DS 2026 Final Exam Prep Questions

NAME: Kalenga Mumba

1. Compare and contrast different definitions of probability, illustrating differences with simple examples

   - Probability is a fundamental mathematical concept that quantifies the likelihood of an event occurring.
   - Classical Probability –> refers to the assumption that all outcomes in a sample space are equally likely. An example would include rolling a 6-sided dice, the probability of rolling a 3 is P(rolling a 3) = 1/6
   - Frequency Probability –> refers to the likelihood of an event as the limit of its relative frequency over many trials. An example includes flipping a coin 100 times and observe it lands on heads 52 times, therefore P(heads) = 52/100 = 0.52
   - Bayesian Probability –> P(A|B) = P(B|A) * P(A) / P(B)

2. Express the rules of probability verbally, mathematically, and computationally

   - The probability of any event is always a non-negative number (falling between 0 and 1) and P(A) > 0
   - The total probability of all possible outcomes in the sample space is equal to 1, if S is sample space, then P(S) = 1
   - The probability that A or B occurs is the sum of their individual probabilities: P(A u B) = P(A) + P(B)
   - The probability that an event A does not occur is equal to one minus the probability that it does occur. This can be stated as: P(A') = 1 - P(A) where A' denotes the complement of A.
   - For two independent events A and B (events where the occurrence of one does not affect the other), the probability that both events occur simultaneously is given by: P(A n B) = P(A) * P(B)

3. Illustrate the rules of probability with examples

   - Addition Rule

     Example: Suppose you have a standard six-sided die. Let event A be rolling a 2, and event B be rolling a 4. Since these two outcomes cannot occur simultaneously (they are mutually exclusive), we can use the addition rule:

     P(A) = Probability of rolling a 2 = 1/6

     P(B) = Probability of rolling a 4 = 1/6

     Using the addition rule:

     P(A u B) = P(A) + P(B) = (1/6) + (1/6) = 2/6 = 1/3

     Thus, the probability of rolling either a 2 or a 4 on a single roll of a die is **1/3**.

- Multiplication rule

  Example: Consider flipping a fair coin and rolling a six-sided die simultaneously. Let event A be getting heads on the coin flip, and event B be rolling a 5 on the die.

  P(A) = Probability of getting heads = 1/2

  P(B) = Probability of rolling a 5 = 1/6

  Since these two events are independent:

  P(A n B) = P(A) × P(B) = (1/2) × (1/6) = 1/12

  Therefore, the probability of getting heads on the coin flip and rolling a five on the die at the same time is **1/12**.

- Complementary Rule

  Example: Using our previous example with a six-sided die, let's find out the probability that you do not roll a number greater than four (i.e., you roll either a 1, 2, or 3).

  First, calculate P(A), where A is rolling greater than four (which includes outcomes {5,6}).

  Outcomes favorable to A: {5,6} → Number of favorable outcomes for A = 2

  Total outcomes when rolling one die = 6

  Thus,

  P(A) = Number of favorable outcomes / Total outcomes = 2/6 = 1/3

  Now applying the complementary rule:

  P(A') = 1 - P(A) = 1 - (1/3) = 2/3

  So, the probability that you do not roll greater than four is **2/3**.

- Conditional Rule

  Example: Given a bag containing 6 red marbles, 3 blue marbles, and 1 green marble, this gives a total of 10 marbles in the bag.

  Event A: Drawing a red marble from the bag.

  Event B: Drawing a marble that is not green.

  To find P(B), determine how many marbles are not green. There are 9 marbles that are not green (6 red + 3 blue).

  So, P(B) = number of non-green marbles / total number of marbles = 9/10

  To find P(A n B), which represents the probability that both events occur, drawing a red marble (A) while also drawing a marble that is not green (B). Since all red marbles are not green, all occurrences of event A fall under event B.

  The number of favorable outcomes for both events happening together is simply the number of red marbles, which is 6.

  So, P(A n B) = number of red marbles / total number of marbles = 6/10 or 3/5

4. Using long-run proportion definition of probability, derive the univariate rules of probability
5. organize/express bivariate random variables in cross tables
6. Define joint, conditional, and marginal probabilities

- **Joint Probability**
  Joint probability refers to the probability of two or more events occurring simultaneously. It is denoted as P(AnB) for events A and B. For example, if we want to find the probability that a card drawn from a standard deck is both red and a four, we can calculate this as follows: there are two red fours in a deck of 52 cards (the 4 of hearts and the 4 of diamonds), so the joint probability would be P(four and red)=2/52=126

- **Marginal Probability**
  Marginal probability is the probability of an event occurring without consideration of any other events. It is often referred to as unconditional probability. For instance, if we want to find the marginal probability of drawing a red card from a deck, we can calculate it as follows: there are 26 red cards in a standard deck of 52 cards, so the marginal probability would be P(red)=26/52=12. This calculation does not depend on any other event.

- **Conditional Probability**
  Conditional probability is the probability of one event occurring given that another event has already occurred. It is denoted as P(A|B), which reads as "the probability of A given B." For example, if we want to find the conditional probability that a card drawn is a four given that it is red, we can calculate it by considering only the red cards. There are 26 red cards total, and among them, there are 2 fours (the 4 of hearts and the 4 of diamonds). Thus, the conditional probability would be calculated as P(four | red)=2/26=1/13

7. Identify joint, conditional, and marginal probabilities in cross tables

8. Identify when a research question calls for a joint, conditional, or marginal probability

A. Joint Probability Scenarios Use joint probability when:

- The question involves determining the likelihood of two specific outcomes happening together.
- Example: "What is the probability that a randomly selected person is both a smoker and has high blood pressure?" Here, you need to calculate P(Smoker n High Blood Pressure).

B. Conditional Probability Scenarios Use conditional probability when:

- The question specifies a condition or context under which an outcome occurs.
- Example: "What is the probability that a person has high blood pressure given that they are a smoker?" In this case, you would calculate P(High Blood Pressure | Smoker).

C. Marginal Probability Scenarios Use marginal probability when:

- The question seeks to find out how likely an event is without regard to any other factors.
- Example: "What is the overall probability that a randomly selected person has high blood pressure?" Here, you would simply look for P(High Blood Pressure) across all individuals in your sample.

9. Describe the connection between conditional probabilities and prediction

Conditional probabilities play a crucial role in the field of prediction, particularly in statistical modeling and machine learning. At its core, conditional probability refers to the likelihood of an event occurring given that another event has already occurred. This concept is mathematically expressed as P(A|B), which denotes the probability of event A occurring given that event B has occurred. In predictive modeling, understanding how different variables interact with one another is essential for making accurate forecasts. For instance, if we want to predict whether a customer will purchase a product (event A), we might consider various factors such as their previous purchasing behavior or demographic information (event B). By calculating the conditional probabilities associated with these factors, we can refine our predictions about future purchases.

10. Derive Bayes rule from cross tables

11. Apply Bayes rules to answer research questions

12. Apply cross table framework to the special case of binary outcomes with special attention to Sensitivity, Specificity, Positive predictive value, Negative predictive value, Prevalence, Incidence

13. Define/describe confounding variables, including Simpson's paradox, DAGs, causal pathway

Definition of Confounding Variables

A confounding variable is an extraneous factor in a statistical model that correlates with both the independent variable (the variable being manipulated) and the dependent variable (the outcome being measured). This correlation can lead to a false assumption about the relationship between the independent and dependent variables. For instance, if researchers are studying the effect of exercise on weight loss, a confounding variable could be diet; individuals who exercise more might also have healthier diets, which independently affects weight loss.

Confounding variables can obscure true relationships and lead to incorrect conclusions if not properly controlled for in research studies. They are particularly problematic in observational studies where random assignment is not possible.

Simpson's Paradox

Simpson's Paradox occurs when a trend appears in several different groups of data but disappears or reverses when these groups are combined. This phenomenon highlights how confounding variables can influence results. For example, consider two hospitals treating patients for a specific condition. If Hospital A has a higher success rate than Hospital B within each gender group (males and females), but when combined, Hospital B shows a higher overall success rate, this could be due to differences in the distribution of genders treated by each hospital (a confounding variable).

This paradox illustrates the importance of considering how data is aggregated and emphasizes that conclusions drawn from aggregated data may not reflect true relationships present within subgroups.

Directed Acyclic Graphs (DAGs)

Directed Acyclic Graphs (DAGs) are graphical representations used to illustrate causal relationships between variables. In a DAG, nodes represent variables while directed edges (arrows) indicate causal influences from one variable to another. The acyclic nature means there are no loops; you cannot return to a node once you leave it.

DAGs help researchers visualize potential confounders and identify pathways through which causation may occur. By mapping out these relationships, researchers can better understand how to control for confounding variables in their analyses. They serve as powerful tools for clarifying assumptions about causality and guiding statistical modeling decisions.

Causal Pathway

A causal pathway refers to the sequence of events or mechanisms through which an independent variable influences a dependent variable. Understanding causal pathways is essential for identifying direct effects versus those mediated by other variables (including confounders).

For example, consider the relationship between education level (independent variable) and income level (dependent variable). The causal pathway might include factors such as job opportunities, skills acquired through education, and social networks that facilitate employment—all of which mediate the relationship between education and income.

Identifying these pathways allows researchers to discern whether observed associations are direct or indirect effects influenced by other factors. It also aids in designing interventions aimed at altering specific parts of the pathway to achieve desired outcomes.

In summary, confounding variables complicate our understanding of relationships between variables; Simpson's Paradox demonstrates how aggregation can obscure true trends; DAGs provide visual frameworks for analyzing causal relationships; and causal pathways elucidate mechanisms underlying observed associations.

14. Describe approaches for avoiding or addressing confounding, including stratification and randomization

- **Identification of Confounders**: The first step is to identify potential confounders through literature review, expert consultation, or exploratory data analysis. This involves understanding the relationships between variables and recognizing which factors may distort the true relationship being studied.
- **Statistical Control**: Once confounders are identified, statistical techniques such as multivariable regression models can be used to adjust for these variables. By including confounders in the model, researchers can isolate the effect of the primary independent variable on the dependent variable.
- **Matching**: In observational studies, matching participants based on confounding variables (e.g., age, gender) ensures that groups are comparable. This can be done through techniques like propensity score matching where individuals with similar characteristics are paired.
- **Restriction**: Researchers can restrict their study population to individuals who fall within certain categories of a confounder (e.g., only studying non-smokers if smoking is a confounder).
- **Stratified Analysis**: After stratifying the data by potential confounders, researchers analyze each stratum separately. This allows for examination of associations within more homogeneous groups and helps clarify whether observed effects differ across strata.
- **Random Assignment**: Participants are randomly assigned to treatment or control groups. This process ensures that any differences between groups are due to chance rather than systematic biases related to confounding variables.
- **Blinding**: Implementing blinding (single or double) minimizes bias in treatment administration and outcome assessment. Participants and/or researchers do not know which group participants belong to, reducing expectations that could influence results.
- **Sample Size Calculation**: Adequate sample size is crucial in randomization studies to ensure sufficient power to detect meaningful differences between groups while minimizing Type I and Type II errors.
- **Intention-to-Treat Analysis**: Analyzing participants based on their original group assignment regardless of adherence helps maintain randomization benefits and reduces bias introduced by dropouts or protocol deviations.

15. List various data types (nominal, ordinal, interval, ratio, discrete, continuous) and match each data type with probability models that may describe it

1. Nominal Data Nominal data represents categories without any intrinsic ordering. Each category is distinct and does not have a numerical value associated with it. Examples include:

- Gender (male, female)
- Eye color (blue, brown, green)

In nominal data, you can count the frequency of occurrences but cannot perform mathematical operations.

- **Multinomial Distribution**: This distribution generalizes the binomial distribution to more than two categories. It describes the probabilities of counts across multiple categories.

2. Ordinal Data Ordinal data involves categories that have a meaningful order or ranking but do not have consistent intervals between them. While you can say one category is higher or lower than another, the difference between ranks is not uniform. Examples include:

- Education level (high school, bachelor's degree, master's degree)
- Satisfaction ratings (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)

In ordinal data analysis, you can determine the median and mode but not the mean due to the lack of consistent intervals.

- **Ordinal Logistic Regression**: This model is used when the dependent variable is ordinal. It estimates the probabilities of different outcomes based on predictor variables.
- **Cumulative Distribution Function (CDF)**: For ordinal data, CDFs can help describe the probability of a variable being less than or equal to a certain value in an ordered set.

3. Interval Data Interval data has both order and equal intervals between values but lacks a true zero point. This means that while you can add and subtract values meaningfully, you cannot multiply or divide them in a way that reflects true ratios. Examples include:

- Temperature in Celsius or Fahrenheit
- Dates on a calendar

For instance, 20°C is not twice as hot as 10°C; it merely indicates a difference in temperature.

- **Normal Distribution**: Often used for interval data when it is assumed that the data follows a bell-shaped curve.
- **T-distribution**: Used particularly when sample sizes are small and population standard deviations are unknown.

4. Ratio Data Ratio data possesses all the characteristics of interval data but includes a true zero point that allows for meaningful comparisons using multiplication and division. This means you can say one value is twice another. Examples include:

- Height
- Weight
- Distance

In ratio data analysis, all arithmetic operations are valid.

- **Normal Distribution**: Like interval data, ratio data can also often be modeled using normal distribution if it meets certain conditions.
- **Exponential Distribution**: This model can describe time until an event occurs (e.g., survival analysis), which is applicable to ratio-scaled measurements.

5. Discrete Data Discrete data consists of distinct or separate values that are countable and often represent whole numbers. There are no intermediate values between two adjacent points in discrete datasets. Examples include:

- Number of students in a classroom
- Number of cars in a parking lot

Discrete variables typically arise from counting processes.

- **Poisson Distribution**: This distribution describes the number of events occurring within a fixed interval of time or space and is suitable for modeling count-based discrete variables.
- **Binomial Distribution**: Used when there are two possible outcomes (success/failure) in a fixed number of trials.

6. Continuous Data Continuous data can take any value within a given range and can be measured rather than counted. It includes fractions and decimals and represents measurements that can be infinitely divided into smaller parts. Examples include:

- Height
- Time
- Temperature

Continuous variables arise from measuring processes where precision may vary.

- **Normal Distribution**: Commonly used for continuous variables due to its properties and central limit theorem implications.
- **Uniform Distribution**: When all outcomes in a range are equally likely, this model applies to continuous variables.
- **Exponential Distribution**: Also applicable here for modeling time until an event occurs in continuous scenarios.

17. Discuss the degree to which models describe the underlying data

Nominal Data -> **Degree of Description:** The multinomial distribution effectively captures the frequency of occurrences across different categories but does not provide insights into relationships or rankings among categories. It simply reflects the proportion of each category in the dataset.

Ordinal Data -> **Degree of Description:** Ordinal logistic regression provides a good fit for ordinal data by acknowledging the inherent order among categories while accounting for varying distances between them. However, it may not fully capture nuances if the intervals between ranks are not uniform.

Interval Data -> **Degree of Description:** The normal distribution can describe interval data well if it follows a bell-shaped curve; however, it assumes symmetry and may not fit skewed distributions accurately. The t-distribution provides a better fit for smaller samples but also relies on assumptions about normality.

Ratio Data -> **Degree of Description:** Both exponential and log-normal distributions can effectively describe ratio data, particularly in contexts where measurements have meaningful zero points and ratios are interpretable. However, their applicability depends on whether the underlying assumptions (e.g., independence, distribution shape) hold true.

Discrete Data -> **Degree of Description:** The Poisson distribution works well for rare events over time or space, while binomial distribution applies to scenarios with fixed trials and constant success probabilities. Both models can accurately reflect discrete datasets but may oversimplify complex behaviors if additional factors influence counts.

Continuous Data -> **Degree of Description:** The normal distribution often describes continuous data effectively; however, it may not fit all datasets, especially those that exhibit skewness or kurtosis beyond what normality allows. The uniform distribution serves well in cases where every outcome has equal likelihood but lacks flexibility in capturing real-world variability.

18. Tease apart model fit and model utility

1. Definition of Model Fit

Model fit refers to how well a statistical model describes the data it is intended to predict. In machine learning, this is often assessed through various metrics that evaluate the accuracy of predictions made by the model compared to actual outcomes. The process of fitting a model involves adjusting its parameters so that it minimizes the difference between predicted values and actual values from the training dataset.

2. Definition of Model Utility

Model utility goes beyond just how well a model fits the training data; it encompasses how useful or effective a model is when applied to new, unseen data. This concept is crucial because a model can have excellent fit on training data but perform poorly on validation or test datasets due to overfitting.

19. Express probability models both mathematically, computationally, and graphically (PMF/PDF CMF/CDF, quantile function, histogram/eCDF)

Probability models can be expressed mathematically through various functions that describe the likelihood of different outcomes in a random experiment. The two primary types of probability distributions are discrete and continuous.

- **Probability Mass Function (PMF)**: For discrete random variables, the PMF is defined as P(X=x), where X is a discrete random variable and x is a specific value. The PMF must satisfy two conditions:
  1. P(X=x) greater than or equal to 0 for all x
  2. P(X=x)=1
- **Cumulative Distribution Function (CDF)**: This function gives the probability that a random variable takes on a value less than or equal to x:

  F(x)=P(X less than or equal to x)

Q. Suppose the yearly hospital charges (in thousands of dollars) for a randomly selected UVA student is a mixture distribution.

For 60% of students, the hospital charges will be \$0. For the remaining 40% of students, the hospital charges are a random variable described by a gamma distribution with shape = 2 and scale = 2. (Again, in thousands of dollars.)

The following function mimics the hospital charge distribution. It generates draws of the random variable. Use the function to generate an expression for the CDF and quantile functions of the random variable.

```
rhc <- function(n){ rgamma(n,shape=2,scale=2)*rbinom(n,1,.4) }
```

```
set.seed(123)
n_samples <- 10000
samples <- rhc(n_samples)

cdf_function <- function(x){
    if (x < 0){
        return(0)
    } else if (x == 0){
        return(0.6)
    } else {
        return(0.6 + (0.4 * pgamma(x, shape=2, scale=2)))
    }
}
```

```r
quantile_function <- function(p){
    if (p < 0.6){
        return(0)
    } else {
        return(qgamma(p - 0.6 / 0.4, shape=2, scale=2))
    }
}

cdf_value <- cdf_function(3)
quantile_value <- quantile_function(0.8)
```

```
## Warning in qgamma(p - 0.6/0.4, shape = 2, scale = 2): NaNs produced
```

```r
print(cdf_value)
```

```
## [1] 0.7768698
```

```r
print(quantile_value)
```

```
## [1] NaN
```

Q. Consider earnings (in thousands of dollars) the first year after graduation from UVA with an undergraduate degree. If X is normal with $\mu = 60$ and $\sigma = 10$, what level of earnings represents the top 90th percentile?

```r
mu = 60
sigma = 10

top_90th_percentile <- qnorm(0.90, mu, sigma)
top_90th_percentile
```

```
## [1] 72.81552
```

Q. Suppose an upcoming election for UVA student body president is between two candidates. In a survey of 30 students, 20 voiced support for candidate A. Use the Desmos calculator (link) to fit a probability model with Bayesian methods for the election, specifically the probability that candidate A is the preferred by the student body. Report the 95% credible interval. (In this calculator, H_1 is the number of supporters for candidate A and T_1 is the number of supporters for candidate B.)

10 for H1 and 4 for T1

Q. Suppose an upcoming election for UVA student body president is between two candidates. In a survey of 30 students, 20 voiced support for candidate A. Use the Desmos calculator (link) to fit a probability model with Maximum Likelihood for the election, specifically the probability that candidate A is the prefered by the student body. Report the 1/20 support interval. (In this calculator, n is the total number of respondants, h is the number that voice support for candidate A.)

n is 21 and h is 1

Q. Repeat the election analysis performed above with additional data. In a survey of 100 students, 60 students voiced support for candidate A. Compare the interval estimates based on the larger dataset to those generated from the smaller dataset. Comment on which analysis you find more persuasive and explain why.

Smaller Dataset Results:

- Bayesian Credible Interval: **(0.617, 0.883)**
- MLE Support Interval: **(0.642, 0.692)**

Larger Dataset Results:

- Bayesian Credible Interval: **(0.558, 0.748)**
- MLE Support Interval: **(0.575, 0.625)**

  When comparing intervals derived from both datasets:

  1. **Width of Intervals**: The intervals generated from the larger dataset are narrower than those from the smaller dataset due to increased sample size leading to more precise estimates.
  2. **Credibility and Confidence**: The larger sample size provides stronger evidence regarding candidate A's support level among students because it reduces uncertainty associated with sampling error.
  3. **Consistency Across Methods**: Both Bayesian and MLE methods yield overlapping intervals in their estimates for candidate A's support in both datasets; however, they are more tightly clustered in the larger dataset.
  4. **Statistical Power**: The larger dataset enhances statistical power and reliability of conclusions drawn about candidate preferences among students.

  Overall, I find the analysis based on the larger dataset more persuasive due to its reduced uncertainty and tighter confidence intervals reflecting a more accurate representation of student preferences regarding candidate A's support.

Q. Going back to the election question, suppose that the support for candidate A was known to be $p = 0.55$. In an election in which 100 students vote, what is the probability that 51 or more votes will be cast for candidate A?

```
n <- 100
probability <- sum(dbinom(51:n, 100, 0.55))
print(probability)
```

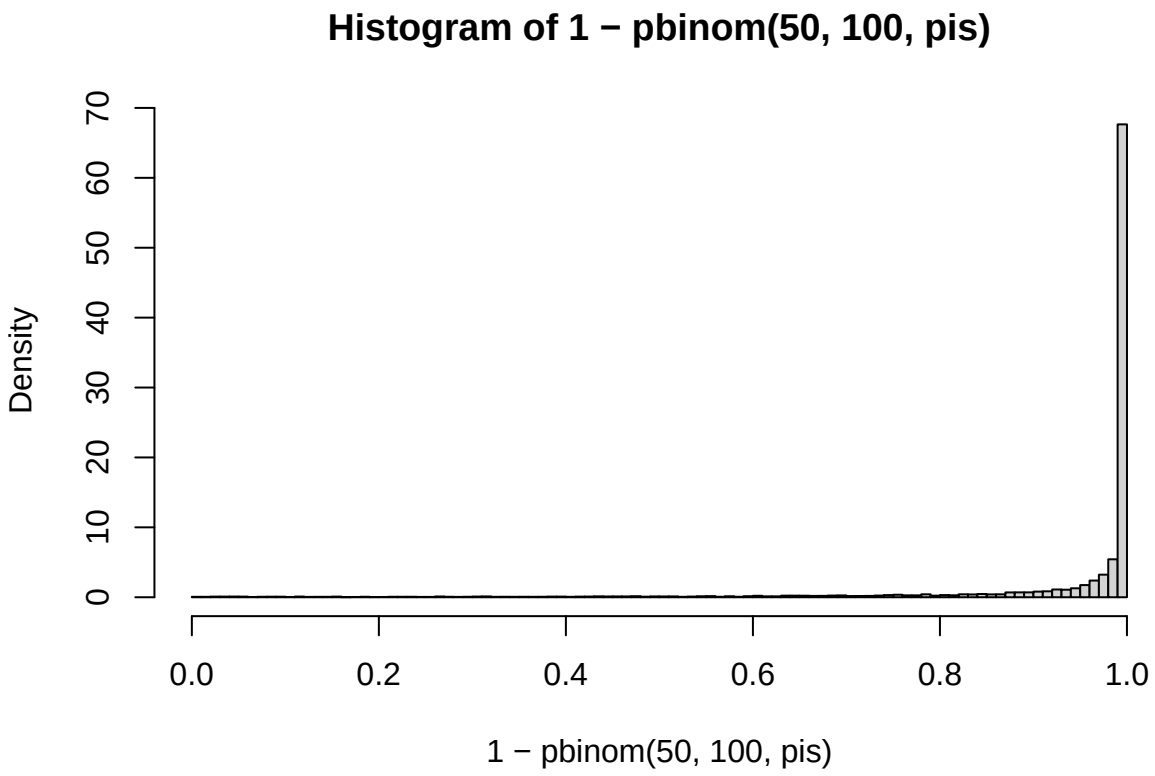
```
## [1] 0.8172718
```

Q. Now suppose the the probability is unknown, and is estimated from data. The following shows the distribution for $P(\text{Votes}>50)$ when estimated from data using a uniform prior and a survey of 30 students with 20 voicing support for the candidate. Add a line to show the solution when p is known. Comment on the uncertainty when p is estimated from data.

```
pis <- rbeta(10000, 21, 11)
mean(1 - pbinom(50, 100, pis))
```

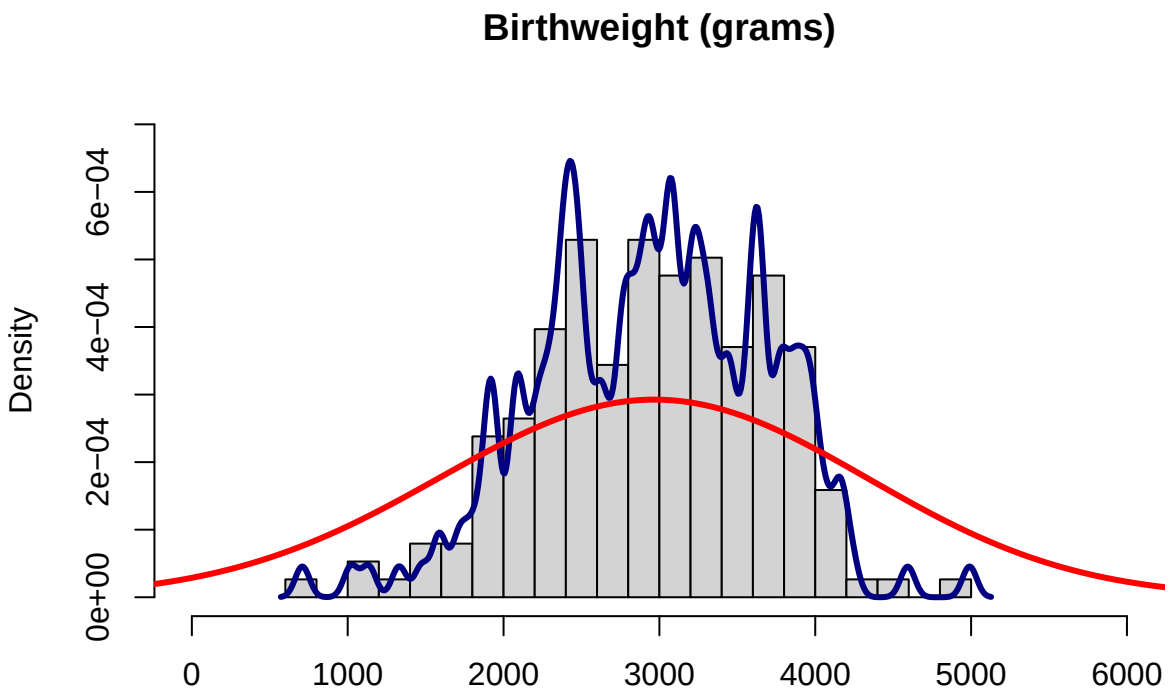
```
## [1] 0.939132
```

```
hist(1-pbinom(50,100,pis), freq=FALSE, breaks=100)
```



Q. Consider the following estimates of the PDF for infant birthweight. Both are poorly fitting estimates. Explain the concepts of overfitting and underfitting in the context of the birthweight data.

```
d1 <- MASS::birthwt
hist(d1$bwt, breaks=20, freq=FALSE, xlim = c(0,6000), ylim = c(0,0.0007), main = "Birthweight (grams)",
lines(density(d1$bwt, adjust = 1/5), lwd = 3, col = "navy")
lines(density(d1$bwt, adjust=5), lwd = 3, col = "red")
```



Overfitting Explained

Overfitting occurs when a model is too complex relative to the amount of data available. In the context of estimating the PDF for infant birthweight, an overfit model would closely follow every fluctuation in the training data, capturing noise rather than the true underlying distribution.

For example, if we use a very flexible density estimation method with a high bandwidth adjustment (like using `adjust=0.1`), the resulting curve might have many peaks and valleys that correspond to random variations in the birthweight data rather than actual trends. This could lead to a model that performs well on the training dataset but poorly on new, unseen data because it fails to generalize beyond the specific instances it was trained on.

In practical terms, an overfit density estimate might show extreme values or irregular shapes that do not reflect realistic birthweight distributions. This could mislead healthcare professionals regarding typical birthweights and associated risks.

Underfitting Explained

Underfitting occurs when a model is too simple to capture the underlying structure of the data adequately. In our case, if we were to use a very rigid method for estimating density, such as assuming a uniform distribution across all observed birthweights—the resulting PDF would likely miss important features of the actual distribution.

For instance, using a very high bandwidth adjustment (like adjust=10) could smooth out all variations in birthweight data, leading to a flat line that fails to represent any meaningful trends or characteristics of infant birthweights. This underfit model would not only provide poor predictions but also fail to inform about critical aspects like average weights or weight distributions across different populations.

Q. Explain the concept of generalizability in the context of the birthweight data.

In the context of birthweight data, generalizability is crucial for understanding how well the insights gained from analyzing a particular sample of infants can inform healthcare practices and policies for all newborns. For instance, if researchers analyze birthweight data from a specific hospital or region, they must consider whether the characteristics of that sample, such as maternal health, socioeconomic status, and environmental factors, are representative of the larger population. If the sample is biased or too homogeneous, any conclusions drawn may not accurately reflect trends in different demographics or geographic areas. This limitation can affect predictions about average birthweights, risk factors for low birthweight, and interventions aimed at improving infant health outcomes.

Q: The Monte Hall problem is a classic game show. Contestants on the show where shown three doors. Behind one randomly selected door was a sportscar; behind the other doors were goats.

At the start of the game, contestants would select a door, say door A. Then, the host would open either door B or C to reveal a goat. At that point in the game, the host would ask the contestant if she would like to change her door selection. Once a contestant decided to stay or change, the host would open the chosen door to reveal the game prize, either a goat or a car.

In this problem, consider a **modified** version of the Monte Hall problem in which the number of doors is **variable**. Rather than 3 doors, consider a game with 4 or 5 or 50 doors. In the modified version of the game, a contestant would select an initial door, say door A. Then, the host would open **one** of the remaining doors to reveal a goat. At that point in the game, the host would ask the contestant if she would like to change her door selection. Once a contestant decided to stay or change, the host would open the chosen door to reveal the game prize, either a goat or a car.

Consider two strategies:

1. Always stay with the first door selected.
2. Always switch to the unopened door.

A. The function `game` below plays a single game of Monte Hall. The function returns a vector of length two, the first element is the prize under strategy 1 and the second element is the prize under strategy 2. The function has a single input parameter, N, which is the number of doors in the game.

Use the `game` function to estimate the probability that both strategies *simultaneously* result in a goat. Let **N=4**. (Note the word *simultaneously*. This means that in the same game, both strategies resulted in a goat.)

```
require(magrittr)
```

```
## Loading required package: magrittr
```

```
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
game <- function(N){
  if(N<3) stop("Must have at least 3 doors")
  prize <- sample(c(rep("goat",N-1),"car"), N)
  guess <- sample(1:N,1)
  game <- data.frame(door = 1:N, prize = prize, stringsAsFactors = FALSE) %>%
    mutate(first_guess = case_when(
      door == guess ~ 1
      , TRUE ~ 0
    )) %>%
    mutate(potential_reveal = case_when(
        first_guess == 1 ~ 0
      , prize == "car" ~ 0
      , TRUE ~ 1
    )) %>%
    mutate(reveal = 1*(rank(potential_reveal, ties.method = "random") == 3)) %>%
    mutate(potential_switch = case_when(
      first_guess == 1 ~ 0
      , reveal == 1 ~ 0
      , TRUE ~ 1
    )) %>%
    mutate(switch = 1*(rank(potential_switch, ties.method = "random") == 3))
  c(game$prize[game$first_guess == 1], game$prize[game$switch == 1])
}

#Estimating probability that both strategies simultaneously result in a goat when N=4
set.seed(123)

num_simulations <- 1000
simultaneous_goat <- 0

for (i in 1:num_simulations) {
  results <- game(4)
  if (all(results == "goat")) {
    simultaneous_goat <- simultaneous_goat + 1
  }
}

prob_result_goat <- simultaneous_goat / num_simulations
prob_result_goat
```

```
## [1] 0.376
```