**Exercise 1**

**Item 1.1**

Answer: The resolution is twofold.

Two Bootstrap tests will be executed to demonstrate whether the data could stem from an exponential distribution with lambda between 0.01 and 0.1.

Having a lambda between 0.01 and 0.1 in an exponential distribution means that the mean for this population would be contained in the interval 1/(0.1) -> 10 and 1/(0.01) -> 100 [10;100].

Thus, the first test will consist of a null hypothesis that lambda = 0.1 (lowerBound ==> mean = 10) with alternative hypothesis that lambda < 0.1 (mean > 10), the idea is to demonstrate that given this alternative hypothesis we reject the null hypothesis in favor of the alternative hypothesis (lambda < 0.1).

The second test will use null hypothesis that lambda = 0.01 (upperBound) with alternative hypothesis that lambda > 0.01 (mean < 100) the idea is to demonstrate that given this alternative hypothesis we reject the null hypothesis in favor of the alternative

hypothesis (lambda > 0.01).

*First test*

Generating the simulations, creating simulated samples Xstars and computing the statistic over that simulation Tstars considering H0 a distribution with lambda = 0.1 (mean = 10) and H1 lambda < 0.1 (mean > 10).

Lambda = 0.1 means that the exp. distribution has an average 1/0.1 = 10 with a sd. = sqr(1/(0.1^2) = 10.

Calculating p-value for this test.

*pLowerBound=sum(tstarLowerBound>t)/B*

P-value upon which we would reject (or fail to reject) the null hypothesis.

This is the probability of computing the statistic with values greater than the observed.

Grater because our alternative hypothesis is that the mean is greater than 10.

As can be seen, p is very low (0), so we reject the null hypothesis in favor of the alternative hypothesis.

*Second test*

Now generating another simulation but considering different null and alternative hypothesis:

H0 distribution is a exp distribution with lambda = 0.01 (mean =100) and H1 lambda is > 0.01 (mean < 100)

Lambda = 0.01 means that the exp. distribution has an average 1/0.01 = 100 with a sd. = sqr(1/(0.01^2) = 100.

*pUpperBound=sum(tstarUpperBound<t)/B*

P-value upon which we would reject (or fail to reject) the null hypothesis.

This is the probability of computing the statistic with values less than the observed.

Less because our alternative hypothesis is that the mean is lower than 100.

As can be seen, p is very low(0), so we reject the null hypothesis in favor of the alternative hypothesis.

**Conclusion:** Given the two tests above is very plausible that lambda is located between 0.01 and 0.1 (in other words, that the mean is located between 10 and 100.

**Item 1.2:**

We generate the bootstrap interval, using the median as location estimator.

Bootstrap interval for the mean of this population with 95% confidence.

97.5% 2.5%

38.12714 48.77792

**Conclusion:** Considering the bootstrap confidence interval, the mean consumption for this population falls in the interval between 38.12714 and 48.77792 (this is in line with the hypothesis test from the previous item), so the marketing manager could potentially focus on consumers within this range.

**Exercise 2**

1. Given the boxplot of the light1879.txt dataset, we can observe small number of outliers and the histogram is quite symmetric. We can presume that this dataset follows a normal distribution. In contrast, 1882 dataset might be not sampled normal distribution as we can observe some outliers in the boxplot diagram and the histogram is right skewed. The boxplot and histogram of last dataset - light.txt, seems to be symmetric but we can see some significant outliers in the first diagram. To evaluate the normality, we additionally used Shapiro-Wilk normality test. With p- value = 2.724e-12, we can reject the null hypothesis is that this dataset follows normal distribution.

|  |  |  |
| --- | --- | --- |
| Dataset | Mean confidence interval | Median confidence interval |
| light1879.txt | [299836.6, 299869.3] | [299830, 299860] |
| light1882.txt | [299713.2, 299800.9] | [299752, 299837] |
| light.txt | [299725.9, 299790.6] | [299742.2, 299772.4] |

1. As seen from the table above, the median confidence intervals of all datasets are smaller the mean confidence intervals respectively, which means that median confidence intervals can have more accurate estimation. This result is fairly reasonable as the estimating statistic x̄ is not robust against outliers that are located in all boxplots.
2. As mentioned above, the light1879.txt dataset can be sampled from normal distribution population. To test whether the result of this measurement is consistent with the currently most accurate value of the speed of light (299792.458), we use t-test with µ= 299792.458 (km/s). With p-value = 4.451e-06 < 0.05, we can reject the null hypothesis that the true mean of speed of light in this experiment is equal to 299792.458. Both light1879.txt and light.txt dataset do not follow the normal distribution so we apply Wilcoxon signed rank test those data. In the former test, the p-value = 0.09736 which means that we can’t reject the null hypothesis that the true mean is equal 299792.458, while the later test with p-value = 4.451e-06 indicates that we can reject the same null hypothsis.

**Exercise 3**

1. According to the QQ-plot of KLM data, the distribution is not taken from normal population. For skewed distribution, the mean is highly influenced by the high/low values. In such cases, it is better to test location in terms of the median, instead of the mean. Therefore, the sign test is the best choice which prerequisites are all satisfied. The data are a random sample from a population with a certain median m. we test null hypothesis . The test statistics is which has the binary (N,0.5)- distribution under . Conclusion: As the p-value is 0.01, we can reject the null hypothesis. It means that the median of this population is bigger than 31.
2. This test is similar to sign test except the probability of success in binary distribution. we test null hypothesis . The test statistics is which has the binary (N,0.1)- distribution under . Conclusion: As the p-value is 0.007, we can reject the null hypothesis. It means that more than 10% of the parts arrives after the maximum delivery period of 72 days.

**Exercise 4**

1. Given the histograms and QQ-plots of seeded and unseeded cloud data, we can't assume that they are in a normal distribution. As a result, we can't trust the result of the two samples t-test in this case because the assumption of the normal distribution in t-test was violated. On the other hand, the Mann- Whitney and the Kolmogorov-Smirnov test can be adopted in this case for the reason that both don't assume observations are from normal distribution. Mann- Whitney test has p-value = 0.01383 < 0.05, then we can reject that the populations of two samples are the same. Similarly, Kolmogorov-Smirnov test has p-value = 0.01905 < 0.05 so we can draw the same conclusion.
2. As we can't assume that the square root of seeded and unseeded cloud data is normal distribution. The assumption of the normal distribution in the two samples t-test was violated so we shouldn't apply the two samples t-test to the data. Mann- Whitney test and the Kolmogorov-Smirnov test can be adopted in this case for the reason that both don't assume observations are from normal distribution. The former test, Mann-Whitney, has p-value = 0.01383 < 0.05, we can conclude that of same populations is rejected. The later test, Kolmogorov-Smirnov, has p-value = 0.01905 < 0.05 so we can also reject the null hypothesis that the populations of two samples are the same.
3. After transformed by square root of the square if the values, the data for both seeded and unseeded clouds seems to follow the normal distribution which leads to the validity of the two samples t-test in the case. With p-value = 0.0124 < 0.05 in the two samples t-test, we can reject the null hypothesis H0 : μ = ν that the means of two populations are the same. The “p-value” values of Mann-Whitney and Kolmogorov-Smirnov test remain the same in comparison with two previous cases (Question 1 and 2), and therefore we have the identical conclusions in those tests.

**Exercise 5**

1. Based on the pairs is possible to see a potential correlation in the following pairs: migration x age, migration x weight, migration x wrist. This conclusion is based on the fact that the plot from these pairs resembles a straight line passing through the origin.
2. Tests will be conducted using Spearman rank correlation test which doesn't assume normality between the two variables.

* Test 5.2.1 (migration x age): As can be seen from the test, p-value = 0.0021, which led us to reject the null hypothesis that rho is equal to 0. In fact, the calculated rho based on the samples is 0.4760.
* Test 5.2.2 (migration x weight): As can be seen from the test, p-value = 0.02861, which led us to reject the null hypothesis that rho is equal to 0. In fact, the calculated rho based on the samples is 0.3506.
* Test 5.2.3 (migration x length): As can be seen from the test, p-value = 0.6087, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence level.

The calculated rho based on the samples is 0.0845 which is very close to 0 indicating a weak correlation between the variables.

* Test 5.2.4 (migration x wrist): As can be seen from the test, p-value = 0.1797, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence.

The calculated rho for the sample is rho = 0.2193.

* Test 5.2.5 (migration x diastolic): As can be seen from the test, p-value = 0.6494, which led us to fail to reject the null hypothesis that rho is equal to 0 - considering a 0.05 confidence level.

The calculated rho based on the samples is 0.0751 which is very close to 0 indicating a weak correlation between the variables.

**Exercise 6**

1. To study the data, we plot the running time before and after both soft and energy drink separately. It includes QQ-plot, histogram and box plot of all of them. The median of running time increased after soft drink. Moreover, the running time after drinking spread in larger range rather than before drinking. As the sample size is really small, it is hard to conclude the normality of population. QQ-plot of before and after for soft drink demonstrates the normality of population while the histograms contradicts this idea. In case of energy drink, the histograms and QQ-plots prove that the sample is not taken from normal population for both before and after energy drink. The median of running time decreased after energy drink. Additionally, the running time after drinking spread in larger range than before.
2. Since we can’t assume normality for both soft drink and energy, we should apply permutation test for two cases separately. We generate 1000 randomly chosen permutations to estimate the distribution of our test statistic under . Conclusion: In case of soft drink, since the p-value is about 0.16 we can’t reject null hypothesis. Therefore, soft drink doesn’t affect the running time. In terms of energy, the p-value is 0.29 and still greater than critical value. So, it means that energy drink can’t influence running time either.
3. In such case, we have two different groups without normal population. Therefore, we apply Man-Whitney test. The sample stems from soft drink population (S) and similarly originates from energy drink (E). We test null hypothesis that the populations are the same. Conclusion: of equal means is not rejected. The underlying distribution of time difference in soft drink is similar to energy drink and drink type doesn’t affect running time.
4. Time differences (after - before) are negative numbers for some cases. It may because of dependency of two measurement (before and after) in a short period of time. Additionally, the sample size for both cases seems small. The distribution of time difference is not normal in soft and energy drink and we can’t use t-samples t-test. **Are these correct?**
5. Yes, we have similar objections here.
6. If we apply t-test in part 3, we should assume normality of time difference in both drink types. Instead, we can apply Man-Whitney test. **(I don’t understand the question)**

**Exercise 7**

1. According to the QQ-plot of these three drugs, we just can assume samples in drug 3(cyclopropane) are taken from normal population. Drug 1 and 2 are certainly not normal.
2. The estimated concentration for drug 1 is equal 0.4340. drug 2 and drug 3 are 0.399 and 0.015 respectively. Since the p-value is 0.11, we reject null hypothesis which means for some (i, j).
3. In this part, the p-value is 0.059. Therefore, the null hypothesis in not rejected (but not with a great difference). . As we concluded in first part, the samples are not taken from normal population. However, ANOVA assumes normality to test the data. Therefore, the conclusion of ANOVA is against Kruskal-Wallis result.