Statistics for Market Research exams

Angry Teachers, Folklore

January 6, 2023

Contents

Description

See updates at https://github.com/bdemeshev/stat4mr_exams.
Click on red hyperlinks inside pdf, you can get to the answers and back!
Any comments? Bugs? https://github.com/bdemeshev/stat4mr_exams/issues/.
The order of topics has changed after the first course iteration in 2020-21. The interested reader may find relevant exercises by looking through all 2020-21 exams.

Greatings to the contributors

Here we describe only the style guidelines and typical erros. For more information on tex one may read the book by K. Vorontsov.

1. Use decimal point as a separator: 3.14 — good style, $3{,}14$ — bad style. This goes against russian tradition, but favors copy-pasting numbers in software for computations.
2. Use `\[...\]` for display math formulas. Do not use $$...$$!
3. Use `cases` for systems of equations, `align*` for multiline formulas, `enumerate` for enumerations.
4. Inside formulas use `\text{...}` to write text.
5. Use `\ldots` for ellipsis.
6. You can find useful macros in the preamble, like `\P, \E, \Var, \Cov, \Corr, \cN`.
7. Use backslash before functions: `\ln, \exp, \cos`...
8. Use booktabs style for tables. You may use online tablesgenerator. Choose booktabs table style instead of default table style.
9. Respect the letter ë! :)
10. Start every sentence in tex source from a new line. There will be no additional newlines in final pdf but tex file will be easier to read.
11. For multiplication use `\cdot`. Please never use `*` :)

1 2022-2023

1.1 2022-12-22 demo

Rules: two parts, two hours in total, one A4 cheat sheet is allowed.

Part 1. Test part, only numerical answers are checked, 6 questions, each question gives 1 point but no more than 4 points in total. This part is very predictable :)

1. (bootstrap) I have a sample X_1, ..., X_{100}.

   I generate one naive bootstrap sample X_1^*, ..., X_{100}^*.

   What is the probability that the first observation will be present in the bootstrap sample 2 times or more?

2. (welch) We have data for an AB-experiment $\bar{X}_a = 10$, $\bar{X}_b = 12$, $n_a = 20$, $n_b = 30$, $\sum(X_i^a - \bar{X}_a)^2 = 100$, $\sum(X_i^b - \bar{X}_b)^2 = 200$.

   Calculate the standard error of $\bar{X}_a - \bar{X}_b$ for the Welch test.

3. (mw test) I have five results of two runners A and B for the 5 km race: 25:12 (A), 26:34 (B), 27:43 (A), 28:12 (A), 29:05 (B).

   Calculate Mann-Whitney statistic U_A that tests the null-hypothesis of equal distributions of time.

   (The statistic U_A should positively depend on the ranks of the runner A).

4. (multiple comparison) I have 100 hypothesis with independent statistics. The null hypothesis for all 100 cases is actually true, but I don't know this.

   I calculate all p-values. If the two lowest p-value are both lower than 0.05 I wrongly conclude that not all H_0 are true. Otherwise I correctly conclude that all H_0 are true.

   What is the probability that I will get the correct conclusion?

5. (sample size) My target variable is binary and I wish minimal detectable effect equal to 0.01, probability of I-error not greater than 0.02, probability of II-error not greater than 0.10, control and experimental group of the same size equal to n.

   What is minimal value of n?

6. (anova 1+2) Vasiliy loves to eat shaurma. He has three local shaurma dealers. Vasiliy bought 7 shaurmas from each dealer. and measured their weight. He would like to test the hypothesis that mean weight is the same for all dealers.

   Total sum of squares is 1000, between sum of squares is 500.

   Calculate the F-statistic to test the hypothesis.

Part 2. Open part, solutions are required, 4 problems, each problem gives 2 points but no more than 6 points in total. This part is almost unpredictable :)

1. Let random variables Y_1, ..., Y_n be iid uniform $U[0;1]$. Consider the naive bootstrap sample Y_1^*, ..., Y_n^*.
   Find $\text{Var}(Y_1^*)$, $\text{Cov}(Y_1^*, Y_2^*)$, $\text{Var}(\bar{Y}^*)$.

2. Winnie-the-Pooh simulteneously tests h null hypothesis using independent samples. All the null hypothesis are true but Winnie does not know it.

   (a) What is the probability that the highest P-value will be greater than 0.95?
   (b) What is the possible range for the probability in point (a) if exactly one null hypothesis is false?

3. The correlation matrix of standardized variables a, b and c is given by

$$C = \begin{pmatrix} 1 & 0.2 & 0 \\ & 1 & 0.2 \\ & & 1 \end{pmatrix}$$

   Let p_1, p_2 and p_3 be the principal components.

   (a) Express p_1 in terms of a, b and c.
   (b) Express b in terms of p_1, p_2 and p_3.
   (c) How would you restore the second observation of variable b if you know that first and second components for the second observation are equal to -1 and 2 respectively?

4. Consider the Mann-Whitney test with possible ties. The variables X_1, X_2, ..., X_{n_x} are iid Poisson with rate $\lambda = 1$. The variables Y_1, Y_2, ..., Y_{n_y} are iid Poisson with the same rate, independent from X sample.

   Let L be the number of all pairs (X_i, Y_j) such that $X_i > Y_j$.

   (a) Find $\mathbb{E}(L)$, $\text{Var}(L)$.
   (b) What is the probability that the ordered sequence of all X_i and Y_j will start with three or more members from X-sample?

1.2 2022-12-22 exam

Part 1. Test part, only numerical answers are checked, 6 questions, each question gives 1 point but no more than 4 points in total.

1. I have a sample $X_1, \dots, X_{90}$. I generate one naive bootstrap sample $X_1^*, \dots, X_{90}^*$.

   What is the probability that the first observation will be present in the bootstrap sample exactly 3 times?

2. We have data of an AB experiment: $\bar{X}_a = 5.4, \bar{X}_b = 6, n_a = 18, n_b = 15, \sum(X_i^a - \bar{X}_a)^2 = 890, \sum(X_i^b - \bar{X}_b)^2 = 800$.

   Calculate the estimate of variance of $\bar{X}_a - \bar{X}_b$ for the Welch test.

3. I have five results of two runners A and B for the 5 km race:

   16:49 (B), 21:17 (A), 18:30 (B), 6:18 (B), 20:16 (A), 15:39 (B).

   Calculate Mann-Whitney statistic U_A that tests the null-hypothesis of equal distributions of time.

   (The statistic U_A should positively depend on the ranks of the runner A).

4. I have 30 hypothesis with independent statistics. The null hypothesis for all 30 cases is actually true, but I don't know this.

   I calculate all p-values. If the 4 lowest p-value are simultaneously lower than 0.01 I wrongly conclude that not all H_0 are true. Otherwise I correctly conclude that all H_0 are true.

   What is the probability that I will get the correct conclusion?

5. My target variable is binary and I wish minimal detectable effect equal to 0.04, probability of I-error not greater than 0.01, probability of II-error not greater than 0.2. The control and experimental group are of the same size equal to n.

   Which minimal value of n is sufficient in the worst case?

6. Vasiliy loves to eat shaurma. He has 5 local shaurma dealers. Vasiliy bought 4 shaurmas from each dealer and measured their weight. He would like to test the hypothesis that mean weight is the same for all dealers.

   Total sum of squares is 600, between sum of squares is 100 Calculate the F -statistic to test the hypothesis.

Part 2. Open part, solutions are required, 4 problems, each problem gives 2 points but no more than 6 points in total.

7. Let random variables Y_1, ..., Y_n be iid uniform $U[0; 1]$. Consider the naive bootstrap sample Y_1^*, ..., Y_n^*.
   Find $\mathbb{P}(Y_1^* = Y_1)$, $\text{Cov}(Y_1^*, Y_2)$, $\mathbb{P}(\max\{Y_1, \dots, Y_n\} = \max\{Y_1^*, \dots, Y_n^*\})$.

8. The eigenvalues of sample correlation matrix are 2.5, 0.3 and 0.2. The eigenvalue $\lambda = 2.5$ corresponds to eigenvector $v = (3, 4, -2)$.

   (a) James Bond predicts every original variable using multivariate regression on the first two components. What is the average value of R^2 he will get?

   (b) Express the first principal component in terms of scaled original variables a, b and c.

9. Winnie-the-Pooh simulteneously tests h null hypothesis using independent samples. All the null hypothesis are true but Winnie does not know it.

   (a) What is the expected value of the lowest P-value?

   (b) What is the expected number of wrongly regected hypothesis if Winnie rejects all the hypothesis with P-value less 0.1?

10. There are three continuosly distributed samples of the same size n, X_1, ..., X_n, Y_1, ..., Y_n, Z_1, ..., Z_n. Imagine that the null hypothesis that all samples have the same distribution is true.

    Consider the random variable R_X — the sum of ranks of the X sample in the pooled sample.

(a) Find the expected value $\mathbb{E}(R_X)$.

(b) What is the probability that R_X will be equal to $n(n+1)/2$?

1.3 2022-12-22 demo solutions

1.4 2022-12-22 exam solutions

1. (a) (2 points) $\mathbb{P}(Y_1^* = Y_1) = 1/n$

   (b) (4 points)

$$\mathrm{Cov}(Y_1^*, Y_2) = \mathrm{Cov}(I_1 Y_1 + \ldots + I_n Y_n, Y_2) = \mathrm{Cov}(I_2 Y_2, Y_2);$$

$$\mathrm{Cov}(I_2 Y_2, Y_2) = \mathbb{E}(I_2 Y_2^2) - \mathbb{E}(I_2 Y_2)\,\mathbb{E}(Y_2) = \frac{1}{n}\,\mathrm{Var}(Y_2) = \frac{1}{12n}$$

   (c) (4 points)

$$\mathbb{P}(\max\{Y_1^*, \ldots, Y_n^*\} = \max\{Y_1, \ldots, Y_n\}) = 1 - \left(\frac{n-1}{n}\right)^n \approx 1 - e^{-1}$$

2. (a) (5 points)

$$\text{average } R^2 = \frac{2.5 + 0.3}{2.5 + 0.3 + 0.2} = \frac{28}{30} \approx 0.93$$

   (b) (5 points)

$$p = \frac{3}{\sqrt{29}}a + \frac{4}{\sqrt{29}}b - \frac{2}{\sqrt{29}}c$$

3. (a) (5 points) One possible solution: obtain the density of minimal p-value (3 points) and calculate expected value (2 poitns):

$$f(m) = h(1-m)^{h-1}, \text{ where } m \in [0;1]$$

$$\mathbb{E}(M) = \int_0^1 m f(m) dm = \frac{1}{h+1}$$

   (b) (5 poinst) Note that $N_{\text{wrong}} \sim \mathrm{Bin}(h, 0.1)$. Hence $\mathbb{E}(N_{\text{wrong}}) = 0.1h$.

4. (a) (5 points)

$$\mathbb{E}(R_X) = n\,\mathbb{E}(R(X_1)) = n\frac{1+3n}{2}$$

   (b) (5 points)

$$\mathbb{P}(\text{X-obs are before other obs}) = \frac{1}{C_{3n}^n} = \frac{n!(2n)!}{(3n)!}$$