Exam Questions: Maximum Likelihood Estimation (Lecture 2)

Question 1: Maximum Likelihood Estimation (MLE)

Question: What is the fundamental idea behind Maximum Likelihood Estimation (MLE)?

Options:

A. Finding parameters that minimize the probability of observed data

B. Maximizing the probability of the observed data given the parameters

C. Finding parameters that match the sample mean exactly

D. Estimating parameters by averaging all possible values

E. Choosing the parameters that yield the largest standard deviation

Correct Answer: B

Explanation: MLE selects the parameter values that maximize the likelihood (i.e., the probability) of the observed data under the assumed statistical model.

Question 2: Likelihood Function for a Normal Distribution

Question: Suppose a dataset consists of independent and identically distributed (i.i.d.) observations following a normal distribution with unknown mean μ and variance σ^2. What is the likelihood function $L(\mu, \sigma^2)$ for this dataset?

Options:

A. The sum of probability densities for all observations

B. The product of probability densities for all observations

C. The difference between the probability densities of the first and last observation

D. The average of probability densities for all observations

E. The cumulative distribution function evaluated at the mean

Correct Answer: B

Explanation: Since the observations are independent, the joint likelihood is given by the product of the individual probability densities.

Question 3: MLE for an Exponential Distribution

Question: Given a dataset with n independent observations $x_1, x_2, \ldots, x_n$, following an exponential distribution with rate parameter λ, what is the Maximum Likelihood Estimator (MLE) for λ?

Options:
(Assuming the intended correct answer is the estimator: $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$)

A. $\hat{\lambda} = \dfrac{n}{\sum_{i=1}^{n} x_i}$

Correct Answer: A

Explanation: Maximizing the likelihood function for an exponential distribution yields $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$.

Question 4: Pearson Correlation Coefficient

Question: A researcher is testing whether two continuous variables X and Y are associated. She computes their Pearson correlation coefficient and finds $r = 0.85$. What does this indicate?

Options:

A. X and Y have a strong negative linear relationship

B. X and Y are independent

C. X and Y have a strong positive linear relationship

D. X is the cause of Y

E. X and Y follow a nonlinear relationship

Correct Answer: C

Explanation: A Pearson correlation of 0.85 indicates a strong positive linear relationship between X and Y (and note that correlation does not imply causation).

Question 5: Pearson vs. Spearman Correlation

Question: Which of the following best describes the key difference between Pearson and Spearman correlation coefficients?

Options:

A. Pearson measures linear relationships, while Spearman measures monotonic relationships.

B. Pearson is only used for categorical data, while Spearman is for continuous data.

C. Spearman considers the mean of the data, whereas Pearson does not.

D. Spearman correlation requires normally distributed data, while Pearson does not.

E. Pearson correlation can only be positive, while Spearman can be negative.

Correct Answer: A

Explanation: Pearson correlation assesses linear relationships, whereas Spearman correlation uses ranked data to measure the strength of a monotonic relationship.

Question 6: Covariance Expression

Question: Which is the mathematical expression of $\mathrm{Cov}(X, Y)$ for two random variables X and Y?

Options:

A. $\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$

B. $\mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[XY]$

C. $\mathrm{E}[X] - \mathrm{E}[Y]$

D. $\mathrm{E}[X] + \mathrm{E}[Y]$

E. $\mathrm{E}[X]\,\mathrm{E}[Y]$

Correct Answer: A

Explanation: By definition, $\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.

Exam Questions: Bivariate Estimates (Lecture 3)

Question 1: Marginal Distributions

Question: Which of the following correctly defines the marginal distribution of X from a joint distribution $f_{X,Y}(x, y)$?

Options:

1. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$
2. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$
3. $f_X(x) = \dfrac{f_{X,Y}(x, y)}{f_Y(y)}$
4. $f_X(x) = \max_y f_{X,Y}(x, y)$

Correct Answer: 1

Explanation: The marginal distribution $f_X(x)$ is obtained by integrating the joint density over all values of y.

Question 2: Independence in Gene Expression

Question: A researcher measures expression levels of two genes across hundreds of cells and finds that the joint probability density function can be expressed as

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

What does this imply about these genes?

Options:

1. They are co-regulated by the same transcription factor
2. They are statistically independent of each other
3. They must have equal mean expression levels
4. The genes are located on the same chromosome

Correct Answer: 2

Explanation: The factorization of the joint density indicates that the gene expression levels are statistically independent.

Question 3: Conditional Distributions

Question: In a bivariate normal distribution with correlation coefficient ρ, what happens to the conditional distribution of Y given $X = x$ as $|\rho|$ increases?

Options:

1. The variance of the conditional distribution increases.
2. The mean of the conditional distribution becomes more dependent on the value of x.
3. The conditional distribution approaches a uniform distribution.
4. The conditional distribution becomes independent of the value of x.

Correct Answer: 2

Explanation: With a higher absolute correlation, the conditional mean of Y given $X = x$ becomes more strongly linked to x, and the conditional variance generally decreases.

Question 4: Bivariate Normal Properties

Question: For a bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and correlation $\rho = 0.5$, what is the probability density at the point $(1, 1)$ relative to the density at the origin $(0, 0)$?

Options:

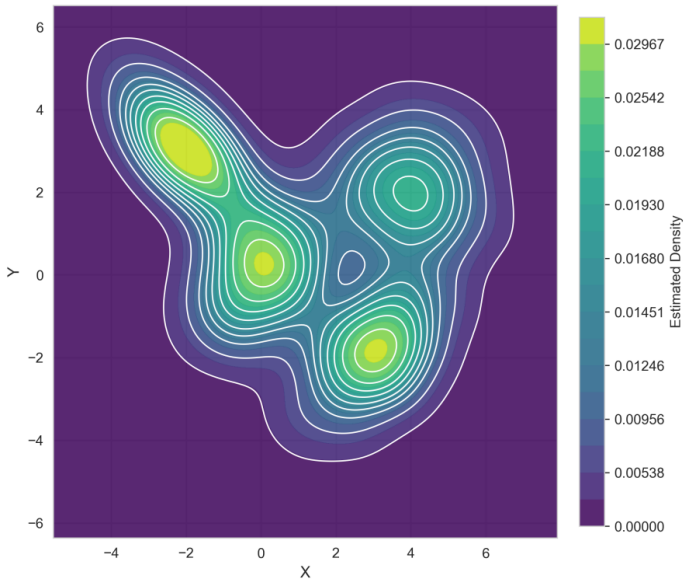
1. 0.223 (about 22.3% of the density at the origin)
2. 0.368 (about 36.8% of the density at the origin)
3. 0.472 (about 47.2% of the density at the origin)
4. 0.607 (about 60.7% of the density at the origin)

Question 5: Understanding KDE Visualization

Question: Based on the KDE contour plot shown, what can we conclude about the underlying bivariate distribution?

Options:

1. It shows independent variables with no correlation.
2. It exhibits a single mode with roughly circular level sets.
3. It shows a strongly bimodal distribution with two distinct clusters.
4. It displays a complex distribution with multiple local maxima and a curved ridge structure.



Correct Answer: 4

Explanation: We can see clearly multiple peaks (high local density) and a complex relationship between the two variables.

Exam Questions: Practical Testing Scenarios (Lecture 4)

Question 1: Standard Error of the Mean (SEM)

Question: What is the standard error of the mean (SEM)?

Options:

A. The standard deviation of the population

B. The standard deviation of the sample

C. The standard deviation of the sampling distribution of the mean

D. The variance of the sampling distribution of the mean

Correct Answer: C

Explanation: The SEM quantifies the variability of the sample mean as an estimator of the population mean.

Question 2: Interpretation of a p-value

Question: A researcher conducts a drug trial and obtains a p-value of 0.03 for the difference in mean response between treatment and control groups. What is the correct interpretation of this p-value?

Options:

A. There is a 3% probability that the drug has no effect.

B. There is a 3% probability that the observed difference occurred by chance.

C. If the drug truly has no effect, there is a 3% probability of observing a difference as large or larger than what was observed.

D. 97% of patients will respond positively to the drug.

Correct Answer: C

Explanation: The p-value is the probability of obtaining results at least as extreme as those observed if the null hypothesis is true.

Question 3: Sampling Distribution of Variance

Question: A sample of size $n = 20$ is drawn from a normal population. The sample variance s^2 is calculated. Which of the following expressions correctly describes the sampling distribution of the quantity

$$\frac{(n-1)s^2}{\sigma^2}\,?$$

Options:

A. A chi-square distribution with n degrees of freedom

B. A chi-square distribution with $n-1$ degrees of freedom

C. A t-distribution with $n-1$ degrees of freedom

D. A normal distribution with mean 0 and variance 1

Correct Answer: B

Explanation: The statistic $\frac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with $n-1$ degrees of freedom.

Question 4: Degrees of Freedom in a Paired t-test

Question: In a paired t-test with 15 pairs of observations, what is the degrees of freedom for the test statistic?

Options:

A. 14

B. 15

C. 28

D. 29

Correct Answer: A

Explanation: For a paired t-test with n pairs, the degrees of freedom is $n-1$. Here, $15-1=14$.

Question 5: Appropriate Test for One-Sample Data

Question: The figure below shows one-sample data with a reference value (dashed line). If you wanted to test whether the sample mean is significantly different from the reference value, what would be the most appropriate test?

Options:

A. Z-test, because the sample size is small

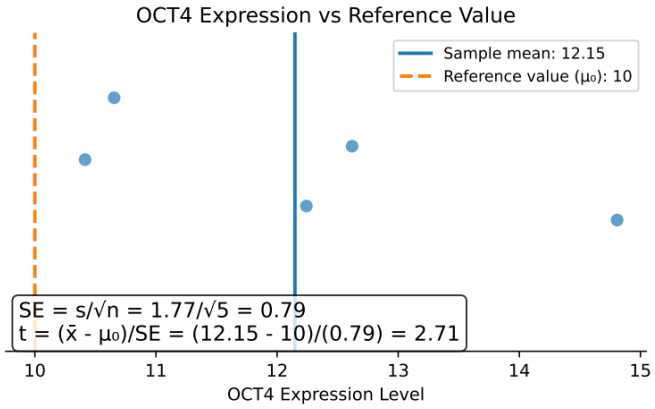
B. Two-sample t-test, because the reference represents a second group

C. One-sample t-test, because we deal with a fixed reference value

D. One-sample Paired t-test, to account for the experimental design

Correct Answer: C

Explanation: When comparing a sample mean to a fixed reference value, a one-sample t-test is the appropriate choice.



Question 6: Multiple Testing – Expected False Positives

Question: A biologist tests 2,000 genes for differential expression between healthy and diseased tissue using a significance threshold of $\alpha = 0.05$, and finds 200 significant genes. If no multiple testing correction is applied, approximately how many false positives would be expected among these 200 genes?

Options:

A. 10 false positives

B. 100 false positives

C. 200 false positives

D. It is impossible to estimate without knowing the true proportion of differentially expressed genes

Correct Answer: B

Explanation: With $\alpha = 0.05$, one would expect about $0.05 \times 2000 = 100$ false positives among all tests, so roughly 100 of the 200 significant genes may be false positives.

Question 7: Bonferroni Correction

Question: When controlling the Family-Wise Error Rate (FWER) at $\alpha = 0.05$ using the Bonferroni correction for 1,000 independent tests, what is the corrected significance threshold for each individual test?

Options:

A. $\alpha_{\text{corrected}} = 0.05$

B. $\alpha_{\text{corrected}} = 0.005$

C. $\alpha_{\text{corrected}} = 0.0005$

D. $\alpha_{\text{corrected}} = 0.00005$

Correct Answer: D

Explanation: The Bonferroni correction sets $\alpha_{\text{corrected}} = \frac{0.05}{1000} = 0.00005$.

Question 8: FWER vs. FDR

Question: What is the key difference between controlling the Family-Wise Error Rate (FWER) and controlling the False Discovery Rate (FDR)?

Options:

A. FWER controls the probability of making at least one false discovery, while FDR controls the expected proportion of false discoveries among all rejected null hypotheses.

B. FWER is applicable only to small numbers of tests, while FDR works for any number of tests.

C. FWER requires independence between tests, while FDR does not make any assumptions about independence.

D. FWER is always more powerful than FDR regardless of the number of tests.

Correct Answer: A

Explanation: FWER limits the chance of any false positives, whereas FDR controls the expected proportion of false positives among the tests declared significant.

Question 9: Inappropriate Use of the Standard Normal Distribution

Question: A researcher calculates a test statistic $t = 2.8$ from a sample of $n = 10$ observations. Under the null hypothesis, this statistic follows a t-distribution with 9 degrees of freedom, yielding a two-tailed p-value of 0.021. What would be the p-value if the researcher had instead (incorrectly) used a standard normal distribution?

Options:

A. $p < 0.005$

B. $0.005 < p < 0.01$

C. $0.01 < p < 0.02$

D. $p > 0.02$

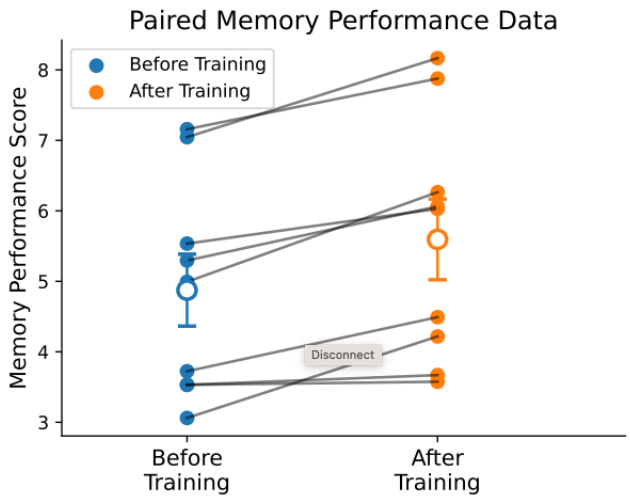
Correct Answer: A

Explanation: Using the standard normal distribution (which has thinner tails than the t-distribution for small samples) would yield a much smaller p-value.

Question 10: Paired vs. Unpaired t-test

Question: Referring to the figure showing paired versus unpaired analysis, why does the paired t-test detect a significant effect while the unpaired t-test does not?

Options:



A. The paired t-test uses more sophisticated statistical methods.

B. The paired t-test accounts for consistent within-subject differences despite high between-subject variability.

C. The paired t-test has a different null hypothesis than the unpaired t-test.

D. The paired t-test requires fewer assumptions about the data distribution.

Correct Answer: B

Explanation: By accounting for the pairing of observations, the paired t-test reduces variability, increasing its sensitivity.

Exam Questions: Simple Linear Regression Analysis (Lecture 5)

Sample Question 1: Error Distribution in Regression

Question: In simple linear regression, the model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

assumes that the error term ε follows which distribution?

Options:

A. Uniform distribution

B. Student's t-distribution

C. Normal distribution with mean 0 and constant variance

D. Chi-square distribution

E. Exponential distribution

Correct Answer: C

Explanation: The classical assumption in linear regression is that the errors are normally distributed with mean 0 and constant variance.

Sample Question 2: Precision of the Slope Estimate

Question: Which of the following would increase the precision (reduce the standard error) of the slope estimate?

Options:

A. Collecting data points with x-values close to the mean $\bar{x}$

B. Increasing the error variance σ^2

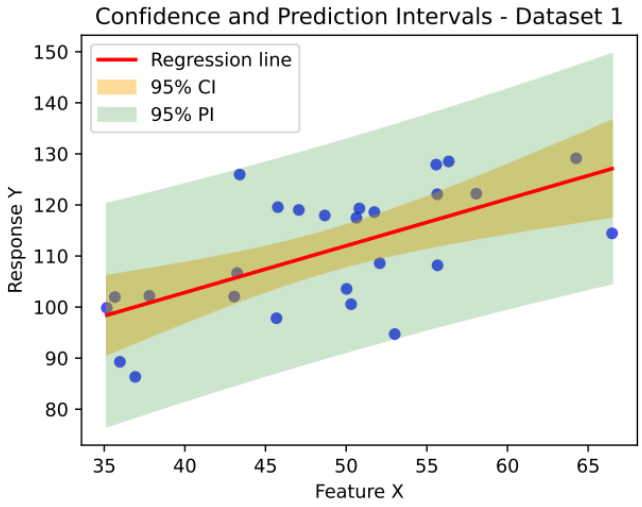
C. Reducing the sample size n

D. Increasing the spread of x-values around their mean

E. Focusing on values of x that produce the largest residuals

Correct Answer: D

Explanation: A larger spread in x-values provides more information about the relationship between X and Y, thereby reducing the standard error of the slope estimate.



Sample Question 3: Regression Bands and Outliers

Question: Consider the regression bands shown in the figure. If a new data point is observed at $X = 30$ with $Y = 110$, which of the following statements is correct?

Options:

A. This observation provides evidence that the regression model is incorrect.

B. This observation falls within the 95% prediction interval but outside the 95% confidence interval.

C. This observation is considered an outlier because it falls outside both intervals.

D. The probability that the true mean response at $X = 30$ equals 15 is 95%.

E. We expect 95% of observations at $X = 30$ to fall within the inner band.

Correct Answer: B

Explanation: The 95% prediction interval covers individual future observations, whereas the 95% confidence interval covers the mean response. An individual data point can lie outside the confidence interval but still be within the prediction interval.

Sample Question 4: Confidence Interval for the Slope

Question: In constructing a confidence interval for the slope parameter in simple linear regression, which factor would make the interval narrower?

Options:

A. Decreasing the sample size

B. Increasing the confidence level from 95% to 99%

C. Smaller variability in the response variable (smaller σ^2)

D. Collecting data points with x-values very close to each other

E. Using a one-tailed rather than a two-tailed test

Correct Answer: C

Explanation: Less variability in the response variable yields a more precise estimate of the slope, resulting in a narrower confidence interval.

Sample Question 5: Prediction vs. Confidence Intervals

Question: A researcher measures enzyme activity (Y) as a function of substrate concentration (X) and fits a simple linear regression model. The 95% prediction interval at $X = 5$ is $[10, 30]$, while the 95% confidence interval for the mean response at $X = 5$ is $[15, 25]$. Which statement is correct?

Options:

A. The confidence interval is wider because it accounts for more sources of uncertainty.

B. The estimate of the mean response at $X = 5$ is 15.

C. If the experiment were repeated many times, about 95% of individual observations at $X = 5$ would fall between 10 and 30.

D. The true mean response at $X = 5$ has a 95% probability of falling between 15 and 25.

E. The prediction interval and confidence interval would become identical with a large enough sample size.

Correct Answer: C, D

Explanation: By definition.