Rules: online test in lms, no proctoring, 20 questions, 60 minutes, only numerical answers are checked, two digits after decimal point are requested, use anything you want (calculators, python/r code, google, ...), don't cheat.

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. (bootstrap) Nature generates random variables X_1, ..., X_{100} independently and uniformly on $[0; 10]$.

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

   Find the variance $\text{Var}(X_1^*)$.

3. (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.

4. (welch) Assume that X_i are independent and identically normally distributed $\mathcal{N}(\mu, \sigma^2)$, sample size is $n = 10$.

   Find $\text{Var}(\sum(X_i - \bar{X})^2/(n-1))$.

5. (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).

6. (mw test) I have five results of two runners A and B for the 5 km race: three results for A and two results for B. Assume that the running time for both runners are continuously distributed and their distribution are equal.

   What is the probability that the running times of the runner A will get the ranks 1 and 5?

7. (cuped) Consider three variables: target variable y_i, predictior x_i and indicator of treatment $z_i \in \{0, 1\}$. The treatment z_i was assigned independently of x_i, total $n = 200$.

   The matrix of all cross products (sums of the form $\sum a_i b_i$) C is provided. The order of variables is y, 1, x and z. For example, $\sum 1 \times y_i = 10$:

$$C = \begin{pmatrix} 500 & 10 & 2 & 8 \\ 10 & 200 & 100 & 100 \\ 2 & 100 & 100 & 40 \\ 8 & 100 & 40 & 100 \end{pmatrix}$$

   Consider CUPED with first regression given by $\hat{y}_i = \hat{\alpha}_1 + \hat{\alpha}_2 x_i$ with residuals $r_i = y_i - \hat{y}_i$.

   What is the cross-product $\sum r_i z_i$?

8. (cuped) Consider three variables: target variable y_i, predictior x_i and indicator of treatment $z_i \in \{0, 1\}$. The treatment z_i was designed to be independent of x_i, but in fact $x_i = f_i \cdot (1 + 0.01 z_i)$.

   We suppose that z_i are Bernoulli with $p = 0.5$, $f_i \sim \mathcal{N}(1; 1)$ and they are independent.

   Find the probability limit
   $$\text{plim} \frac{\sum (x_i - \bar{x})(z_i - \bar{x})}{n - 1}.$$

9. (matching) Vasiliy uses knn with 1 neighbour to match observations. Here z is treatment indicator, x is predictor and y is target variable.

i	y_i	x_i	z_i
1	6	6	0
2	6	1	0
3	6	3	0
4	6	7	0
5	6	2	1
6	6	5	1
7	6	9	1
8	6	1	1

   Which indivial will be matched with individual number 3?

10. (matching) The indicator of treatment z is Bernoulli with $p = 0.5$. The conditional distribution of variables $y(0)$ (hypothetical outcome if $z = 0$) and $y(1)$ (hypothetical outcome if $z = 1$) is given in two tables

Condition $z = 0$	$y(0) = 0$	$y(0) = 1$
$y(1) = 0$	0.1	0.5
$y(1) = 1$	0.3	0.1

Condition $z = 1$	$y(0) = 0$	$y(0) = 1$
$y(1) = 0$	0.4	0.1
$y(1) = 1$	0.2	0.3

   What is the average treatment effect $\mathbb{E}(y(1) = y(0))$?

11. (multiple comparison) I do 100 independent tests at significance level 5%. The null hypothesis for all 100 tests is actually true, but I don't know this.

    What is the probability that I will receive at least 2 significant results?

12. (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 lowes 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?

13. (sample size) My target variable y is continuous. I have two stratas of respondents with preliminary estimates $\hat{\sigma}_s$, total strata size N_s, and cost per one observation c_s.

My total budget is 5000. I wish to estimate $\mathbb{E}(y_i)$.

How much observations I should sample from the first strata?

Strata s	$\hat{\sigma}_s$	N_s	c_s
1	20	10^5	4
2	30	$2 \cdot 10^5$	1

14. (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?

15. (contingency table) I eated 10 M&Ms: 2 green, 1 red, 4 yellow, 1 green, 2 red.

    Only these three colors are possible. I assume that yellow and green colors are equally probable.

    Calculate the maximal log likelihood for my model.

16. (contingency table) Consider the following contingency table

	$B = 1$	$B = 2$
$A = 1$	10	20
$A = 2$	30	40

    Calculate LR statistic that checks the hypothesis that A and B are independent against dependency alternative.

17. (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.

18. (anova 1+2) Vasiliy loves to eat shaurma. He has three local shaurma dealers with two types of shaurma. Vasiliy bought 7 shaurmas of each type from each dealer and measured their weight. He would like to test the hypothesis that weight depenends on the shaurma type and not on the dealer. He assumes no interaction.

    Total sum of squares is 1000, sum of squares explained by dealers is 300, sum of squares explained by type is 500.

    Calculate the F-statistic to test the hypothesis.

19. (partial correlation) The variables X and Y are jointly normal with zero means, unit variances and $\text{Corr}(X, Y) = 0.8$.

    Find α such that $X^* = X - \alpha Y$ is not correlated with Y.

20. (partial correlation) The variables X_1, X_2, ... are independent and identically distributed with mean 5 and variance 7.

    Find $\text{pCorr}(X_1, X_2; S)$ where $S = X_2 + X_3$.