DS 2026 Final Exam Prep Questions

Stefan Regalia

NAME: Please add your name here

NOTE: The `.rmd` version of the file is available here: (link)

Instructions

Please prepare reponses/solutions for the following questions. On the day of the exam, you will be given a new set of questions. You will use the solutions you've prepared for this exam during the exam.

During the exam, you will also be permitted to access the internet for publicly available content. You will not be allowed to communicate with anyone via the internet or any other means during the exam. This includes, but is not limited to:

- No messaging, emailing, or using social media to contact others.
- No posting questions or seeking answers on forums, chat rooms, chat bots (including large language models like ChatGPT), or any collaborative platforms.
- No sharing or discussing exam content with peers through any online or electronic medium.

You may **NOT** discuss any aspect of the exam or prep questions with anyone other than the instructor or TA. You may **NOT** share code or documents.

Submission instructions

1. Within your course repo, create a folder called `final-exam`
2. Within the folder, create the script file `exam-prep.rmd` with your solutions. Create a rendered report in `.pdf` output.
3. Add, commit, and push to your repo on github.com.

Questions

Exam questions are organized into sections corresponding to the learning outcomes of the course.

> compare and contrast different definitions of probability, illustrating differences with simple examples

Long-run proportion:

The long-run proportion is how often something happens when you repeat an experiment many times. For example, if you roll a die 10,000 times and it lands on the number 6 a total of 1,660 times, the long-run proportion of rolling a 6 is 1,600/10,000. The more you roll the die, the closer the long-run proportion will get to the true probability of 1/6. This type of probability is different from personal beliefs because it can always be proven by experiments and data.

Personal beliefs:

Probability that relies on personal beliefs is unlike long-run proportion probability because it can not always be proven by experiments and data, it is subjective. For example, if I believe there is an 30% probability that I go to dinner with my friends on a random night, I can not prove this because I can not conduct many repeated trials of an experiment.

Combination of beliefs and data:

There are certain probabilities that combine personal beliefs and data. For example, if you believe there is a 70% chance of rain tomorrow, this belief may be informed by weather forecasts and data (e.g., meteorological models and past weather patterns). However, unlike long-run proportions, this probability cannot be tested by repeated trials because tomorrow's weather is a one-time event. Advanced technologies, like weather prediction models, use data to provide a highly accurate estimation, but your belief might still adjust this based on your personal judgment or experience.

express the rules of probability verbally, mathematically, and computationally

AND probability: Refers to the probability of two events occurring simultaneously. For example, P(E and S) = P(E) x P(S) if the two events are independent.

```r
E <- 0.6
S <- 0.3

prob_E_and_S <- E * S
prob_E_and_S
```

```
[1] 0.18
```

OR probability: Refers to the probability of one event E occurring or another event S occurring, or both occurring simultaneously. P(E or S) = P(E) + P(S) - P(E and S) if they are not mutually exclusive. If they are mutually exclusive then P(E or S) = P(E) + P(S).

```r
prob_E_or_S <- E + S - prob_E_and_S
prob_E_or_S
```

```
[1] 0.72
```

Total probability: The law of total probability states that all probabilities involving a certain event E under different scenarios add up to 1.

Complement probability: The complement of event E is the probability that event E does not occur. It is denoted by E^c and is 1 - P(E).

Relative simulation error: The relative simulation error measures the accuracy of an estimate obtained through simulation relative to the true or expected value. It provides a sense of how close the simulated result is to the actual probability, normalized by the actual value.

(|P(sim) - P(true)|)/P(true)

```r
sim <- 0.3
true <- 0.5

sim_error <- abs(sim - true)/true
sim_error
```

```
[1] 0.4
```

Absolute simulation error: The absolute simulation error measures the difference between the simulated probability and the true probability. It gives a direct indication of how much the simulated result deviates from the actual value without considering its size relative to the true probability.

|P(sim) - P(true)|

using long-run proportion definition of probability, derive the univariate rules of probability

1. P(E) is between 0 and 1
2. P(Sample space) = 1
3. P(E^c) = 1 - P(E)

define joint, conditional, and marginal probabilities

Joint probability: the probability of two events occurring together.

Conditional probability: the probability of one event occurring given that another event has already occurred.

Marginal probability: The sum of all joint probabilities and is the probability of a single event occurring.

identify when a research question calls for a joint, conditional, or marginal probability

QUIZ ME

describe the connection between conditional probabilities and prediction

Conditional probabilities are foundational to prediction because they quantify the likelihood of an event occurring given that some related information is already known. In predictive contexts, the given information (conditioning variable) serves as the basis for making informed estimates about future or unknown outcomes.

derive Bayes rule from cross tables

QUIZ ME

apply Bayes rules to answer research questions

QUIZ ME

> 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

Sensitivity: True Positive Rate (Test positive given they have the disease)

Specificity: True Negative Rate (Test negative given they do not have the disease)

Positive Predictive Value: The probability an individual actually has a disease given they test positive.

Negative Predictive Value: The probability an individual does not have a disease given they test negative.

Prevalence: The proportion of individuals in the population who have the disease.

Incidence: the number of new cases of the disease in a given population over a specific time period.

define/describe confounding variables, including Simpson's paradox, DAGs, causal pathway

Confounding variables: A type of outside variable that influences the outcome of an experiment. Confounding variables affect the relationship between the independent and dependent variables.

Example of a Confounder:

- Suppose we study the relationship between coffee drinking and heart disease.
- **Observation**: Coffee drinkers have a higher rate of heart disease.
- **Confounder**: Smoking is associated with coffee drinking (exposure) and also increases the risk of heart disease (outcome).

Simpson's Paradox:

Simpson's Paradox is a phenomenon where an observed association between two variables reverses or changes when data is stratified by a confounding variable.

Example:

- **Data**: Two hospitals treat patients, and Hospital A appears to have a higher survival rate.
- **Stratification**: When stratified by patient severity:
  - Hospital B performs better for both severe and mild cases.
- **Confounder**: Hospital A treats mostly mild cases, creating a misleading aggregate survival rate.

Simpson's Paradox highlights the importance of stratifying data to account for confounders.

Directed Acyclic Graphs (DAGs)

DAGs are diagrams used to represent causal relationships between variables. They help identify confounding, mediating, and colliding variables.

- **Example DAG**:
  - Smoking (C) → Coffee Drinking (X)
  - Smoking (C) → Heart Disease (Y)

In this DAG, **smoking** is a confounder because it affects both coffee drinking and heart disease.

Causal Pathway

A causal pathway represents the sequence of events or variables that connect an independent variable (exposure) to a dependent variable (outcome).

Example:

- **Causal Pathway**: Physical Activity → Weight Loss → Lower Risk of Diabetes.
  - Weight loss is a **mediator** on the causal pathway.
- **Confounding Pathway**: Socioeconomic Status affects both Physical Activity and Risk of Diabetes, creating an alternate explanation.

> describe approaches for avoiding or addressing confounding, including stratification and randomization

Randomization: Randomization assigns participants to different groups (e.g., treatment and control) randomly, ensuring that confounding variables are evenly distributed across groups.

Stratification: Stratification involves dividing the data into subgroups (strata) based on levels of a confounding variable, then analyzing the association between the exposure and the outcome within each stratum.

list various data types (nominal, ordinal, interval, ratio, discrete, continuous)

match each data type with probability models that may describe it

Data Type	Definition	Examples	Probability Models
Nominal	Categories with no order.	Gender, blood type.	Multinomial, Categorical, Bernoulli.
Ordinal	Ordered categories without consistent intervals.	Education level, satisfaction.	Ordinal Logistic, Proportional Odds.
Interval	Numeric, meaningful intervals, no true zero.	Temperature (°C), years.	Normal, T-Distribution, GLMs.
Ratio	Numeric, meaningful intervals, true zero.	Weight, height, income.	Normal, Log-Normal, Exponential, Poisson.
Discrete	Countable, separate values.	Number of students, calls.	Binomial, Poisson, Geometric, Hypergeometric.
Continuous	Any value within a range.	Time, distance, temperature (K).	Normal, Exponential, Uniform, Weibull.

discuss the degree to which models describe the underlying data

tease apart model fit and model utility

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

Feature	**Discrete (PMF)**	**Continuous (PDF)**	**Cumulative (CDF/eCDF)**	**Quantile Function**

Mathematical	P(X=x)	f(x)	F(x)=P(X x)	Q(p)=F^−1(p)

Computational	`pmf(x)`	`pdf(x)`	`cdf(x)`	`ppf(p)` (percent-point function)

Graphical	Bar plot	Curve	Line plot (eCDF approximates CDF)	Plot cumulative probability vs. values

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.

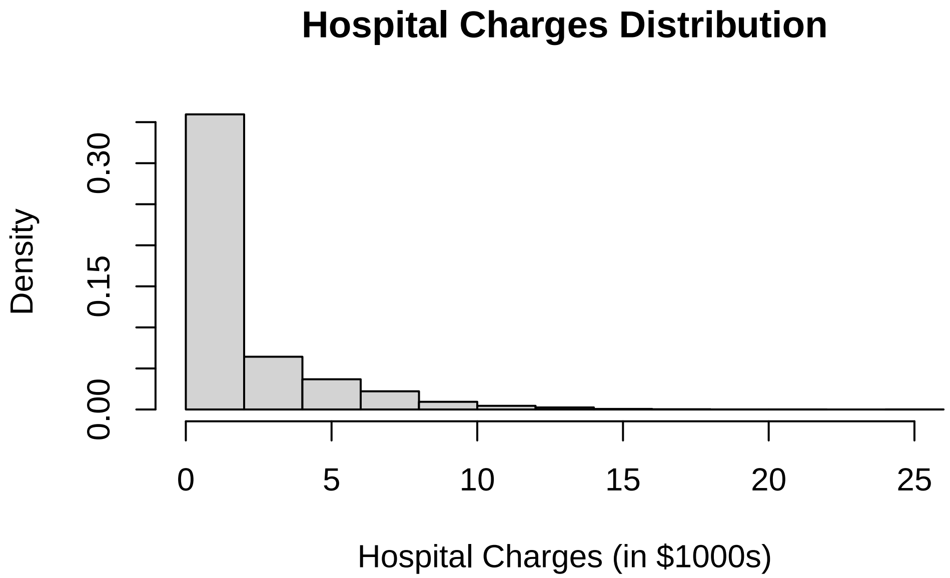
```r
rhc <- function(n){ rgamma(n,shape=2,scale=2)*rbinom(n,1,.4) }
```

```r
gamma_cdf <- function(x, shape = 2, scale = 2) {
  pgamma(x, shape = shape, scale = scale)
}

hospital_cdf <- function(x) {
  ifelse(
    x == 0,
    0.6,
    0.6 + 0.4 * gamma_cdf(x, shape = 2, scale = 2)
  )
}
```

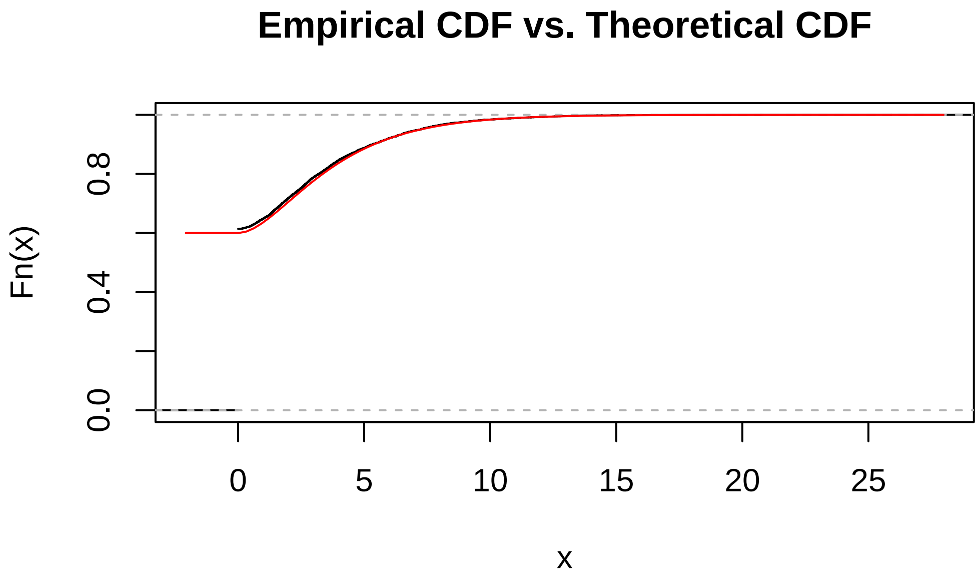
```r
gamma_quantile <- function(p, shape = 2, scale = 2) {
  qgamma(p, shape = shape, scale = scale)
}

hospital_quantile <- function(p) {
  ifelse(
    p <= 0.6,
    0,
    gamma_quantile((p - 0.6) / 0.4, shape = 2, scale = 2)
  )
}
```

```r
set.seed(378)

n <- 10000
samples <- rhc(n)
```

```
ecdf_samples <- ecdf(samples)

hist(samples, probability = TRUE, main = "Hospital Charges Distribution",
     xlab = "Hospital Charges (in $1000s)")
```



```
plot(ecdf_samples, main = "Empirical CDF vs. Theoretical CDF")
curve(hospital_cdf(x), col = "red", add = TRUE)
```


Empirical CDF vs. Theoretical CDF

Summary

- **CDF**: Combines the constant probability at X=0 and the gamma distribution for X>0.
- **Quantile Function**: Combines the fixed quantile for p 0.6 and the gamma quantile for p>0.6.
- Computationally, the CDF and quantile functions are implemented using pgamma and qgamma for the gamma portion, scaled by the mixture weights.

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
qnorm(0.9, 60, 10)
```

```
[1] 72.81552
```

explain different approaches for fitting probability models from data (Maximum likelihood, Bayesian posterior, Method of Moments, Tuning, Kernel Density Estimation)

REVIEW COURSE NOTES

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.)

After changing H1 to 20 and T1 to 10 and changing the interval to 95%, the credible interval became [0.49, 0.81]

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.)

After changing n to 30 and h to 20, as well as k to 0.05, the support interval became [0.44, 0.85] explore how to communicate uncertainty when constructing models and answering research questions (support intervals, credible intervals)

1. Types of Intervals to Communicate Uncertainty

a. Support Intervals (Likelihood-Based)

- **Definition**: A support interval is a range of parameter values where the likelihood remains above a certain fraction (e.g., 1/20) of the maximum likelihood.
- **Purpose**: Reflects the range of plausible parameter values based on the likelihood function.
- **Use Cases**:
  - Suitable for frequentist models.
  - Highlights regions of parameter space strongly supported by the data.
- **Example**:
  - For a binomial likelihood, the 1/20 support interval shows the range of probabilities for which the likelihood is at least 1/20 of its peak value.

b. Credible Intervals (Bayesian)

- **Definition**: A credible interval is the range of parameter values that contain a specified proportion (e.g., 95%) of the posterior probability mass.
- **Purpose**: Reflects uncertainty about parameters using Bayesian posterior distributions.
- **Use Cases**:
  - Suitable for Bayesian models.
  - Directly interpretable: "There is a 95% probability that the parameter lies within this interval."
- **Example**:
  - For a Beta distribution describing support for a candidate, the 95% credible interval may be [0.49,0.81].

2. Methods for Communicating Uncertainty

a. Numerical Representation

Provide explicit ranges with interpretative context:

- Example for support intervals: "The likelihood-based support interval for the probability of candidate A being preferred is [0.46, 0.84]."
- Example for credible intervals: "The Bayesian 95% credible interval for the probability of candidate A being preferred is [0.49, 0.81]."

b. Visual Representation

1. **Confidence/Support/Credible Interval Plots**:
   - Overlay the interval on the likelihood or posterior curve.
   - Highlight the interval bounds using vertical lines or shaded regions.
   - Example: Plot a posterior distribution with a 95% credible interval shaded in blue.
2. **Error Bars**:
   - Add error bars to parameter estimates or predictions to show the range of uncertainty.
3. **Prediction Intervals**:
   - Plot predicted outcomes with shaded regions to indicate variability.

c. Textual Interpretation

Explain the interval in plain language:

- Example: "Based on the data, we estimate that the true proportion of students supporting candidate A is likely between 46% and 84% (support interval), or between 49% and 81% (credible interval)."

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.

Dataset	**Sample Size**	**MLE** (p^hat)	**Support Interval**	**95% Credible Interval**
Smaller Dataset	n=30	0.6667	[0.46,0.84]	[0.49,0.81]
Larger Dataset	n=100	0.6	[0.51,0.68]	[0.51,0.68]

Conclusion

- **Smaller Dataset**: Higher uncertainty, wider intervals, less persuasive.
- **Larger Dataset**: More precise, narrower intervals, and more robust conclusions.

The larger dataset's interval estimates of [0.51,0.68] provide a clearer picture of the proportion of students supporting candidate A. This reinforces the importance of larger sample sizes in reducing uncertainty and enhancing the reliability of statistical inferences.

propagate uncertainty in simulations, visualize the uncertainty inherent in fitting probability models from data

Propagating uncertainty in simulations involves incorporating the variability of input parameters and observing how this variability affects the outcomes. When fitting probability models, uncertainty arises due to limited data, model assumptions, and random variation. Properly visualizing this uncertainty ensures transparent communication of the model's limitations.

---

1. Propagating Uncertainty in Simulations

Steps for Propagation

1. **Define Input Distributions**:
   - Specify distributions for uncertain parameters (e.g., means, variances, probabilities).
   - For example, if you're estimating the mean of a normal distribution, use a posterior distribution from Bayesian analysis or confidence intervals from frequentist analysis.
2. **Generate Random Samples**:
   - Draw samples from the input distributions to reflect uncertainty.
   - Use Monte Carlo methods to run multiple simulations, incorporating this uncertainty.
3. **Simulate Outcomes**:
   - Use the sampled input parameters to simulate the outcomes of the model.
   - Record the variation in outputs.
4. **Analyze Results**:
   - Summarize results using histograms, confidence intervals, and other uncertainty measures.

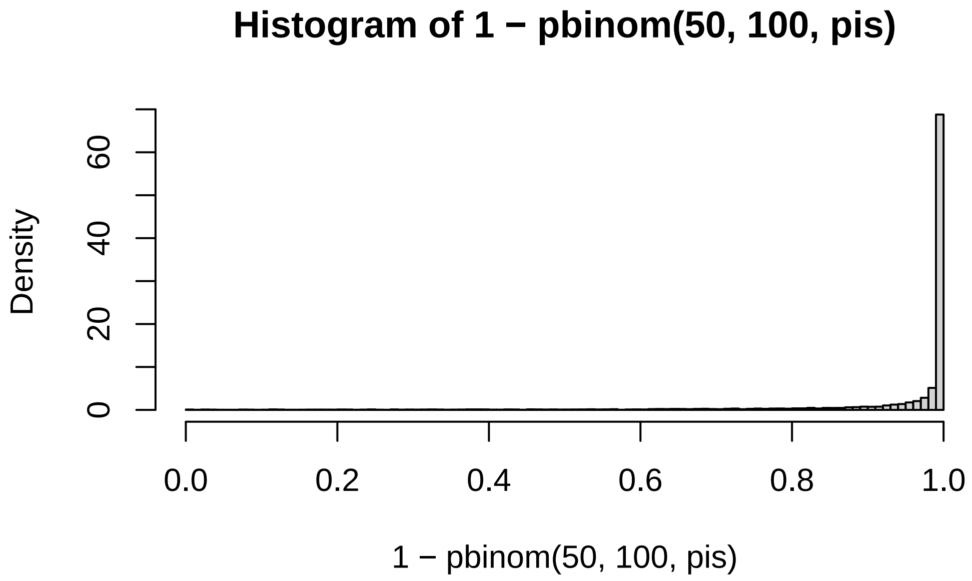
Q. Going back to the election question, suppose that the support for candidate A was known to be $p = 0.55$. In an election in which 100 students vote, what is the probability that 51 or more votes will be cast for candidate A?

```r
prob <- 1 - pbinom(50, 100, 0.55)
prob
```

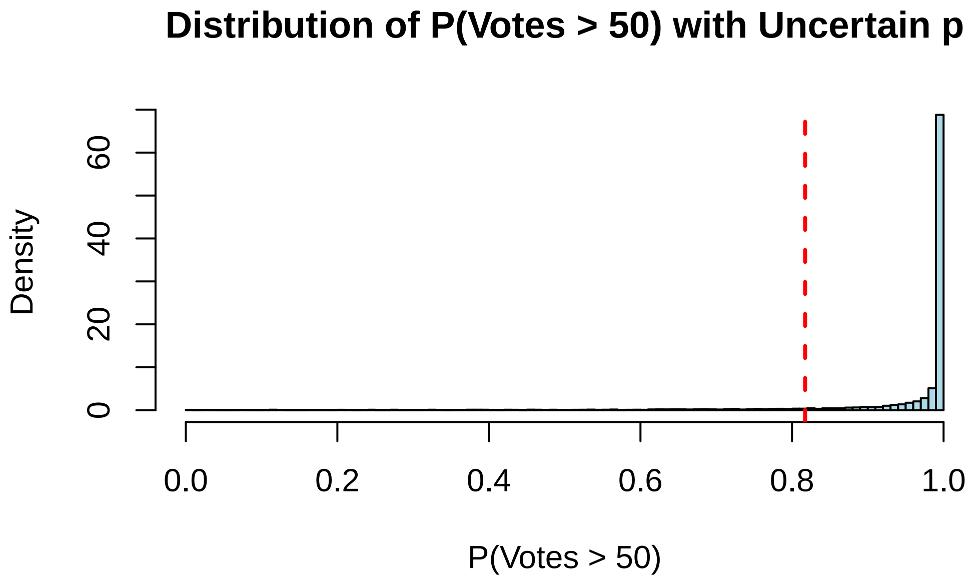
```
[1] 0.8172718
```

Q. Now suppose the the probability is unknown, and is estimated from data. The following shows the distribution for P(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)
hist(1-pbinom(50,100,pis), freq=FALSE, breaks=100)
```



```
unknown_p <- 1 - pbinom(50, 100, pis)

hist(unknown_p, freq = FALSE, breaks = 100,
     main = "Distribution of P(Votes > 50) with Uncertain p",
     xlab = "P(Votes > 50)", col = "lightblue")

abline(v = 0.8172718, col = "red", lwd = 2, lty = 2)  # Known P(Votes > 50)
```


Distribution of P(Votes > 50) with Uncertain p
Density
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.0
P(Votes > 50)

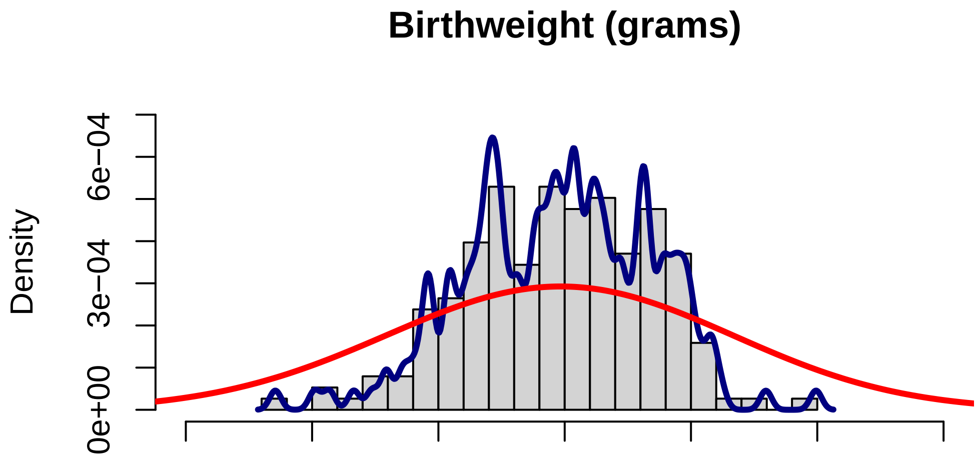
explore the trade-offs of model complexity and generalizability

The trade-off between model complexity and generalizability is critical in model building. Striking the right balance depends on:

Data Size: Complex models require larger datasets to avoid overfitting. Domain Knowledge: Simpler models may suffice if relationships are well-understood. Purpose: If interpretability is crucial, prioritize simpler models. The best model is not always the most complex but the one that captures the essence of the data while maintaining the ability to generalize to unseen scenarios.

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.

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



The goal is to estimate the probability density function (PDF) of infant birthweight using kernel density estimation (KDE). In the given code:

The blue line (adjust = 1/5) represents a KDE with low smoothing (potentially overfitting).

The red line (adjust = 5) represents a KDE with high smoothing (potentially underfitting).

Concepts of Overfitting and Underfitting

1. Overfitting

Definition: Overfitting occurs when a model captures noise or random fluctuations in the data rather than the underlying pattern.

In this context:

The blue line (adjust = 1/5) is an overfit estimate.

The low smoothing factor causes the KDE to closely follow every fluctuation in the data, creating a highly irregular curve.

This results in spiky behavior that reflects random variations in the sample rather than the true birthweight distribution.

Indicators of Overfitting:

High variance: The estimate changes significantly with small changes in the data.

Poor generalization: The model performs poorly on new data.

2. Underfitting

Definition: Underfitting occurs when a model is too simple to capture the underlying structure of the data.

In this context:

The red line (adjust = 5) is an underfit estimate.

The high smoothing factor causes the KDE to oversimplify the distribution, resulting in a flat and overly smoothed curve.

Important features of the data (e.g., multiple modes or skewness) are ignored.

Indicators of Underfitting:

High bias: The estimate fails to capture key patterns in the data.

Misrepresentation: The estimate may fail to reflect critical aspects of the true distribution, such as peaks or asymmetry.

Interpreting the Birthweight Data

True Birthweight Distribution:

Infant birthweights typically follow a skewed distribution with a peak around the median weight and a tail for lower weights (e.g., preterm births).

Overfit Estimate (Blue Line):

The blue line shows many sharp peaks and valleys that align too closely with random sampling variation in the dataset. This results in an unrealistic representation of the true population distribution.

Underfit Estimate (Red Line):

The red line is overly smoothed, losing critical features of the distribution, such as the peak and skewness.

This results in a flat, non-representative estimate of the birthweight data.

Balancing Overfitting and Underfitting

Goal: Find a balance where the KDE captures the main patterns of the data while ignoring noise.

Adjusting Smoothing:

The adjust parameter in density() controls the bandwidth (smoothing level) of the KDE:

Small values (e.g., adjust = 1/5) reduce smoothing, increasing variance (overfitting).

Large values (e.g., adjust = 5) increase smoothing, increasing bias (underfitting).

An optimal value lies somewhere in between, reflecting the true birthweight distribution.

Conclusion

The blue line (overfit) captures too much noise, leading to a highly irregular and unrealistic PDF.

The red line (underfit) oversimplifies the data, losing important features of the true distribution.

The trade-off between bias (underfitting) and variance (overfitting) is central to selecting an appropriate level of smoothing for KDE.

An intermediate smoothing factor (e.g., adjust = 1 or adjust = 2) might better capture the birthweight distribution.

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

Generalizability in the context of the birthweight data ensures that the estimated PDF accurately represents the broader population, not just the sample data. Overfitting and underfitting both harm generalizability, but by balancing model complexity and smoothing, we can construct a reliable PDF that captures the true structure of the birthweight distribution. This enables robust insights and reliable predictions for new data.

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.)

```r
require(magrittr)
```

```
Loading required package: magrittr
```

```r
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
```

```
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])
}
```

Solution:

```
set.seed(22)

simulations <- replicate(10000, game(4))

count <- 0
for (i in 1:ncol(simulations)) {

  if (simulations[1, i] == "goat" && simulations[2, i] == "goat") {
    count <- count + 1
  }
}

probability <- count / 10000
probability
```

```
[1] 0.3762
```