Mathematical Statistics Exam, September 10th, 2019, Set A

Fill in the dotted spaces [“..........”]. 1 question (•) = 1 point; maximum = 16 points. Only responses in the specified places will be checked, but **you need to include your notes with calculations when you return your exam**. Fill in your responses **after having verified them**; if illegible or larded with corrections and crossings-out, the answers will be treated as wrong. You can use a simple calculator, statistical tables and one a4 sheet of paper with helpful formulas. **Communication with the rest of the world is not allowed**.

NAME: .... student’s number ....

Signature ....

1. The changes in the number of users of a streaming service provider for the years 2014-2018 are described by certain characteristics in the table below:

Users \ Year	2014	2015	2016	2017	2018
Number of male users index, 2014 as the base year	1	1.2	1.8	2.4	3
Number of female users, in millions	40	50	60	80	100

   - The average yearly growth rate for the years 2014-2018 for the number of male users amounts to ...., and for the number of female users amounts to ....
   - Assuming that in 2018 the number of male users amounted to 90 million and that the average growth rates calculated above will continue into the future, we can predict that:
     - the number of male users in 2019 will amount to ....
     - the number of female users in 2019 will amount to ....
     - the number of male users will surpass the number of female users in the year .... /NEVER (fill in the blank or underline the appropriate).

2. A streaming service provider wishes to determine whether the length of a free trail period translates to a tendency in paid subscriptions. The data from a randomized sample is summarized in the table below:

Length of free trial period offered, in months	1	2	3	4
Number of users who continue with paid subscription	60	70	50	50
Number of users who do not continue with paid subscription	40	30	50	50
Number of users who were offered such trial length	100	100	100	100

   - We conduct a chi-squared test of independence to determine whether the length of the free trial period determines the willingness to continue with a paid subscription. The value of the appropriate chi-squared test statistic amounts to ...., the critical REGION for a 5% significance test is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis of the lack of influence of the length of the free trial period on paid subscriptions.

- We conduct a simple test to compare whether the fractions of users who continue with a paid subscription are the same for the one month and two months free trial periods. The value of the appropriate test statistic amounts to ...., the critical VALUE for a 5% significance test is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis of the equality of fractions.

3. Based on data presented in the table above, we construct confidence intervals for the fraction of users who continue and do not continue with a paid subscription after a free trial period.

   - Looking at the group of individuals who were offered a one month free trial period, the realization of a 95% confidence interval for the fraction who continue with a paid subscription is ....
   - Looking at the group of individuals who were offered a one month free trial period, the realization of a 95% confidence interval for the fraction who did not continue with a paid subscription is .... This confidence interval is WIDER /NARROWER /OF THE SAME LENGTH (underline the appropriate) as the confidence interval calculated in the previous point (•), and this result is A COINCIDENCE /A RULE (underline the appropriate).

4. A streaming service provider granted access to its services in a new country, where the fraction of individuals in the population who speak English is lower than average and wishes to determine whether this fact has impact on the number of movies that the users watch. In the new country, in a sample of 36 new subscribers, the average monthly number of movies watched was 6.4. In general, the streaming service provider observed on average 8 movies watched per user per month, with a standard deviation equal to 3. We make a simplifying assumption that the number of movies watched follows a normal distribution.

   - The value of the appropriate test statistic is equal to ...., the p-value of this result is equal to ...., so for a 10% significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis that the average for the new users is the same as in general.
   - Additionally, it was calculated that the value of the standard deviation in the new country sample amounted to 2.7. We verify whether the variance of the number of movies watched in the new country is the same as in the general population for a significance level of 10%. The value of the appropriate test statistic amounts to .... The critical VALUE of the appropriate test amounts to .... and comes from a distribution with .... degrees of freedom. Based on these results, we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis.

5. In some countries, a streaming service provider offers an additional option of movie subtitles in the native language of the subscribers, but in some countries this feature is not available. A survey conducted with random 1000 users in each of the two groups of countries showed that in countries where native language subtitles are available, the average amount of time spent on receiving streams amounts to 300 minutes weekly, with a standard deviation equal to 50 minutes, while in countries where the native language subtitles are not available, the average amount of time spent amounts to 200 minutes, with a standard deviation equal to 40 minutes.

- We conduct a test for the equality of mean stream duration for the two groups, against the alternative that in the group without subtitles the average is lower. The value of the appropriate test statistic is equal to ...., the critical REGION of the appropriate test for a 1% significance level is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis.
- Let us assume that the same values of averages and standard deviations were obtained for sample sizes of 10000 users in each of the two groups, respectively. In such a case, the value of the test statistic calculated for the larger sample would be HIGHER /LOWER /THE SAME (underline the appropriate) as in the previous point, while the p-value of the test statistic calculated for the larger sample would be HIGHER /LOWER /THE SAME (underline the appropriate).

6. Let us assume that X, the time (in minutes) spent on watching streamed movies, follows a distribution with density $f(x) = \frac{1}{\lambda}e^{-\frac{x-c}{\lambda}}$ for $x > c$ and 0 otherwise, where $\lambda > 0$ is an unknown parameter and c is a constant which is known to the researcher (it differs for various countries). We have that for this distribution, $EX = c + \lambda$ and $VarX = \lambda^2$. Let us assume that we have a sample of n observations: $X_1, X_2, \ldots, X_n$ from this distribution.

   - The method of moments estimator for λ, based on the mean, is equal to $\hat{\lambda}_M =$ ....
   - The bias of the estimator $\hat{\lambda}_M$ is equal to $b =$ ...., the variance of this estimator is equal to $Var =$ ...., and the MSE of this estimator is equal to $MSE =$ ....

7. We continue with the assumptions from the previous problem.

   - The maximum likelihood estimator for λ for a sample size of n is equal to $\hat{\lambda}_{ML} =$ ....
   - The Fisher information connected with a single observation from the distribution provided is equal to ...., so for the whole sample it amounts to ....

8. Continuing with the assumptions presented in problem 6, let us assume that the researcher has two independent random samples available: $X_1, X_2, \ldots, X_{10}$ for country A, where the constant c is equal to 15, and $Y_1, Y_2, \ldots, Y_{20}$ for country B, where the constant c is equal to 30. The researcher constructs three estimators of the unknown parameter λ:

$$\hat{\lambda}_1 = \frac{X_1 + X_2 + \ldots + X_{10}}{10} - 15, \; \hat{\lambda}_2 = \frac{Y_1 + Y_2 + \ldots + Y_{20}}{20} - 30, \text{ and}$$

$$\hat{\lambda}_3 = \left(\frac{X_1 + X_2 + \ldots + X_{10}}{10} - 15\right) \cdot \frac{1}{2} + \left(\frac{Y_1 + Y_2 + \ldots + Y_{20}}{20} - 30\right) \cdot \frac{1}{2}$$

.

   - From among the three estimators, the estimator with the lowest bias is: λ_1 /λ_2 /λ_3 (underline ALL the appropriate)
   - From among the three estimators, the estimator with the lowest variance is λ_1 /λ_2 /λ_3 (underline ALL the appropriate), so using the standard criteria of estimator quality we conclude that from among the three presented, the best estimator is λ_1 /λ_2 /λ_3/IT IS IMPOSSIBLE TO DETERMINE (underline ALL the appropriate).

Mathematical Statistics Exam, September 10th, 2019, Set B

Fill in the dotted spaces [".........."]. 1 question (•) = 1 point; maximum = 16 points. Only responses in the specified places will be checked, but **you need to include your notes with calculations when you return your exam**. Fill in your responses **after having verified them**; if illegible or larded with corrections and crossings-out, the answers will be treated as wrong. You can use a simple calculator, statistical tables and one a4 sheet of paper with helpful formulas. **Communication with the rest of the world is not allowed**.

NAME: .... student's number ....

Signature ....

1. A streaming service provider wishes to determine whether the length of a free trail period translates to a tendency in paid subscriptions. The data from a randomized sample is summarized in the table below:

Length of free trial period offered, in months	1	2	3	4
Number of users who continue with paid subscription	50	60	70	60
Number of users who do not continue with paid subscription	50	40	30	40
Number of users who were offered such trial length	100	100	100	100

   - We conduct a chi-squared test of independence to determine whether the length of the free trial period determines the willingness to continue with a paid subscription. The value of the appropriate chi-squared test statistic amounts to ...., the critical VALUE for a 1% significance test is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis of the lack of influence of the length of the free trial period on paid subscriptions.
   - We conduct a simple test to compare whether the fractions of users who continue with a paid subscription are the same for the one month and two months free trial periods. The value of the appropriate test statistic amounts to ...., the the p-value for this result is equal to ...., so for a 1% significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis of the equality of fractions.

2. Based on data presented in the table above, we construct confidence intervals for the fraction of users who continue and do not continue with a paid subscription after a free trial period.

   - Looking at the group of individuals who were offered a two months free trial period, the realization of a 90% confidence interval for the fraction who continue with a paid subscription is ....
   - Looking at the group of individuals who were offered a two months free trial period, the realization of a 90% confidence interval for the fraction who did not continue with a paid subscription is .... This confidence interval is WIDER /NARROWER /OF THE SAME LENGTH (underline the appropriate) as the confidence interval calculated in the previous point (•), and this result is A COINCIDENCE /A RULE (underline the appropriate).

3. The changes in the number of users of a streaming service provider for the years 2014-2018 are described by certain characteristics in the table below:

Users \ Year	2014	2015	2016	2017	2018
Number of male users, in million	30	40	50	60	80
Number of female users index, 2014 as the base year	1	1.2	1.6	2.4	3.2

- The average yearly growth rate for the years 2014-2018 for the number of male users amounts to ...., and for the number of female users amounts to ....
- Assuming that in 2018 the number of female users amounted to 75 million and that the average growth rates calculated above will continue into the future, we can predict that:
  - the number of male users in 2019 will amount to ....
  - the number of female users in 2019 will amount to ....
  - the number of female users will surpass the number of male users in the year .... /NEVER (fill in the blank or underline the appropriate).

4. Let us assume that Y, the time (in minutes) spent on watching streamed movies, follows a distribution with density $f(y) = \frac{1}{\mu}e^{-\frac{y-a}{\mu}}$ for $y > a$ and 0 otherwise, where $\mu > 0$ is an unknown parameter and a is a constant which is known to the researcher (it differs for various countries). We have that for this distribution, $EY = a + \mu$ and $VarY = \mu^2$. Let us assume that we have a sample of n observations: $Y_1, Y_2, \ldots, Y_n$ from this distribution.

- The method of moments estimator for μ, based on the mean, is equal to $\hat{\mu}_M =$ ....
- The bias of the estimator $\hat{\mu}_M$ is equal to $b =$ ...., the variance of this estimator is equal to $Var =$ ...., and the MSE of this estimator is equal to $MSE =$ ....

5. We continue with the assumptions from the previous problem.

- The maximum likelihood estimator for μ for a sample size of n is equal to $\hat{\mu}_{ML} =$ ....
- The Fisher information connected with a single observation from the distribution provided is equal to ...., so for the whole sample it amounts to ....

6. Continuing with the assumptions presented in problem 4, let us assume that the researcher has two independent random samples available: $X_1, X_2, \ldots, X_{20}$ for country A, where the constant a is equal to 10, and $Y_1, Y_2, \ldots, Y_{20}$ for country B, where the constant a is equal to 30. The researcher constructs three estimators of the unknown parameter μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + \ldots + X_{20}}{20} - 10, \; \hat{\mu}_2 = \frac{Y_1 + Y_2 + \ldots + Y_{20}}{20} - 30, \text{ and}$$

$$\hat{\mu}_3 = \left(\frac{X_1 + X_2 + \ldots + X_{20}}{20} - 10\right) \cdot \frac{1}{2} + \left(\frac{Y_1 + Y_2 + \ldots + Y_{20}}{20} - 30\right) \cdot \frac{1}{2}$$

.

- From among the three estimators, the estimator with the lowest bias is: μ_1 /μ_2 /μ_3 (underline ALL the appropriate)
- From among the three estimators, the estimator with the lowest variance is μ_1 /μ_2 /μ_3 (underline ALL the appropriate), so using the standard criteria of estimator quality we conclude that from among the three presented, the best estimator is μ_1 /μ_2 /μ_3/IT IS IMPOSSIBLE TO DETERMINE (underline ALL the appropriate).

7. A streaming service provider granted access to its services in a new country, where the fraction of individuals in the population who speak English is lower than average and wishes to determine whether this fact has impact on the number of movies that the users watch. In the new country, in a sample of 25 new subscribers, the average monthly number of movies watched was 12.5, with a standard deviation (based on the unbiased estimator of the variance) of 3. In general, the streaming service provider observed on average 16 movies watched per user per month. We make a simplifying assumption that the number of movies watched follows a normal distribution.

   - The value of the appropriate test statistic is equal to ...., the critical VALUE of the appropriate test statistics for a 5% significance level is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis that the average for the new users is the same as in general.
   - Additionally, we know that the value of the standard deviation in the general population amounts to 4. We verify whether the variance of the number of movies watched in the new country is the same as in the general population, against the alternative that it is lower, for a significance level of 5%. The value of the appropriate test statistic amounts to .... The critical REGION of the appropriate test amounts to .... and comes from a distribution with .... degrees of freedom. Based on these results, we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis.

8. In some countries, a streaming service provider offers an additional option of movie subtitles in the native language of the subscribers, but in some countries this feature is not available. A survey conducted with random 1000 users in each of the two groups of countries showed that in countries where native language subtitles are available, the average amount of time spent on receiving streams amounts to 400 minutes weekly, with a standard deviation equal to 60 minutes, while in countries where the native language subtitles are not available, the average amount of time spent amounts to 360 minutes, with a standard deviation also equal to 60 minutes.

   - We conduct a test for the equality of mean stream duration for the two groups. The value of the appropriate test statistic is equal to ...., the critical REGION of the appropriate test for a 10% significance level is equal to ...., so for this significance level we REJECT /DO NOT HAVE GROUNDS TO REJECT (underline the appropriate) the null hypothesis.
   - Let us assume that the same values of averages and standard deviations were obtained for sample sizes of 100 users in each of the two groups, respectively. In such a case, the value of the test statistic calculated for the smaller sample would be HIGHER /LOWER /THE SAME (underline the appropriate) as in the previous point, while the p-value of the test statistic calculated for the smaller sample would be HIGHER /LOWER /THE SAME (underline the appropriate).