Q1. A probability P(x) is:

- any real number
- a number between 0 and 1 (inclusive)
- always strictly greater than 0
- always an integer

Q2. The p-value is:

- the probability that H_0 is true
- the probability of observing an event at least as extreme as the one observed, assuming H_0 is true
- the probability that H_1 is true
- the probability of committing a Type II error

Q3. In a one-sample t-test, the main assumption is:

- independence and approximate normality
- equal variances between groups
- the variable of interest must be categorical
- sample size must always be > 100

Q4. The Shapiro-Wilk test is used to:

- check homoscedasticity
- test normality
- test independence between two categorical variables
- compare means across more than two groups

Q5. Spearman correlation is preferred over Pearson correlation when:

- the relationship is perfectly linear
- assumption of normality is violated
- both variables are categorical
- you aim to build a predictive model

Q6. In the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, the standard null hypothesis is:

- $H_0: \beta_0 = 0$
- $H_0: \varepsilon = 0$
- $H_0: \beta_1 = 0$
- $H_0: x = y$

Q7. In one-way ANOVA, the F-statistic compares:

- means vs medians
- slope / intercept
- Type I error / Type II error
- between-group variance / within-group variance

Q8. In PCA, the loadings represent:

- coordinates of observations in the PC space
- p-values associated with variables
- weights/contributions of original variables to each principal component
- Euclidean distances between observations

Q9. Main difference between a matrix and a data.frame in R:

- a matrix can store different types in different columns, whereas the data.frame is always numeric
- a matrix has always the number of rows equal to the number of columns
- a matrix is homogeneous (one type), a data.frame can be heterogeneous
- no difference

Q10. In R, scale() is used to:

- compute variance of x
- standardize variables (centering and scaling)
- convert the matrix x into a data frame
- subset a dataset

Q11. Which command fits a simple linear regression in R?

- cov(y ~ x)
- t.test(y, x)
- cor(y, x)
- lm(y ~ x)

Q12. Which function gives the full output of an lm model?

- coef()
- str()
- summary()
- class()

Q13. If ANOVA assumptions are strongly violated (non-normal residuals + heteroscedasticity), a reasonable alternative is:

- always use Tukey HSD
- increase the number of factor levels
- use Pearson correlation
- use Kruskal-Wallis test

Q14. Which function computes a Euclidean distance matrix between rows?

- dist()
- cmdscale()

- prcomp()
- length()

Q15. In dudi.pca() function, which arguments make PCA non-interactive and set the number of components?

- center=TRUE, scale=TRUE
- scannf=FALSE, nf=k
- k=2
- method = "euclidean"

Q16. Which function performs classical (metric) MDS in 2D?

- cmdscale()
- metaMDS()
- prcomp()
- hclust()

Q17. List and briefly explain the four main assumptions underlying the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$.

Q18. Point out differences and similarities between a PCA and an MDS.

Q19. (Regression – dataset "trees")

The dataset *trees* contains measurements collected on 31 black cherry trees.

Variables are:

- Girth: tree diameter (in inches)
- Height: tree height (in feet)
- Volume: timber volume (in cubic feet)

Questions:

1. Interpret the coefficient of Girth keeping Height constant
2. Which predictors are statistically significant at $\alpha = 0.05$? What about $\alpha = 0.01$?
3. Explain what the global F-test is testing in this multiple regression model (write the null hypothesis)
4. Briefly comment on the Adjusted R-squared value: what does it tell you about the model?

Q20. (Two-way ANOVA – dataset "plants")

Variables are:

- growth: quantitative response (plant growth)
- light: factor with two levels (light conditions)
- soil: factor with two levels (soil type)

Questions:

1. Is there evidence that light has a significant effect on plant growth? If so, describe the direction and justify your answer
2. Is there evidence that soil type has a significant effect on plant growth? If so, describe the direction and justify your answer
3. Is the interaction light:soil significant? What does it mean conceptually?
4. From the interaction plot: what graphical feature indicates interaction? Justify