Homework 5

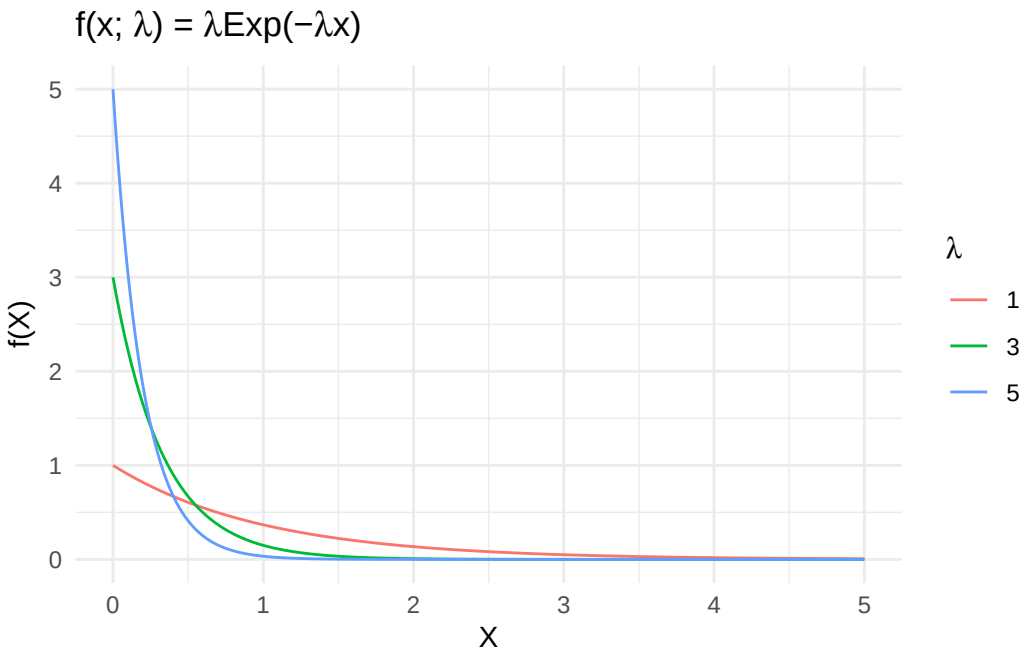
Ashutosh Ekade

Question 1: MLE for Exponential Distribution

1.a) Suppose data X has an exponential distribution with rate λ. The density of X is given below. Plot $f(X;\lambda)$ letting X range from 0 to 5 for $\lambda = 1, 3, 5$. (This requires you to plot three curves on the same graph or make 3 separate plots.)

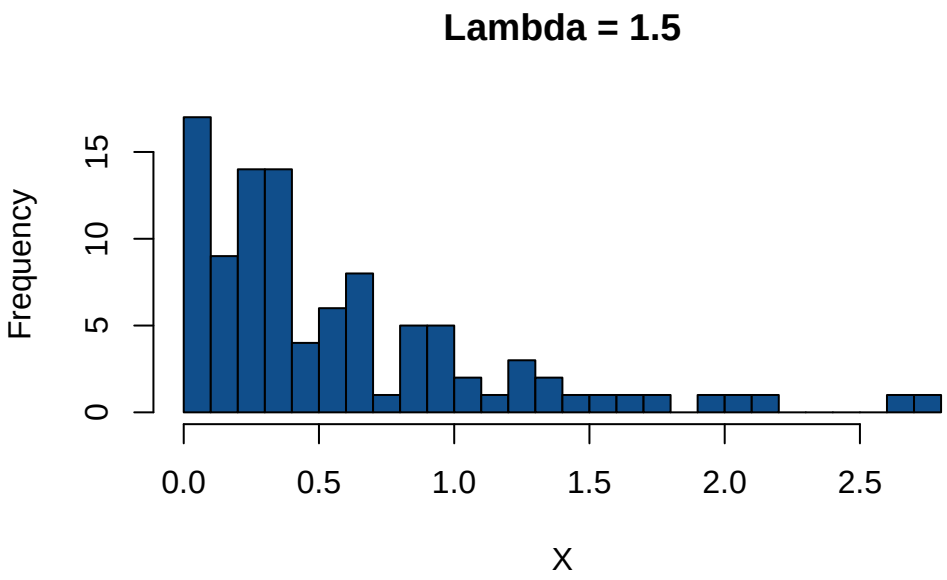
$$f(x;\lambda) = \lambda e^{-\lambda x}$$

```
library(ggplot2)
# Define the function with parameter lambda
f <- function(x, lambda) lambda * exp(-lambda * x)
# Define lambda values
lambd_vals <- c(1, 3, 5)
# Create a data frame for plotting
x_values <- seq(0, 5, length.out = 1000)
data <- expand.grid(x = x_values, lambda = lambd_vals)
data$y <- with(data, f(x, lambda))
# Plot using ggplot
ggplot(data, aes(x = x, y = y, color = as.factor(lambda))) +
geom_line() +
labs(x = "X", y = expression(paste("f(X)", sep = "")),
title = expression(paste("f(x; ", lambda, ") = ", lambda, "Exp(-", lambda, "x)", sep = "")
scale_color_discrete(name = expression(lambda)) +
theme_minimal()
```



1.b) Suppose the true value of λ is 1.5. Set the seed to 3344 and generate 100 values from an exponential distribution with rate 1.5. Plot the histogram of the data.

```r
set.seed(3344)
true_lambd <- 1.5
x_vals <- seq(0, 5, length.out = 1000)
# Compute y-values based on the function
y_vals <- rexp(100,true_lambd)
# Plot a histogram of y-values
hist(y_vals, breaks = 25, col = "dodgerblue4", border = "black",
xlab = "X", ylab = "Frequency",
main = paste("Lambda = 1.5"))
```



1.c) Create an `R` function to calculate the log-likelihood. The function should have two arguments: `lamba` and the vector `x`. Use the data generated in 1b above, find the MLE $\hat{\lambda}_{\text{MLE}}$ (either analytically, or via `optimize()` which is the one-parameter version of `optim()`, or via `optim()` with `method="BFGS"`). (Hint: The week 7 slides may be helpful.)

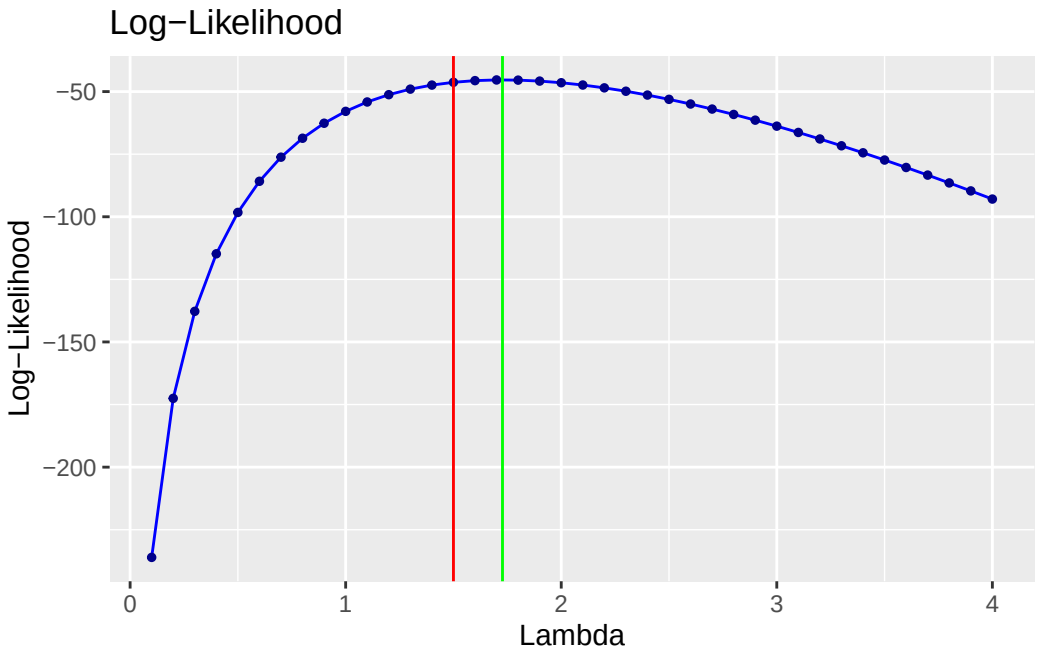
...add answer here...

```
log_likelihood <- function(lambda, data) {
n <- length(data)
n*log(lambda) - n*lambda*mean(data)
}
results <- optim(par = 1, fn = log_likelihood, data = y_vals, hessian = TRUE, method = "BF
mle_lambda <- results$par
cat("Maximum Likelihood Estimate for lambda:", mle_lambda)
```

```
Maximum Likelihood Estimate for lambda: 1.727454
```

1.d) Use the data generated in 1b above to plot the log-likelihood function for values of λ ranging from 0.1 to 4. Add a red vertical line at the true value of λ; add a green vertical line at the value $\hat{\lambda}_{\text{MLE}}$.

```r
lambd_vals <- seq(from = 0.1, to = 4, by = 0.1)
log_ll_vals <- rep(NA, 40)
for(i in 1:length(lambd_vals)) {
log_ll_vals[i] = log_likelihood(lambd_vals[i], y_vals)
}
# Create a dataframe
data <- data.frame(lambda = lambd_vals, log_likelihood = log_ll_vals)
# Plot log-likelihood against lambda using ggplot
ggplot(data, aes(x = lambda, y = log_likelihood)) +
  geom_line(color = "blue") +
  geom_point(color = "darkblue", size = 1) + # Add points (you can adjust size and color)
  labs(x = "Lambda", y = "Log-Likelihood",
  title = "Log-Likelihood") +
  geom_vline(xintercept = true_lambd, color = "red", linetype = "solid") +
  geom_vline(xintercept = mle_lambda, color = "green", linetype = "solid")
```



1.e) Use the asymptotical normal approximation for the MLE to calculate a 95% confidence interval for λ. This requires you to first calculate the standard error of the MLE. (Hint: The week 8 slides may be helpful, but note the difference in the definition of the exponential density function used in the slides. Alternatively, you may use the numerical approximation to the hessian returned by `optim()`.)

```r
results <- optim(par = 1, fn = log_likelihood, data = y_vals, hessian = TRUE, method = "BF
mle_lambd = results$par
std_err <- sqrt(diag(-1*solve(results$hessian)))
alpha <- 0.05 # Significance level
z_score <- qnorm(1 - alpha / 2)
lb <- mle_lambda - z_score * std_err
ub <- mle_lambda + z_score * std_err
cat("Maximum likelihood Estimate for lambda:", mle_lambd)
```

```
Maximum likelihood Estimate for lambda: 1.727454
```

```r
cat("Standard Error:", std_err)
```

```
Standard Error: 0.1727453
```

```r
cat("95% confidence intervals for MLE(lambda):", lb, ub, "\n")
```

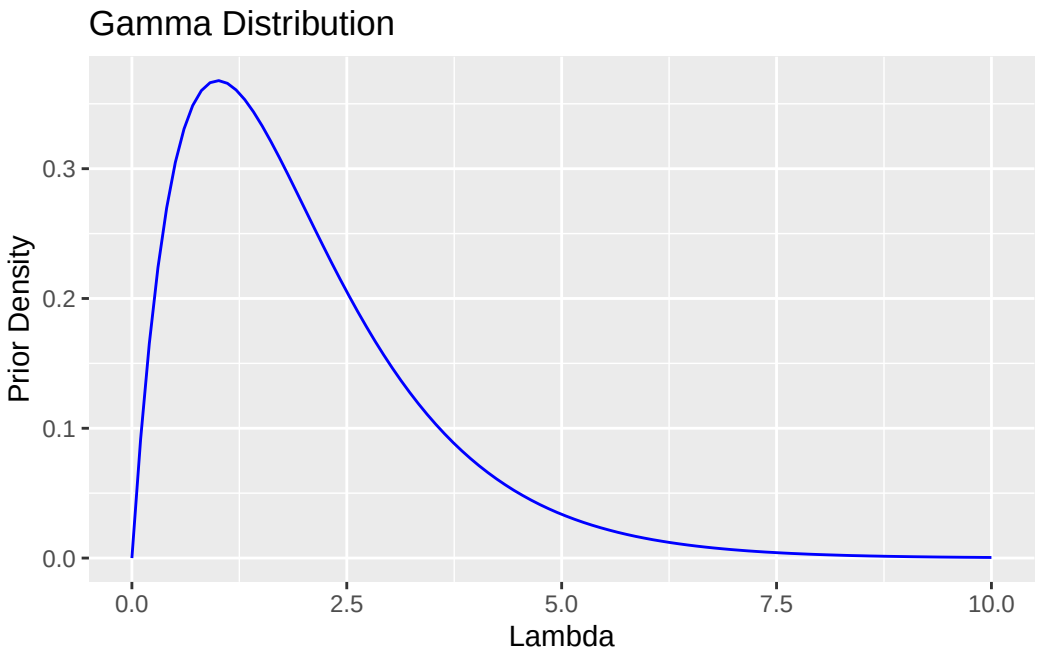
```
95% confidence intervals for MLE(lambda): 1.388879 2.066029
```

Question 2: Bayesian Inference for Exponential Distribution

2.a) Suppose the prior distribution for λ is a gamma distribution with shape parameter $\alpha = 2$ and rate parameter $\beta = 1$ given below. Plot the prior density for λ. (Hint: use the built-in `R` function `dgamma()` or code your own density using `gamma()` for $\Gamma(\cdot)$).

...add answer here...

```r
alpha <- 2
beta <- 1
lambda_values <- seq(0, 10, length.out = 100)
prior_density <- dgamma(lambda_values, shape = alpha, rate = beta)
prior_data <- data.frame(lambda = lambda_values, density = prior_density)
ggplot(prior_data, aes(x = lambda, y = density)) +
  geom_line(color = "blue") +
  labs(x = "Lambda", y = "Prior Density",
      title = "Gamma Distribution")
```



2.b) Analytically solve for the posterior distribution for λ given the data generated in 1b above. Note that Gamma distribution is conjugate to the Exponential likelihood function, so the posterior distribution is also a gamma distribution.

```
round(sum(y_vals),2)
```

```
[1] 57.89
```

- Prior distribution: Gamma($\alpha = 2$, $\beta = 1$) (shape $= \alpha$, rate $= \beta$)
- Likelihood function: Exponential($\lambda = 1.5$)

$$P(X = x|\lambda) = \lambda^n e^{-\lambda n\bar{x}}$$

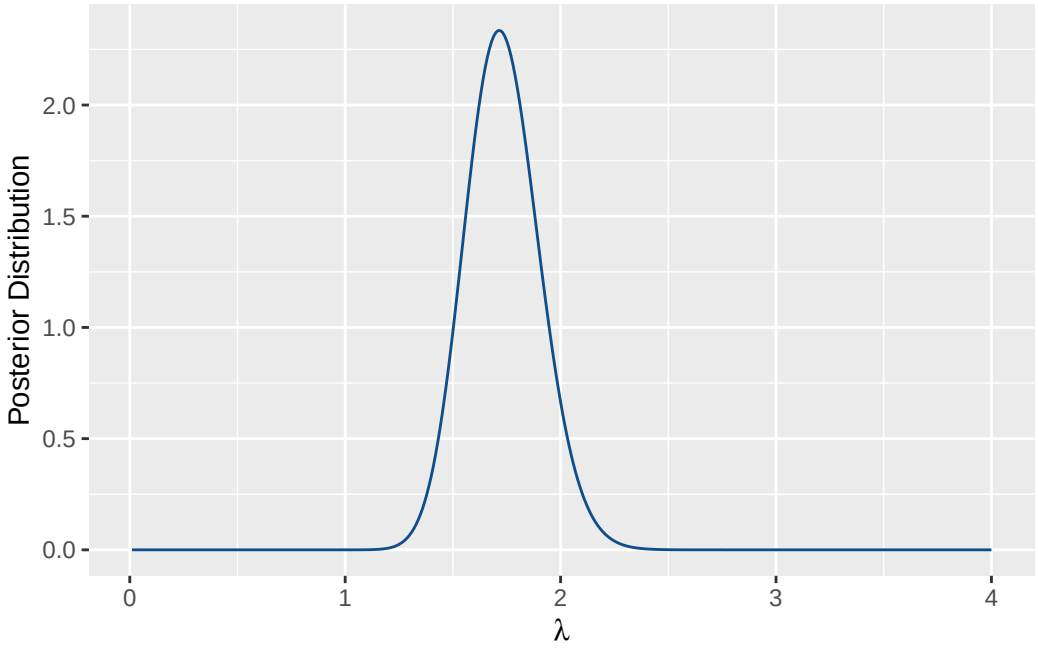
$$P(\lambda|\alpha,\beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

$$P(\lambda|x) = \frac{\lambda^n e^{-\lambda n\bar{x}} \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$
$$\propto \lambda^{n+\alpha-1} e^{-\lambda(n\bar{x}+\beta)}$$

$$\begin{aligned}\text{Posterior distribution} &= \text{Gamma}(\alpha' = \alpha + n, \beta' = \beta + \sum_{i=1}^{n} x_i) \text{ (shape} = \alpha', \text{ rate} = \beta') \\ &= \text{Gamma}(\alpha' = 2 + 100, \beta' = 1 + 57.89) \\ &= \text{Gamma}(\alpha' = 102, \beta' = 58.89)\end{aligned}$$

2.c) Plot the posterior distribution calculated in 2b above usind the `dgamma()` function. Let the horizontal axis for λ range from 0.01 to 4.

```
post_alpha = 102
post_beta = 58.89
lambdas <- seq(from = 0.01, to = 4, by = 0.01)
post_dist <- dgamma(lambdas, shape = post_alpha, rate = post_beta)
data <- data.frame(lambda = lambdas, posterior_dist = post_dist)
ggplot(data, aes(x = lambda, y = posterior_dist)) +
  geom_line(color = "dodgerblue4") +
  labs(x = expression(lambda), y = "Posterior Distribution")
```



2.d) Use the posterior distribution calculated in 2b above to calculate a 95% credible interval for λ. Use the built-in `R` function `qgamma()` and select the interval that leaves 2.5% of the posterior probability in each tail.

```
lq <- qgamma(0.025, shape = post_alpha, rate = post_beta)
uq <- qgamma(0.975, shape = post_alpha, rate = post_beta)
cat("95% Credible Interval for lambda:", lq, uq, "\n")
```

```
95% Credible Interval for lambda: 1.412274 2.083956
```

2.e) Suppose you were unable to derive the posterior distribution for λ analytically. "By hand," code up a Metropolis-Hastings algorithm to sample from the posterior distribution. Specifically, your code should do the following:

1) initialize λ to $\lambda_1 = 4$ (let's intentionally pick a somewhat bad starting value)

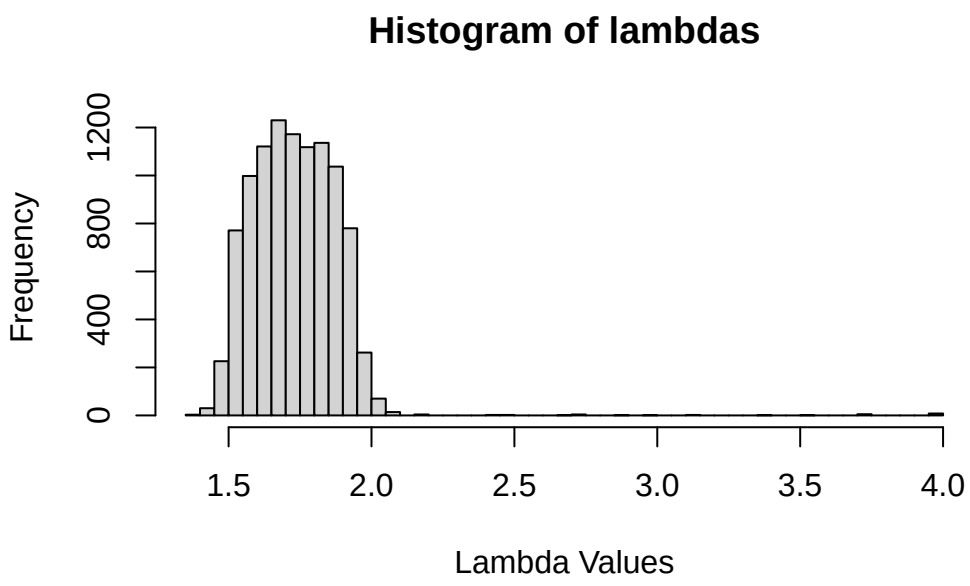
2) for $t = 2$ to $t = 10,000$, do the following

3) draw a candidate value for λ_{t+1} (we'll call it λ^*) from a Normal distribution centered at λ_t with standard deviation 0.2 (if λ^* is less than or equal to zero, which is highly unlikely but possible, draw again)

4) move to λ^* with probability $p = f(\lambda^*|x)/f(\lambda_t|x)$ so that $\lambda_{t+1} = \lambda^*$ or remain where you are with probability $1 - p$ so that $\lambda_{t+1} = \lambda_t$

When the algorithm is complete, you should have a vector of 10,000 samples from the posterior distribution for λ. Plot a histogram of those 10,000 values and confirm that the histogram approximates the posterior distribution plotted in 2.c above.

```
lambd <- 4
lambdas <- c(lambd)
for (t in 2:10000) {
  new_lambda <- -1
  while (new_lambda <= 0) {
    new_lambda <- rnorm(1, lambd, 0.2)
  }
  p <- dgamma(new_lambda, shape = post_alpha, rate = post_beta) / dgamma(lambd, shape = po
  if (p > 1 - p)
    lambd <- new_lambda
  lambdas <- c(lambdas, lambd)
}
hist(lambdas, breaks = 50, xlab = "Lambda Values", ylab = "Frequency")
```



2.f) Calculate the posterior mean for λ using the 10,000 samples generated in 2e above.

```
round(mean(lambdas),3)
```

```
[1] 1.732
```

2.g) Calculate a 95% credible interval for λ using the 10,000 samples generated in 2e above. (Hint: sort the samples from smallest to largest and select the 250th and 9,750th values.)

```
lambdas <- sort(lambdas)
credible_interval <- lambdas[c(250, 9750)]
cat("Credible Interval: (", credible_interval[1], ",", credible_interval[2],")\n")
```

```
Credible Interval: ( 1.498545 , 1.968852 )
```

Question 3: Causal Inference via Diff in Diff

In this problem, you will analyze data from the Card and Krueger paper discussed in lecture.

> Card, D. and A. B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,'' *American Economic Review*, vol. 84, 772-793. [(link)[https://davidcard.berkeley.edu/papers/njmin-aer.pdf]]

Card and Krueger are interested in estimating the impact of minimum wage on teenage employment. Conventional economic wisdom states that raises in minimum wages hurt employment, especially teenage employment, which often takes wages that will be affected by minimum wage law. However, empirical analysis has failed to find evidence of employment responses to raises in minimum wages. In 1992, New Jersey's minimum wage increased from $4.25 to $5.05 while the minimum wage in Pennsylvania remained at $4.25. The authors used data on employment at fast-food establishments in New Jersey and Pennsylvania before and after the increase in the minimum wage to measure the impact of the increase in minimum wage on teenage employment. Download and import the dataset, `card_krueger.RData`, from the course website.

```
load("card_krueger.Rdata")
```

The dataset contains the following variables, with one row corresponding to a single fast-food restaurant:

- `state`: New Jersey or Pennsylvania
- `chain`: the fast-food chain to which the restaurant belongs
- `wage_pre`: starting wage in February 1992, in dollars per hour
- `wage_post`: starting wage in November 1992, in dollars per hour
- `emp_pre`: employment in February 1992, in number of full-time equivalent employees
- `emp_post`: employment in November 1992, in number of full-time equivalent employees
- `closed`: whether the store was closed in November 1992

Assume that the fast-food restaurants surveyed by Card and Krueger represent a random sample from a larger population of all fast-food restaurants in New Jersey and eastern Pennsylvania. Consider the estimands in the table below, which correspond to the mean level of full-time equivalent (FTE) employment for population subgroups (restaurants within a given state-time). For example, $\alpha = \mathbb{E}[Y_i|T = \textsf{February}, \textsf{state} = \textsf{NJ}]$.

	Feb 1992	Nov 1992
New Jersey	α	β
Pennsylvania	γ	δ

3.a) Define the difference-in-differences (DID) estimand in terms of these values, and concisely explain what this represents.

discussed and attempted this section with my study group

```
nj_pre <- mean(ckdata$emp_pre[ckdata$state == "nj"], na.rm = T)
nj_post <- mean(ckdata$emp_post[ckdata$state == "nj"], na.rm = T)
pa_pre <- mean(ckdata$emp_pre[ckdata$state == "pa"], na.rm = T)
pa_post <- mean(ckdata$emp_post[ckdata$state == "pa"], na.rm = T)
DID_est <- (nj_post - pa_post) - (nj_pre - pa_pre)
cat("Difference-in-Differences (DID) Estimand:", DID_est, "\n")
```

```
Difference-in-Differences (DID) Estimand: 2.753606
```

- DID estimand: Compares employment change in treatment (New Jersey) vs. control (Pennsylvania) during the same period.
- Measures differential impact of a policy change on employment.

3.b). Calculate the DID estimate and a 95% confidence interval for it. Interpret your results. Note that you may need to reshape the data in order to call `lm()`. (Hint: You can check your estimate against row 3, column (iii) of Table 3 in the Card and Krueger paper. I was able to replicate their point estimate within a rounding error, but I found a slightly larger standard error than they reported.)

```
suppressPackageStartupMessages(library(dplyr))
library(tidyr)
data <- ckdata %>%
  gather(key = "kp", value = "f", emp_pre, emp_post) %>%
  mutate(t = ifelse(kp == "emp_pre", "pre", "post"))
data$r <- ifelse(data$state == "nj", 1, 0)
data$t_ind <- ifelse(data$t == "post", 1, 0)
model <- lm(f ~ r * t_ind, data = data)
did_estimate <- summary(model)$coef[4]
cat("The DID estimate is: ", did_estimate)
```

```
The DID estimate is:  2.753606
```

```r
ci <- confint(model, "r:t_ind", level = 0.95)
cat("\nConfidence Interval (95%) for DID estimate: ", ci)
```

```
Confidence Interval (95%) for DID estimate:  -0.560693 6.067905
```

- Regression's DID estimate matches calculated differences.
- Wide confidence interval includes zero at 95% confidence.
- Lack of significance suggests no impact on mean employment.