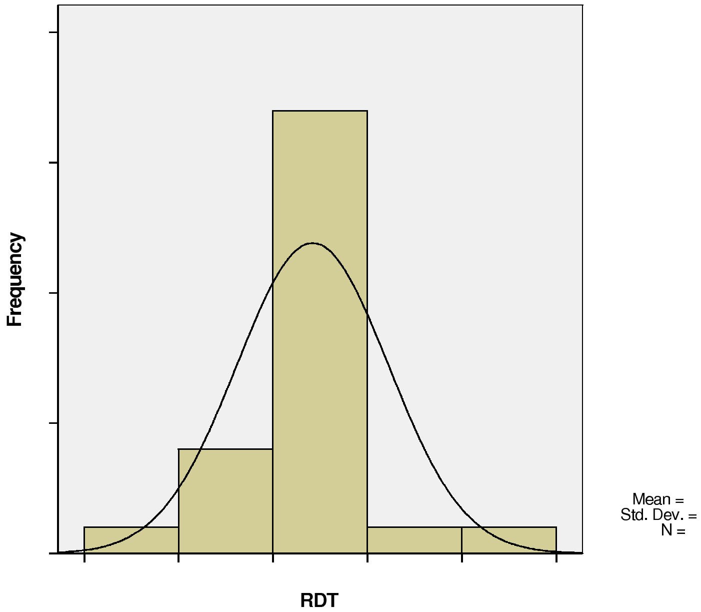
*** 2. Compare the SD (calculated last week) with the mean of the deviations. For what two reasons (two differences in the way they are calculated) do they differ?**

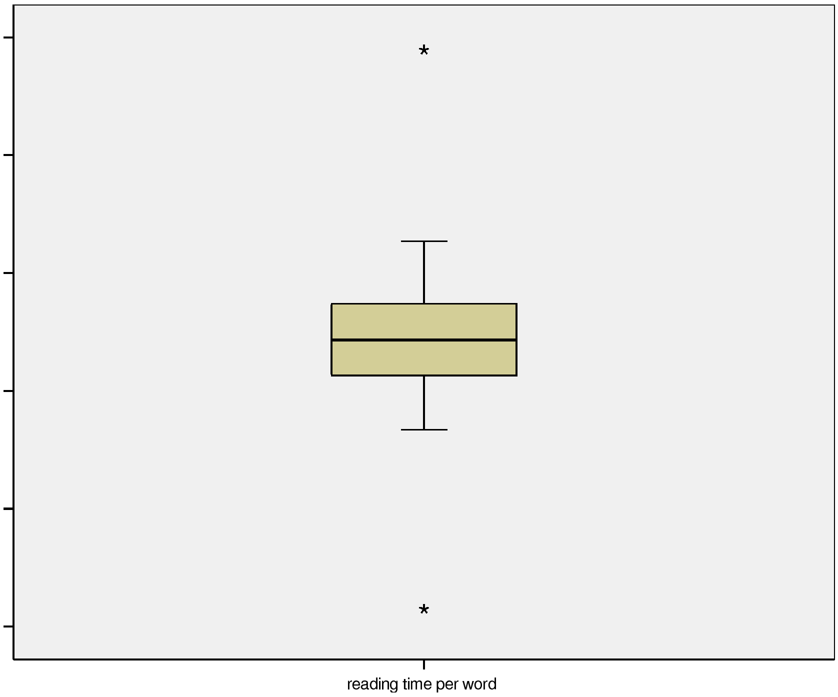
Std. Dev. of MLU: 2.48.
This value differs from the 2.16 given at question 1 for two reasons:
1. we used then the absolute values of deviations, instead of squares of deviations;
2. we divided by *n*, and not by *n-1*.

*** 3. Create a histogram of LONG_MLU, and copy it to your report.**



*** 4. Create a histogram including a Normal curve, as well as a boxplot of RDT. Copy it to your report.**





Please note that the longer **label** you have specified in Variable View (and not the shorter **variable name**) is used as the title of the boxplot.

*** 5. You can find two outliers among your data. Which are they, and what kind of explanation(s) could you provide to explain them?**

Two (possible) outliers can be observed in the boxplot, as they lie outside of the "whiskers" (cf. 1.5 × IQR rule). These are case nr. 5 (having value 30), and case nr. 23 (having value 979). Their case numbers are given next to the star, whereas all other data points are replaced by the boxplot.

Possible explanations for the outliers:
- These words are special in some sense, e.g., case 23 is an extremely rare word, and case 5 is a function word (quickly read). Actually, if this is the case, then the experiment is badly designed. Moreover, I don't think that you can read a function word in 30 msec.
- Errors in the measurement procedure: The subject typed the key by mistake twice, and therefore did not actually read word 5. Whereas for word 23, the subject might have paused for a moment (look elsewhere, ask a question, etc.).

*** 6. In case you decide to remove these cases from your data set, do you expect the mean or the standard deviation to change more? Why?**

Each of them has an important influence on the mean, which is not a resistant measure. Still, the two outliers are sort of "symmetrically" distributed around the other data, so they neutralize somehow each other's effect on the mean.
They both contribute largely to the standard deviation. So removing them will significantly reduce the standard deviation.

*** 7. What can you observe, as compared to your previous results?**

Mean changed only slightly, from 484 to 482.
Standard deviation changed drastically, from 161 to 82.

*** 8. You know the size of the sample, and you know its standard deviation. What is the standard error, then? Calculate it both by hand (give details of your calculation in the report) and let SPSS calculate it for you. Are the two values the same?**

Standard error: standard dev / √ sample size = 82.288/√22 = 17.5439
(I have used the calculator of Windows).
SPSS returns similarly 17.544.

*** 9. Determine the confidence interval for the mean of the sample using the Student-t-statistic. Provide details of your calculations in your report.**

The degree of freedom is $df = n\text{-}1 = 22\text{-}1 = 21$.
For C = 0.95, Table D gives $t^* = 2.080$ (row with $df = 21$)

The confidence interval is: mean ± $t^* \cdot SE$ =
= *481.77* ± 2.080 · *17.544* = *481.77* ± 36.492
In other words, the confidence interval is [445.28, 518.26], or simply [445, 519].

(NB: It does not make sense to keep too many digits, so I have rounded off the values. Yet, in the case of a confidence interval, it is always safe to round values "away" from the mean, that is, to have a slightly larger interval.)

*** 10. What is the meaning of this confidence interval?**

We are 95% confident that the population mean of variable RDT is larger than 445 ms and smaller than 519 ms.

What does it mean that we are "95% confident"? It means that if we repeat this statistical procedure many times on different samples drawn from the same populations, then the real population mean will fall in 95% of the cases within the confidence interval thus calculated. The confidence interval calculated for each sample will differ, but 19 out of 20 will contain the unknown population mean.

*** 11. Why have we used the t-statistic and not the z-statistic?**

Because we do not know the standard deviation σ of the population. So we rather use the t-statistic, which employs the standard deviation s of the sample, instead of the standard deviation σ of the population. But therefore the t-statistic follows a different sampling distribution (a t-distribution, and not a Normal-distribution).

*** 12. Suppose we know that the population standard deviation happens to be the same as the standard deviation of the sample. Determine the confidence interval using the z-statistic for this case.**

Then $z^* = 1.960$ (last row of Table D, worth remembering, very close to 2).
We use the following formula:

$$\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}} = 481.77 \pm 1.960 \cdot 17.544 = 481.77 \pm 34.39$$

So we obtain the interval [447.38; 516.16], which is slightly narrower than the interval obtained at question 9. This is so because the fact that we now know the standard deviation of the population allows us to use the z-test, whose critical value is slightly less (1.96, as opposed to 2.08), so we can be more certain about our conclusions. Knowing the standard deviation σ of the population helps us, and so the same confidence level can already be achieved by a narrower interval. Yet, the difference is not large, because the t-distribution with a degree of freedom of 21 is already very close to the standard Normal distribution.

*** 13. Now have the confidence interval calculated for you by SPSS. Copy the values returned by SPSS to your report. Is it different from your calculations?**

To get the confidence interval, it is best to set the Test Value to 0. I got the following table:

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	T	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
reading time per word	27,461	21	,000	481,773	445,29	518,26

I have to copy the last two columns to my report: 445.29 and 518.26, these are the lower and upper boundaries of the confidence interval. These are the same values as those I got at question 9, so I am very happy.

*** 14. Add again to your report the higher and lower values between which the population mean must lie. Why *and* how is this confidence interval different from the previously calculated one?**

I have changed the confidence level in the "Options" window. Do not forget to change it back afterwards…

One-Sample Test

	Test Value = 0					
					99% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
reading time per word	27,461	21	,000	481,773	432,10	531,45

The new interval is now [432.10, 531.45], which is larger than the 95% interval. This is so, because this time we wanted to be on the safe side: for the same set of observations, the interval must be larger in order to contain the population mean with a larger certainty, a larger "confidence".

*** 15. In each of the two cases, what is the null hypothesis exactly, and what is the alternative hypothesis? (In words/one full sentence, please.)**

The mean of our sample (482 ms) contradicts both the FRT ("Fast Reading Theory": at most 440 ms) and SRT ("Slow Reading Theory": at least 505 ms). The hypothesis we are going to test in both cases is that the theory is true and the extreme value of our sample is only due to chance.

FRT: Original theory: the population mean of RT is 440 ms *or less*.
Null hypothesis in statistical test: the population mean of RT is 440 ms.
Alternative hypothesis: the population mean is *more* than 440 ms.

SRT: Original theory: the population mean of RT is 505 ms *or more*.
Null hypothesis in statistical test: the population mean of RT is 505 ms.
Alternative hypothesis: the population mean is *less* than 505 ms.

For both theories we employ the most extreme value (exactly 440 ms/505 ms) still coherent with the theory (FRT/SRT). If our sample is consistent with this extreme value, then our simple will be consistent with the theory. (In other words, for example in the case of SRT, our sample will not refute the theory that RT can be 440 ms; even if the data may refute a more severe version of SRT that would claim that RT is at most 430 ms.)

On the other hand, if our sample provides sufficient argument to reject the null hypothesis for 440 ms (505 ms), it will even more so reject the more severe possibilities allowed for by the theories (RT lower than 440 ms/higher than 505 ms).

* 16. Are you using a one-sided or a two-sided test?

For both theories (SRT, FRT) we use a one-sided (one-tailed) *t*-test, as is obvious from the formulations of the alternative hypotheses above. Namely, the other tails of the distribution are in harmony with the original theories (FRT, SRT).

NB: we use a *t*-test, because we do not know the standard deviation of the population.

* 17. Perform the test for both cases by hand, and describe the steps of your calculation.

FRT: *First,* calculate *t*-statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{\bar{x}-\mu_0}{SE} = \frac{481.77-440}{17.544} = 2.381$.

SRT: *First,* calculate *t*-statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{\bar{x}-\mu_0}{SE} = \frac{481.77-505}{17.544} = \text{-}1.324$.

As the t-distributions are symmetric, we can use the absolute values of the t-values calculated (hence, 1.324 in the second case).

Second, we remember that *df = n-1 = 22-1 = 21.*

Third, we remember that we are performing a one-sided t-test with a significance level of 5%, whose critical *t** value is the same as the critical *t** for a two-sided test with $\alpha = 0.10$ or $C = 0.90$. That is what we have to look up in Table D. We find: $t^* = 1.721$ (for $df = 21$, and for $C = 90\%$ or upper tail probability = 0.05).

Fourth, we compare the calculated *t*-statistics to the critical t-value: 2.38 > 1.72 > 1.32. We can reject the null hypothesis if and only if $t \geq t^*$, that is, if our sample has a statistic that is as extreme as, or more extreme than the critical value.

Fifth step: we draw the conclusions:

FRT: We <u>have</u> sufficient evidence to reject the null hypothesis at the $\alpha = 0.05$ level.
SRT: We <u>do not have</u> sufficient evidence to reject the null hypothesis at the $\alpha = 0.05$ significance level.

Sixth step: see question 20.

*** 18. Let SPSS calculate the test for you and copy the results.**

FRT:

One-Sample Test

	Test Value = 440					
	t	df	Sig. (2-tailed)	Mean Difference	90% Confidence Interval of the Difference	
					Lower	Upper
reading time per word	2,381	21	,027	41,773	11,58	71,96

Note that the *p*-value of a one-tailed test is the half of a two-tailed test given by SPSS. Therefore, we conclude that $p = 0.014$, which is below the 5% significance level.

SRT:

One-Sample Test

	Test Value = 505					
	t	df	Sig. (2-tailed)	Mean Difference	90% Confidence Interval of the Difference	
					Lower	Upper
reading time per word	-1,324	21	,200	-23,227	-53,42	6,96

Note that the *p*-value of a one-tailed test is the half of a two-tailed test given by SPSS. Therefore, we conclude that $p = 0.100$, which is still above the 5% significance level.

*** 19. What is the meaning of the P-value in each of the two cases? (Hint: the probability of exactly what is it?) Please write one full sentence for each case in your report.**

FRT: The *p*-value is the probability of drawing a simple random sample (SRS) whose sample mean is 481.77 ms or higher (whose t-statistic is 2.381 or higher), *providing that* the population mean is 440 ms.

SRT: The *p*-value is the probability of drawing a simple random sample (SRS) whose sample mean is 481.77 ms <u>or lower</u> (whose t-statistic is -1.324 <u>or lower</u>), *providing that* the population mean is 505 ms.

*** 20. For each case, provide *the* key sentence summarizing the results of the statistical analysis, as it is done in scientific papers. That is, either "*Based on our data, we can reject the null-hypothesis at a significance level alpha = 0.05, that is, we can conclude that [the alternative hypothesis in words] is true (t = ..., df = ..., P = ...)*" or "*our data do not provide sufficient evidence to reject the null hypothesis, that is, to conclude that...*".**

FRT:
Based on our data, we can reject the null hypothesis at a 0.05 significance level that the RT (reading time) is 440 ms or lower (t = 2.38, df = 21, p = 0.014). We conclude that the Fast Reading Theory of the phenomenon being examined cannot be true, and we apply to the Scientific Research Fund to provide us with money to come up with a better theory.

SRT:
Our experiments do not provide sufficient evidence to reject the Slow Reading Theory of the phenomenon being examined (t = -1.32, df = 21, p = 0.10). However, our data suggest that RT cannot be much higher that 505 ms, and might even be less than 505 ms; a larger sample could refute the SRT hypothesis. Thus, more experiments are needed to approximate better the real value of RT, and so we urge the Scientific Research Found to provide us with more money.