BIOENG-210: Biological Data Science I: Statistical Learning

BIOENG-210 Mock Exam
Prof. Gioele La Manno

April 2024

You have 60 minutes to answer all questions. Each question is worth 1 point. Only one answer per question is correct.

Question 1 What is the evaluation at 0 of the probability density function of a normally distributed random variable centred on 0 with a variance of 6?

(a) about 0.40

(b) about 0.17

(c) about 0.07

(d) about 0.28

Question 2 A random variable generating probabilities can usually be fit with the following probability distribution:

(a) normal

(b) gamma

(c) beta

(d) Poisson

Question 3 What are we trying to find when performing a maximum likelihood estimation?

(a) The family of distribution that best fits the data

(b) The sample size needed to accurately estimate the distribution parameters

(c) The set of parameters that minimizes the likelihood function

(d) The set of parameters that minimizes the negative log-likelihood function

Question 4 An agronomist is trying to boost milk production on her farm. She fed one herd of cows with the usual feed, and another with a special feed enriched in certain nutrients. After two weeks, she measures milk production from each animal. Which of the following statements can you draw from the following boxplot:

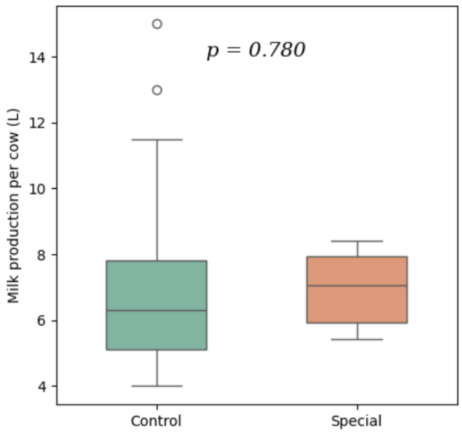

Figure 1: The p-value at the significance level of α=0.05 is plotted on top of the plot. Outliers are circles in white.

(a) The average milk production for cows fed on the special diet is bigger than the one for cows fed on the control diet.

(b) A quarter of the cows fed on the control diet produced less milk than the lowest producing cow in the group fed with the special diet.

(c) The special feed significantly increases milk production at the significance level of α=0.05.

(d) It is not possible to conclude whether the special feed increases milk production from this data.

Question 5 You are the CEO of a pharma company that is currently undergoing heavy budget reductions. You can continue testing only one of two promising drugs for mononucleosis, A or B. Your team sends you a probability density graph of the test statistic under the (complex) null distribution, plus the test statistic values for A and B. You recall from BIOENG-210 that you should look at the areas under the curve (the p-values), after deciding on a criterion (α) for the false positive rate. You cannot risk more than 1% chance of getting a false positive. You observe that the area under the test statistic distribution from the test statistic value for drug A is 0.009, and for drug B it is 0.1. Which of the two drugs should you keep developing?

(a) Drug A

(b) Drug B

(c) None of them

(d) If I had the money, both of them

(e) There is not enough information to answer this question

Question 6 A researcher is measuring the time it takes a particular type of cell to divide. To do so, she measures n times the process and records each corresponding time so that she obtains the set of measurements $T = \{t_i\}_{i=1}^n$. She wants to model the probability distribution of this variable and, since it is a time variable, she decides to use an exponential distribution:

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

This distribution depends on a single parameter $\tau > 0$ and she wants to estimate it using MLE. What would be the MLE estimate of τ as a function of the set of data $\{t_i\}_{i=1}^n$ and n?

(a) $\tau = \frac{1}{n} \sum_{i=1}^n \log t_i$

(b) $\tau = \prod_{i=1}^n t_i^{-n}$

(c) $\tau = \frac{1}{n} \sum_{i=1}^n t_i$

(d) $\tau = \sqrt{\frac{1}{n} \sum_{i=1}^n t_i^2}$

(e) There is no closed form solution, therefore there is no way of estimating τ

(f) There is no closed form solution, she should use numerical methods.

Question 7 The KDE of the two observed variables x and y is shown in the figure.

What do you expect the correlation and the mutual information between x and y to be?

(a) $\rho(x,y) = -0.62, I(x,y) = 0.64$

(b) $\rho(x,y) = -0.02, I(x,y) = 0.74$

(c) $\rho(x,y) = 0.83, I(x,y) = 0.01$

(d) $\rho(x,y) = 0.07, I(x,y) = -0.05$

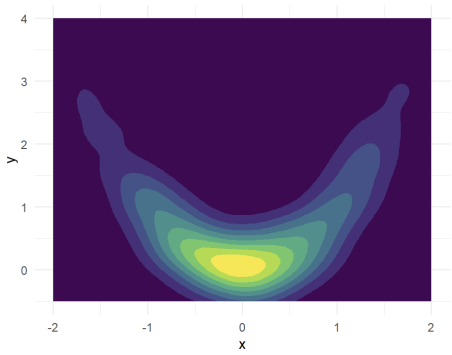

Figure 2: Kernel Density Estimate (KDE) of x and y.

Question 8 The p-value is:

(a) The evaluation of the test statistic at an arbitrarily defined threshold, e.g. 0.05.

(b) The area under the curve of the probability density function of the test statistic defined under the null hypothesis, beyond a certain threshold defined from the data.

(c) The evaluation of the test statistic computed from the sampled data.

(d) The probability of incorrectly rejecting the null hypothesis.

Question 9 The null hypothesis:

(a) is assumed to be false a priori.

(b) is always deduced from the data.

(c) defines the distribution of the test statistic for a one-sample t-test.

(d) is said to be true if the p-value is bigger than a certain threshold, e.g. 0.05.

Question 10 You ran a statistical test, observing a p-value of 0.02, so you rejected your null hypothesis. Which of these p-value intervals was/were **more likely** under the null if the null was true?

(a) 0.58–0.60

(b) 0.99–1.00

(c) 0.00–0.02

(d) A and C

(e) B and C

Question 11 You have two alternative models for the same variable. The following hypothesis tests are performed:

$$M_1: \begin{cases} H_0 : x \text{ is generated by } M_1 \\ H_1 : x \text{ is not generated by } M_1 \end{cases}$$

$$M_2: \begin{cases} H_0 : x \text{ is generated by } M_2 \\ H_1 : x \text{ is not generated by } M_2 \end{cases}$$

You observe these p-values: $p_1 = 0.279, p_2 = 0.567$. What can you conclude?

(a) We don't have sufficient evidence to reject either of the two models.

(b) The p-values are too high, so both models are rejected.

(c) $p_2 > p_1$, hence M_2 is better than M_1.

(d) $p_2 > p_1$, hence we reject M_2.

Question 12 Which of the following is **not** a valid probablity density function in the given domain of x:

(a) $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+, x \in \mathbb{R}$

(b) $f(x) = 1, x \in [0, 1]$

(c) $f(x) = \sqrt{2} - x, x \in [0, \sqrt{2}]$

(d) $f(x) = \frac{1}{b} e^{-\frac{|x-\mu|}{b}}, b \in \mathbb{R}, \sigma \in \mathbb{R}^+, x \in \mathbb{R}$

Question 13 In figure 3 we show the the probability density function of a bivariate gaussian with 0 mean and unknown covariance matrix (Σ).

Knowing that the absolute value of the correlation between x and y is $\rho_{xy} = 0.9$. Which of the following could correspond to the covariance matrix Σ:

(a) $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$

(b) $\Sigma = \begin{pmatrix} 3 & -2.7 \\ -2.7 & 1 \end{pmatrix}$

(c) $\Sigma = \begin{pmatrix} 3 & 2.7 \\ 2.7 & 1 \end{pmatrix}$

(d) $\Sigma = \begin{pmatrix} 3 & 1.6 \\ 1.6 & 1 \end{pmatrix}$

(e) $\Sigma = \begin{pmatrix} 3 & -2.7 \\ 1.6 & 1 \end{pmatrix}$

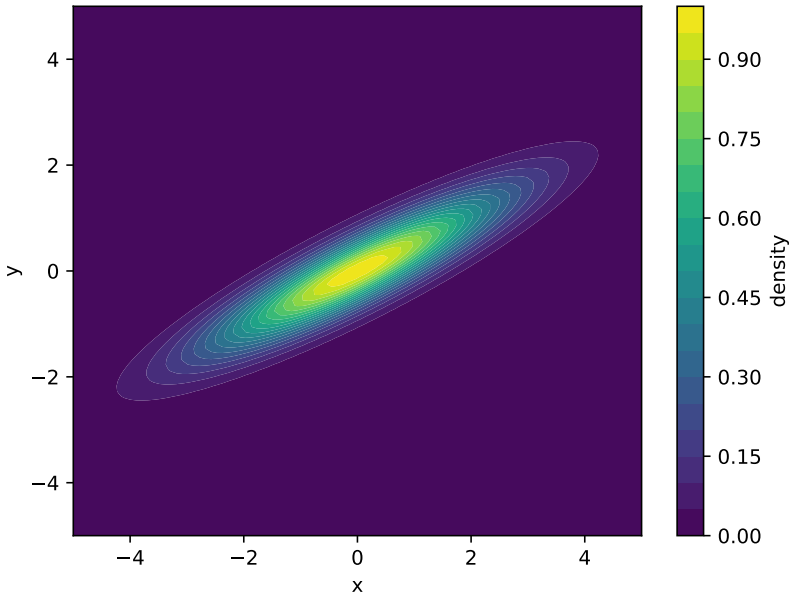

Figure 3: Probability density function of 2d gaussian with 0 mean and unknown covariance.

(f) $\Sigma = \begin{pmatrix} 3 & 0.9 \\ 0.9 & 1 \end{pmatrix}$

(g) $\Sigma = \begin{pmatrix} 3 & -0.9 \\ -0.9 & 1 \end{pmatrix}$

Question 14 You have to model three variables: x, the RNA-seq expression, y, the survival time after a diagnosis, and z, the number of mutated cells in a population with known and finite N. What distributions would you choose to model them?

(a) x Poisson, y Beta, z Bernoulli

(b) x Poisson, y Normal, z Gamma

(c) x Negative-Binomial, y Exponential, z Binomial

(d) x Negative-Binomial, y Dirichlet, z Gamma

Question 15 We have collected n samples from a random variable X of unknown mean and variance. We would like to test whether the mean of X, μ_X is equal to some reference value $\hat{\mu_X}$. Which of the following statements is **false**:

(a) If X is normally distributed, we can use a t-test and compare the t-statistic to a t-distribution centered at $\hat{\mu_X}$ with $n-1$ degrees of freedom.

(b) If X is normally distributed and n is very large, the result of a suitable t-test and a z-test will be almost identical.

(c) Regardless of the distribution of X, the central limit theorem tells us that if n is sufficiently large we can use a z-test because the distribution of X converges to a $\mathcal{N}(0,1)$.

(d) Even if X is normally distributed, since the variance of X is unknown we can not directly use a z-test to compare the mean.

Question 16 What is the primary assumption of multiple linear regression?

(a) The predictors are categorical

(b) The residuals are normally distributed

(c) The response variable is binary

(d) There are no outliers

Question 17 A marine biologist is studying the factors affecting coral reef health. She collects data on several variables:

- Water temperature (°C): ranging from 18-32
- pH level: ranging from 7.2-8.4
- Nitrate concentration (mg/L): ranging from 0.01-0.5
- Light penetration (meters): ranging from 1-25
- Current velocity (m/s): ranging from 0.05-2.5

She fits a multivariate linear regression model to predict coral cover percentage based on these variables. When examining the coefficients, she notices:

- Water temperature is 0.002
- pH level is 2.87
- Nitrate concentration is -9.45
- Light penetration is 0.32
- Current Velocity is -0.93

The biologist is surprised by how small the coefficient for water temperature is, despite knowing from previous research that temperature has a significant impact on coral health.

What is the most likely reason for this unexpected result?

(a) The data is not normally distributed.

(b) The variables are on different scales.

(c) The sample size is too large, diluting the effect of temperature.

(d) The data is not linearly related.

Question 18 What is the primary purpose of the design matrix in regression analysis?

(a) To store response values

(b) To compute confidence intervals

(c) To summarize residuals

(d) To organize predictor variables

Question 19 In multiple linear regression, what does the null hypothesis H_0 typically state for an individual predictor variable?

(a) The predictor's coefficient is greater than 1

(b) The predictor explains all the variance in the response

(c) The predictor has no effect on the response variable

(d) The predictor follows a normal distribution

Question 20 You are studying the effect of pregnancy on the mouse brain. You measure 10 genes, and you ask: "Which genes are changed by pregnancy?" For 5 genes, your alpha is 0.05; for the other 5 genes, your alpha is 0.01. If you do **not** do multiple testing correction, what is your probability of having at least one false positive?

(a) 9.6%

(b) 30.0%

(c) 26.4%

(d) 99.9%

(e) 51.7%

Question 21 You are a plant biologist using multiple linear regression. You want to model the expression of gene A as a function of: condition (rainy or sunny), age (continuous), tissue (leaf, stem, or root), genotype (5 possible genotypes). You made 8 measurements. What is the size of your design matrix if you want to ensure there is no collinearity?

(a) 8x10

(b) 8x5

(c) 10x8

(d) 8x13

(e) 8x11

Question 22 If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, what is the distribution of Y?

(a) $Y \sim \mathcal{N}(\mu + b, a\sigma^2)$

(b) $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

(c) $Y \sim \mathcal{N}(\mu + b, a^2\sigma^2)$

(d) $Y \sim \mathcal{N}(a\mu + b, a\sigma^2 + b)$

Question 23 Suppose we have a covariance matrix

$$\Sigma = \begin{pmatrix} 4 & a \\ a & 5 \end{pmatrix}$$

What is the set of values that a can take on such that Σ is a valid covariance matrix?

(a) $a \in (-\infty, \infty)$

(b) $a \in [-\sqrt{20}, \sqrt{20}]$

(c) $a \geq 0$

(d) $a \in (-\sqrt{20}, \sqrt{20})$

Question 24 A molecular biologist is analyzing gene expression data comparing two experimental conditions. He calculates a 95% confidence interval for the difference in mean expression levels between the two groups and obtains the interval [1.2, 2.5]. Which of the following interpretations of this confidence interval is correct?

(a) There is a 95% probability that the true difference in mean expression lies between 1.2 and 2.5.

(b) If the experiment were repeated many times, 95% of the calculated confidence intervals would contain the true difference in mean expression.

(c) There is a 95% chance that any individual measurement of gene expression difference will fall between 1.2 and 2.5.

(d) 95% of the observed differences in gene expression in this study lie between 1.2 and 2.5.

Question 25 In Figure 4 we show the residuals obtained after fitting a linear model, that is, for every pair (x_i, y_i):

$$\epsilon_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients.

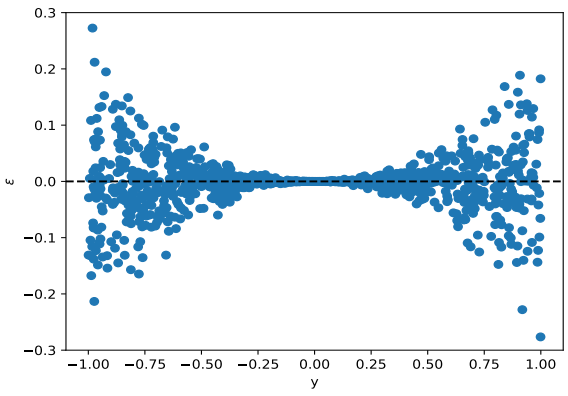

Figure 4: Residuals obtained from fitting a linear model

By looking at the plot, which of the following assumptions of linear regression in this particular data you think is most likely to not fullfilled:

(a) Linearity

(b) Homoscedasticity

(c) Normality of the conditional distribution

(d) All assumptions seem to be fullfilled