Midterm

IDS 702 - Fall 2024

First name: ________________________ Last name: ________________________
Net ID: ________________________

I hereby state that I have not communicated with or gained information in any way from my classmates or unauthorized materials during this exam, and that all work is my own.

Signature: ____

Any potential violation of Duke's policy on academic integrity will be reported to the Office of Student Conduct & Community Standards. All work on this exam must be your own.

1. You have 75 minutes to complete the exam.
2. You are **not** allowed a cell phone (even if you intend to use it for checking the time), music device or headphones, notes, books, or other resources, or to communicate with anyone other than the professor or TAs during the exam.
3. Write clearly.

The exam has 17 questions: 15 multiple choice questions and 2 short answer questions. Question 16 has parts a-d and question 17 has parts a-e. **Before you begin, make sure your exam has all questions.**

Formula Page

$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$

$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Independence: $Pr(A|B) = Pr(A)$, $Pr(A \cap B) = Pr(A)Pr(B)$

Bayes Theorem: $Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$

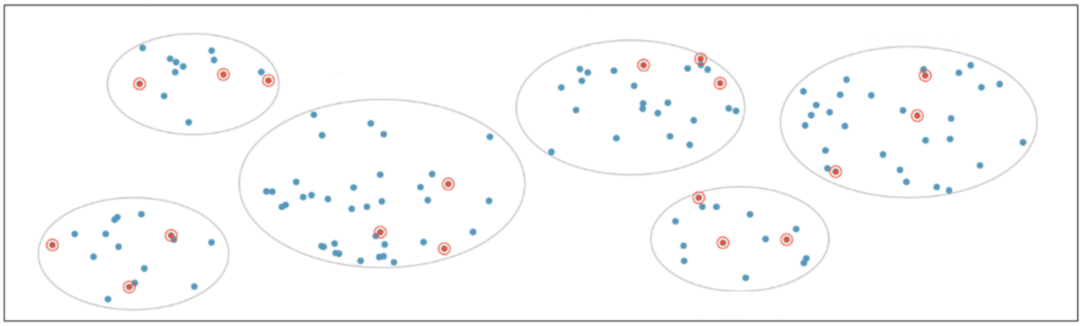
Total probability: $Pr(A) = Pr(A|B)Pr(B) + Pr(A|B^c)Pr(B^c)$

General formula for confidence interval: $\bar{x} \pm z * \times \sigma / \sqrt{n}$

Multiple Choice (3 points each)

Circle the letter to mark your answer choice.

1. Which sampling scheme is presented in the image below? (*Note for 2025: I will not ask about sampling schemes because we did not cover this like I did in 2024*)



   a. Simple random sampling

   b. Stratified sampling

   c. Cluster sampling

   d. Multilevel sampling

2. What is the distinction between standard deviation and standard error?

   a. Standard error is the standard deviation of the sampling distribution

   b. Standard deviation is the standard error of the sampling distribution

   c. Standard error is the standard deviation of the population distribution

   d. Standard deviation is the standard error of the population distribution

3. Let $X_1, X_2, ..., X_n \stackrel{iid}{\sim} Exp(\lambda)$, where the PDF of the exponential distribution is given by $f(x_i; \lambda) = \lambda e^{-\lambda x_i}$ for $x_i > 0$. The derivative of the log-likelihood, which is called the score function, is $\frac{n}{\lambda} - \sum_{i=1}^n x_i$. What is the maximum likelihood estimator $\hat{\lambda}$?

   a. $\frac{\sum_{i=1}^n x_i}{\lambda}$

   b. $\boxed{\frac{n}{\sum_{i=1}^n x_i}}$

   c. $\frac{\sum_{i=1}^n x_i}{n}$

   d. The MLE cannot be calculated because x_i cannot be negative.

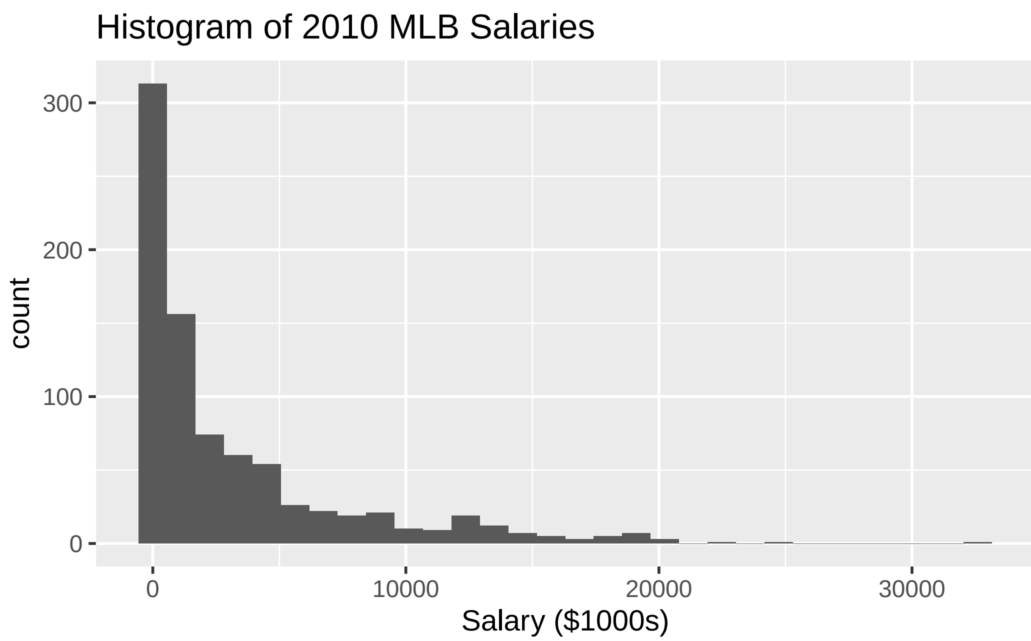
① PDF
② Joint distribution
→ written in terms of the parameter is the likelihood
③ Take the log
④ Differentiate w.r.t. parameter
⑤ Set to 0 and solve for the parameter

Start here in this problem

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$$

$$\frac{n}{\lambda} = \sum_{i=1}^n x_i \qquad \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i}$$

4. The dataset `mlb` has the 2010 salary information (in \$1000s) for 828 Major League Baseball players. Based on the histogram of the `mlb` salaries, can you characterize the distribution of the sample mean of MLB salaries? Assume that individual salaries are independent.



Central Limit Theorem!!

a. No, we would need to know the population distribution of the sample mean to characterize the distribution of the sample mean

b. Yes, the sample mean would have a right-skewed distribution

c. Yes, the sample mean would have a normal distribution

d. No, we would need to know the population distribution of MLB salaries to characterize the distribution of the sample mean

5. A factory has two machines, A and B. Machine A produces 60% of the items, while Machine B produces 40%. The probability that an item produced by Machine A is defective is 2%, and the probability that an item produced by Machine B is defective is 3%. If a randomly selected item is found to be defective, what is the probability that it was produced by Machine A?

a. $\frac{(0.6)(0.4)}{(0.02)(0.03)}$

b. $\frac{(0.6)(0.4)}{(0.02)(0.03)+(0.6)(0.4)}$

c. $\frac{(0.02)(0.6)}{(0.02)(0.6)+(0.03)(0.4)}$

d. $\frac{(0.02)}{(0.4)}$

Bayes Theorem!!

$P(A) = 0.6$ $\quad P(B) = 0.4$ $\quad P(D|A) = 0.02$

$P(D|B) = 0.03$

want to know $P(A|D)$

$P(A|D) = \frac{P(D|A)\,P(A)}{P(D)}$ Bayes Theorem

$P(D) = P(D|A)\,P(A) + P(D|B)\,P(B)$ Total law of probability

6. Which of the following is true?

   a. Independent events are mutually exclusive

   b. If two events are mutually exclusive, then $P(A \cap B) = P(A)P(B)$

   c. If two events are independent, then $P(A \cap B) = 0$

   d. If two events are mutually exclusive, then $P(A \cap B) = 0$

Mean = CLT

7. In which of the following cases would bootstrapping be most useful?

   a. You collect data on 250 workers' salaries across two companies and you want to compare median salary for the two companies.

   b. You collect data on 2500 workers' salaries across two companies and you want to compare mean salary for the two companies.

   c. You collect data on 250 workers' salaries in 2024 and compare the mean salary to the same workers' mean salaries in 2020.

   d. You collect data on 2500 workers' salaries across five companies and you want to compare mean salary for the five companies.

8. You perform an experiment to test the effect of a protein supplement on workout performance. You measure 100 runners' mile times once without using the protein supplement, and then again one week later after using the protein supplement for seven days. You want to compare the mean mile time with and without the supplement. Is a two-sample t-test appropriate in this case?

   a. Yes, because the runners between and within groups are likely independent and the sample size is sufficiently large.

   b. No, because the runners between the groups are not independent.

   c. No, because the runners within the groups are not independent.

   d. No, because we do not have evidence that the mile times are normally distributed.

9. Which of the following does NOT affect the calculation of a p-value?

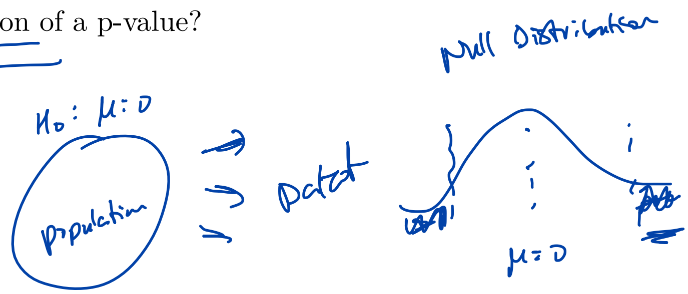
   a. Null hypothesis

   b. Alternative hypothesis

   c. Observed statistic

   d. Significance level α

interpretation but not calculation



10. What is the difference between β and $\hat{\beta}$?

a. β is an estimate that describes the population, while $\hat{\beta}$ is an unknown quantity using the sample

b. β is an unknown quantity that describes the sample, while $\hat{\beta}$ is an estimate using the population

c. β is an estimate that describes the sample, while $\hat{\beta}$ is an unknown quantity using the population

d. β is an unknown quantity that describes the population, while $\hat{\beta}$ is an estimate using the sample

population
parameter

11. Which of the following is not considered a fixed quantity in the theoretical representation of the simple linear regression model?

a. Y

b. β_1

c. X

d. σ^2

Fixed vs. Random

Random = probability distribution

$\boxed{Y} \sim N(\beta_0 + \beta_1 X, \sigma^2)$

$Y = \boxed{\beta_0} + \boxed{\beta_1}\boxed{X} + \varepsilon, \quad \boxed{\varepsilon} \sim N(0, \boxed{\sigma^2})$

$R^2 = 1 - \boxed{\frac{SSR}{SST}}$ variance of Y

12. Which statement is true about R^2?

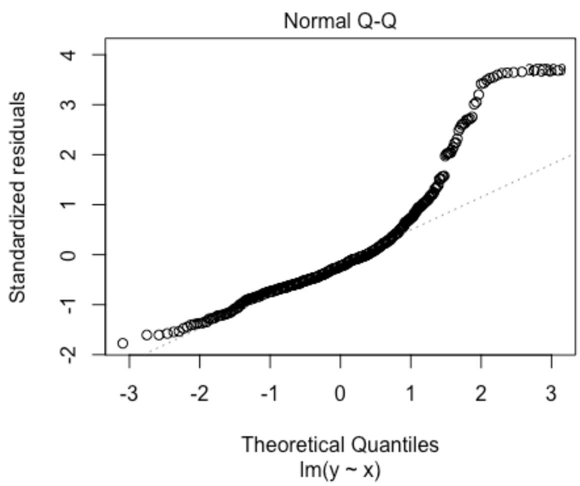
a. R^2 is a proportion because the total variance of Y is always greater than or equal to the sum of squared residuals

b. R^2 will always decrease when adding predictors to a regression model

c. An R^2 value of 0 indicates a perfect model and an R^2 value of 1 indicates a poor model fit

d. R^2 is interpreted as the proportion of variability in the predictor(s) that is explained by the outcome

13. The plot below indicates that which assumption of linear regression may be violated?



a. Linear relationship between the outcome and predictors

b. Independence of error terms

c. Errors are normally distributed

d. Errors have equal variance

14. Which of the following is NOT true about interaction terms in a linear model?

a. Interaction terms assess the effect of one predictor on the outcome based on the value of a second predictor

b. Interaction terms should be used if two predictors both have an effect on the outcome

c. Interaction terms can generate different slopes for a continuous predictor for different levels of a categorical predictor

d. Interaction terms can be included in a model based on the research question or exploratory data analysis

15. In the model below, x_1 is a numeric (continuous) variable, while x_2 and x_3 are indicators for a categorical variable. What is the correct interpretation of the $\widehat{\beta}_4$ coefficient estimate?

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \boxed{\beta_4 x_1 x_2} + \beta_5 x_1 x_3 + \epsilon$$

a. The difference between the effect of x_2 on Y for x_1 and the effect of x_2 on Y for the reference level

b. The difference between the effect of x_1 on Y for x_2 and the effect of x_1 on Y for the reference level

c. The multiplicative effect of the numeric variable x_1 and the categorical variable on Y

d. The average increase of Y per unit increase of x_2 compared to the reference level of x_1

$x_2 = 1$: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1 + \varepsilon$
$x_3 = 0$

reference level: $Y = \beta_0 + \beta_1 x_1 + \varepsilon$

Short Answer

Question 16 (12 points)

You conduct a two-sample t-test to compare the mean full-value property tax rate per $10,000 for Boston homes that are bordered by the Charles River vs. those that are not. You obtain the following results:

Mean tax rate for homes that do not border the river: 409.9

Mean tax rate for homes that border the river: 386.3

Sample statistic: 23.6

p-value: 0.42

95% confidence interval: [-34.9, 82.2]

significance level: 0.05

a. State whether or not the p-value interpretations given below are correct or incorrect. If the interpretation is incorrect, identify why that is the case.

i. The probability that the null hypothesis is true is 0.42. Therefore, we can accept the null hypothesis that the mean tax rate is equal for houses that border the river and houses that do not border the river.

① p-value is not the prob. that null is true

② we do not accept the null, we fail to reject

ii. Assuming that there is a difference in the mean tax rate between houses that border the river and houses that do not, the probability of observing data equal to or more extreme than what we observe is 0.42. Therefore, we reject the null hypothesis that the mean tax rate is equal for houses that border the river and houses that do not border the river.

① Assume NO difference

② 0.42 > 0.05, so we would not reject the null

b. State whether or not the 95% confidence interval interpretations given below are correct or incorrect. If the interpretation is incorrect, identify why that is the case.

i. There is a 95% probability that the true difference in mean tax rate between houses that do not border the river and houses that do is between -34.9 and 82.2.

Incorrect, CI is not a probability

ii. There is a 95% probability that the observed difference in mean tax rate between houses that do not border the river and houses that do is between -34.9 and 82.2.

Incorrect, observed difference is in the interval

Question 17 (15 points)

A dataset contains information about 134 selected countries. The variables in the dataset are listed below. You want to understand the factors that are related to life expectancy.

Codebook:

Life.expectancy Average life expectancy (years)

infant.deaths Number of infant deaths per 1000

Hepatitis.B Hepatitis B immunization rate in infants (%)

Status Economic categorization (developed or developing)

Measles Number of reported measles cases per 1000

First, you fit a model regressing life expectancy on infant deaths, Measles, Hepatitis B, and Status. Use the output shown below to answer questions a and b.

```
                      Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)       71.6389061410 2.4781592912 28.908112 3.287085e-58
infant.deaths     -0.0296026821 0.0105436221 -2.807639 5.766316e-03
Measles            0.0002334783 0.0001108739  2.105801 3.715997e-02
Hepatitis.B        0.0945546739 0.0225639711  4.190516 5.126730e-05
StatusDeveloping  -9.6781992792 1.6134882158 -5.998308 1.879674e-08
```

a. Write the fitted model.

$\hat{Y} = 71.6 - 0.03(\text{infant deaths}) + 0.0002(\text{measles}) + 0.09(\text{HepB}) - 9.7(\text{Status = Developing})$

b. Write an interpretation for the coefficient estimate of the Status variable.

Holding all else constant, average life expectancy in years is 9.7 years lower in developing countries than in developed countries on average.

Next, you fit a model regressing life expectancy on infant deaths, Measles, Hepatitis B, status, and an interaction term for infant deaths and status. Answer questions c-e based on this model.

c. Write the full theoretical model. Be sure to define each term (e.g., "Y = ——")

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_4 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Y = Avg life expectancy (years)

X_1 = infant deaths

X_2 = Measles

X_3 = Hep B

X_4 = 1 if status = developing, 0 otherwise

d. Write the implied theoretical models for developed and developing countries. You do not need to define the terms again.

Developed: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

$x_4 = 0$

Developing: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 + \beta_5 x_1 + \varepsilon$

$x_4 = 1$ $\quad Y = (\beta_0 + \beta_4) + (\beta_1 + \beta_5) x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

e. What is the dimension of the design matrix for this model?

sample size n $\times$ # parameters

134×6

Blank Page - End of exam