12th_april_design_inference_winter_project

Taiwo Ogunkeye

2025-04-12

```r
# loading libraries
library(mvtnorm)
```

```
## Warning: package 'mvtnorm' was built under R version 4.4.3
```

```r
library(ggplot2)
library(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
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Load data with proper column names and whitespace handling
df <- read.table(
  file = "interexp.dat",
  header = TRUE,
  na.strings = c("NA", ""),
  col.names = c("yA", "yB"),  # Explicitly set column names
  strip.white = TRUE,         # Remove extra whitespace
  sep = ""                    # Any whitespace separator
)

# Check column names
cat("Column names:", colnames(df), "\n")
```

```
## Column names: yA yB
```

```r
# Check first 6 rows
cat("\nFirst 6 rows:\n")
```

```
##
## First 6 rows:
```

```r
print(head(df))
```

```
##      yA    yB
## 1 25.33 26.45
## 2 26.77 27.53
## 3 22.76 20.02
## 4 20.94 22.83
## 5 25.40 28.05
## 6 22.49 23.67
```

```r
# Check missing values
cat("\nMissing values per column:\n")
```

```
##
## Missing values per column:
```

```r
print(colSums(is.na(df)))
```

```
## yA yB
## 17 15
```

```r
# EM Imputation Function that uses bivariate normal data
em_imputation <- function(data, max_iter = 500, tol = 1e-8) {

    imputed <- data.frame(
        yA = ifelse(is.na(data$yA),
                    median(data$yA, na.rm = TRUE),  # More robust initialization
                    data$yA),
        yB = ifelse(is.na(data$yB),
                    median(data$yB, na.rm = TRUE),
                    data$yB)
    )

    # Convert to matrix for numerical stability
    imputed_mat <- as.matrix(imputed)
    n <- nrow(imputed_mat)
    mu <- colMeans(imputed_mat)
    cov_mat <- cov(imputed_mat)

    # create initial covariance regularization
    cov_mat <- cov_mat + diag(1e-5, ncol(cov_mat))

    for (i in 1:max_iter) {
        sum_z <- matrix(0, 2, 1)
        sum_zz <- matrix(0, 2, 2)

        for (j in 1:n) {
            obs <- !is.na(data[j, ])
            miss <- is.na(data[j, ])

            if(any(miss)) {

                cov_sub <- cov_mat[obs, obs, drop = FALSE]
                cov_sub_reg <- cov_sub + diag(1e-6, nrow(cov_sub))
                cov_inv <- ginv(cov_sub_reg)

                #conditional mean calculation
                mu_cond <- mu[miss] +
                    (cov_mat[miss, obs, drop = FALSE] %*% cov_inv) %*%
                    (imputed_mat[j, obs] - mu[obs])

                # Updating the imputation
                imputed_mat[j, miss] <- mu_cond
            }

            # collect the statistics by using the updated imputation
            z <- matrix(imputed_mat[j, ], ncol = 1)
            sum_z <- sum_z + z
            sum_zz <- sum_zz + tcrossprod(z)
        }

        # updating the parameters mu and covariance
        new_mu <- sum_z/n
        new_cov <- (sum_zz/n) - tcrossprod(new_mu)
        cov_mat <- new_cov + diag(1e-5, 2)

        # Checking the convergence condition
        if (i > 1 && norm(mu - new_mu, "F") < tol) break
        mu <- new_mu
    }

    list(
        imputed_data = data.frame(imputed_mat),
        mu = mu,
        sigma = cov_mat,
        iterations = i
    )
}


# running the imputation above
set.seed(42)  # For reproducibility
result <- em_imputation(df)
df_imputed <- result$imputed_data

# creating file called imputed.csv
```

```r
write.csv(df_imputed, "imputed_data.csv", row.names = FALSE)

# printing the pre-imputation statistics
cat("Pre-imputation summary:\n")
```

```
## Pre-imputation summary:
```

```r
print(summary(df))
```

```
##        yA              yB       
##  Min.   :20.40   Min.   :20.02  
##  1st Qu.:22.41   1st Qu.:23.16  
##  Median :24.34   Median :24.99  
##  Mean   :24.20   Mean   :24.81  
##  3rd Qu.:25.52   3rd Qu.:26.62  
##  Max.   :29.09   Max.   :28.05  
##  NA’s   :17      NA’s   :15     
```

```r
#printing the post imputation statistics
cat("\nPost-imputation summary:\n")
```

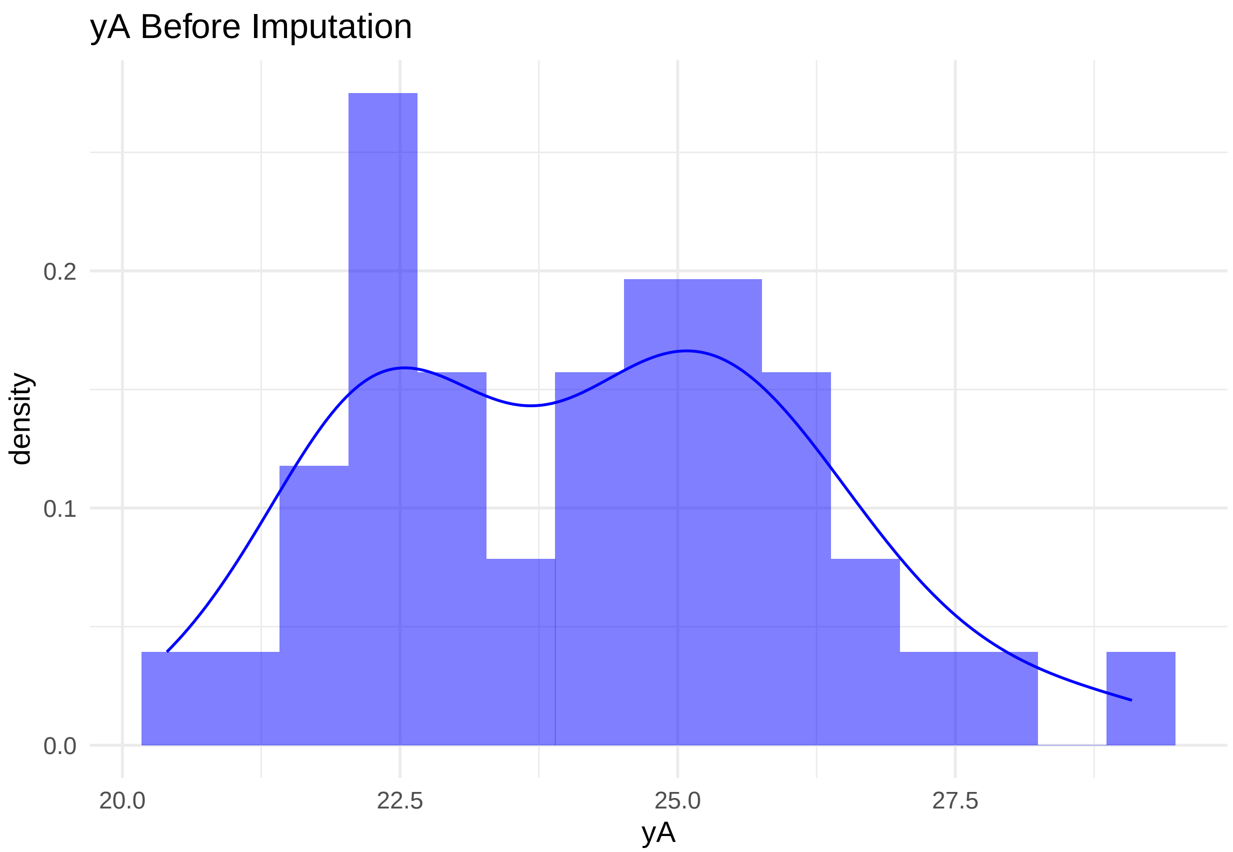
```
## 
## Post-imputation summary:
```

```r
print(summary(df_imputed))
```

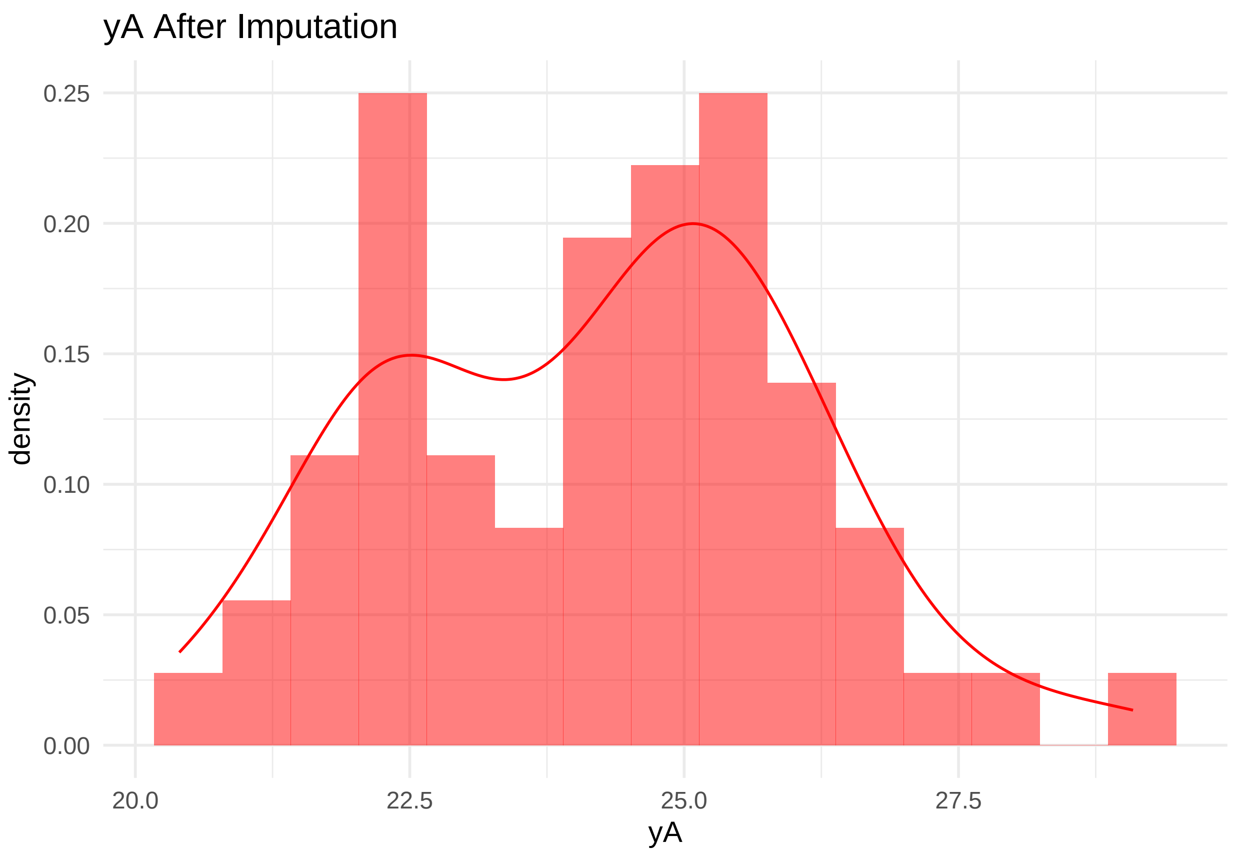
```
##        yA              yB       
##  Min.   :20.40   Min.   :20.02  
##  1st Qu.:22.52   1st Qu.:23.50  
##  Median :24.36   Median :24.98  
##  Mean   :24.21   Mean   :24.83  
##  3rd Qu.:25.38   3rd Qu.:26.26  
##  Max.   :29.09   Max.   :28.05  
```

```r
# Plotting the distributions in individual density plots
plot_distribution_single <- function(var, dataset, title_text, fill_color) {
    ggplot() +
        geom_histogram(
            data = dataset, aes(x = !!sym(var), y = ..density..),
            bins = 15, fill = fill_color, alpha = 0.5
        ) +
        geom_density(data = dataset, aes(x = !!sym(var)), color = fill_color) +
        ggtitle(title_text) +
        theme_minimal()
}

# Plotting the distributions in the same plot to check the visual changes
plot_distribution_combined <- function(var) {
    ggplot() +
        geom_histogram(
            data = df, aes(x = !!sym(var), y = ..density..),
            bins = 15, fill = "blue", alpha = 0.3
        ) +
        geom_histogram(
            data = df_imputed, aes(x = !!sym(var), y = ..density..),
            bins = 15, fill = "red", alpha = 0.3
        ) +
        geom_density(data = df, aes(x = !!sym(var)), color = "blue") +
        geom_density(data = df_imputed, aes(x = !!sym(var)), color = "red") +
        ggtitle(paste("Distribution of", var, "(Combined)")) +
        theme_minimal()
}


plot_distribution_single("yA", df, "yA Before Imputation", "blue")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Removed 17 rows containing non-finite outside the scale range
## (‘stat_bin()‘).

## Warning: Removed 17 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```

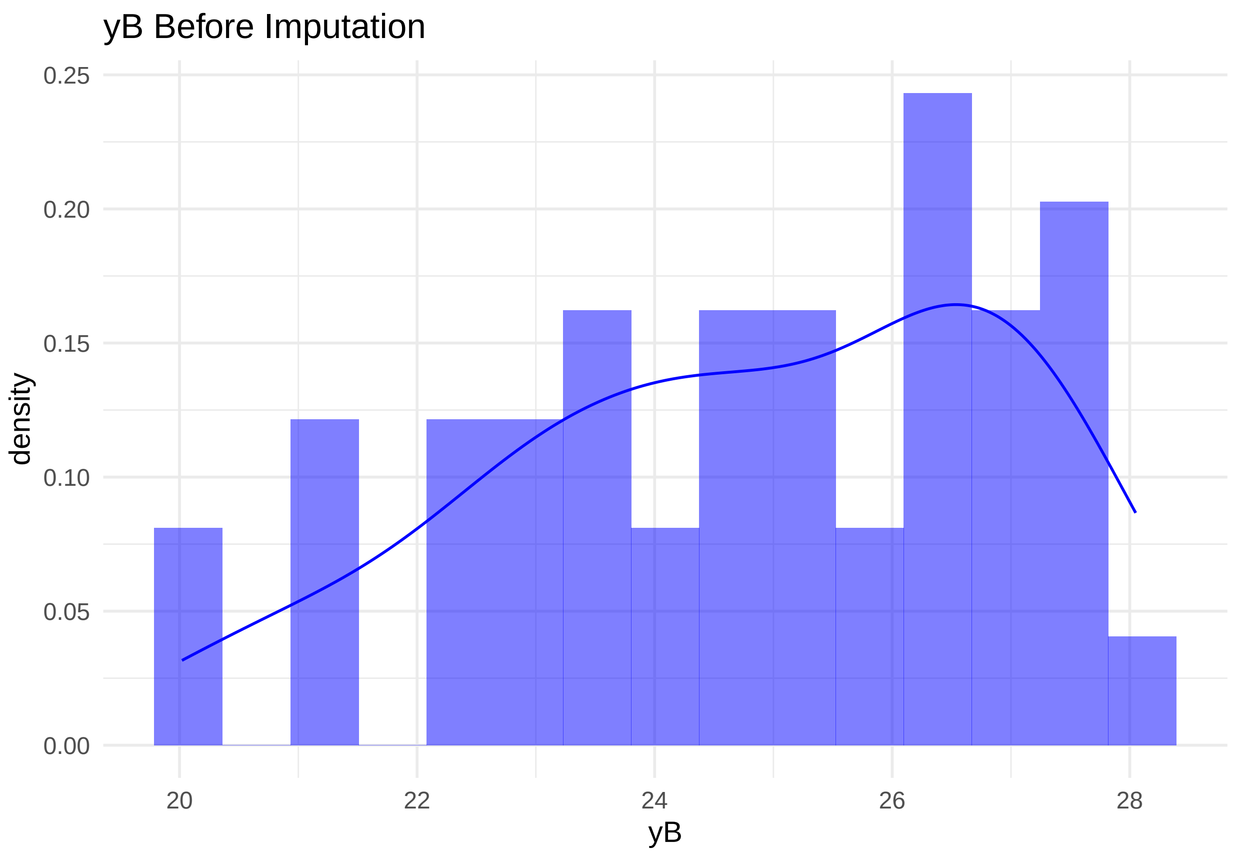


```
plot_distribution_single("yA", df_imputed, "yA After Imputation", "red")
```

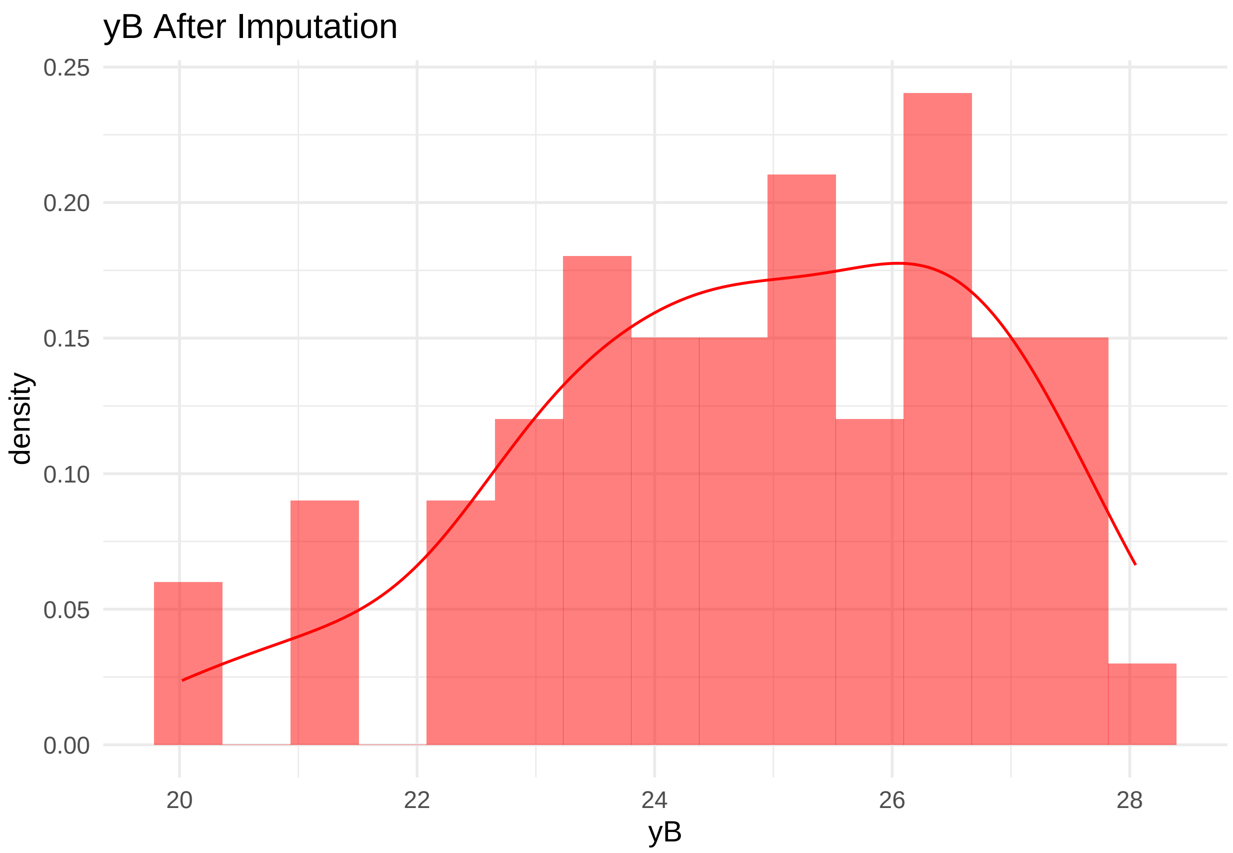


```
plot_distribution_single("yB", df, "yB Before Imputation", "blue")
```

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_bin()‘).

## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```

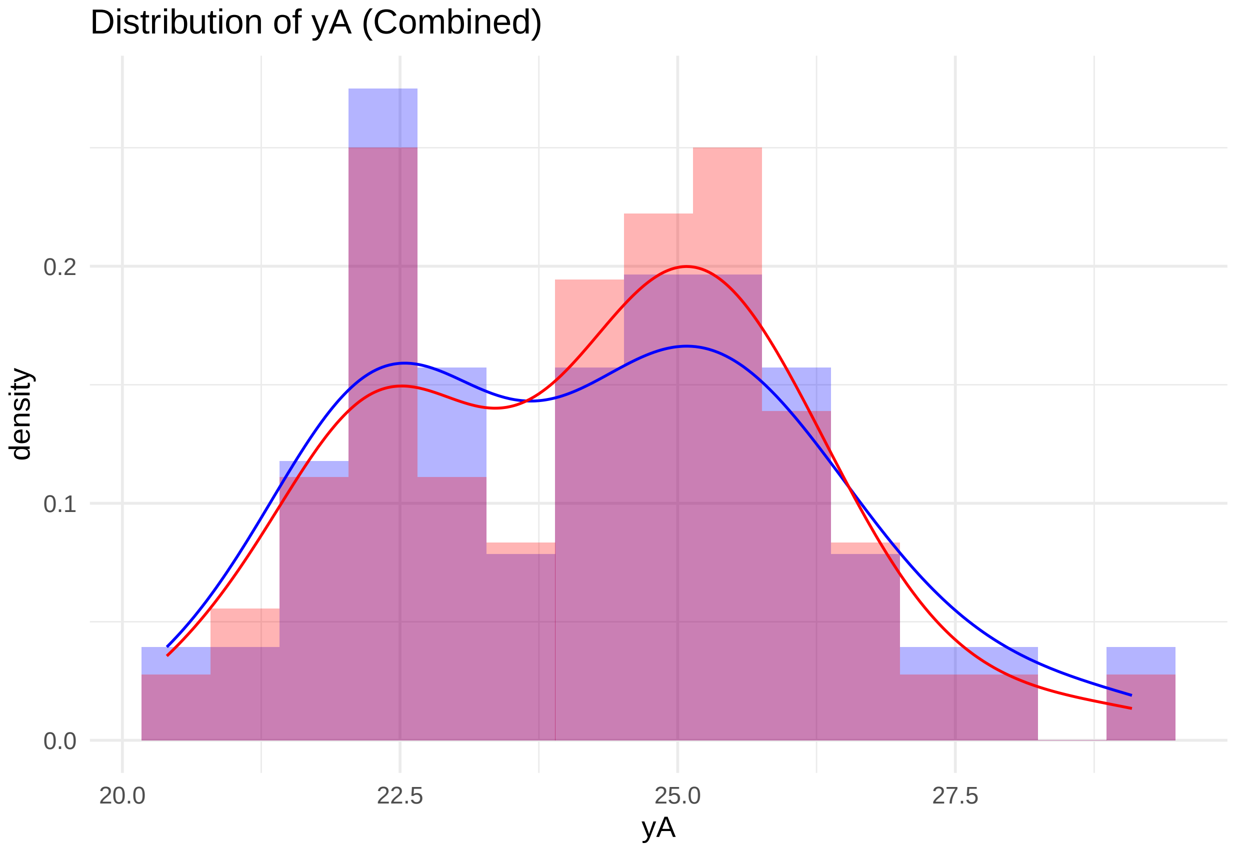


```
plot_distribution_single("yB", df_imputed, "yB After Imputation", "red")
```



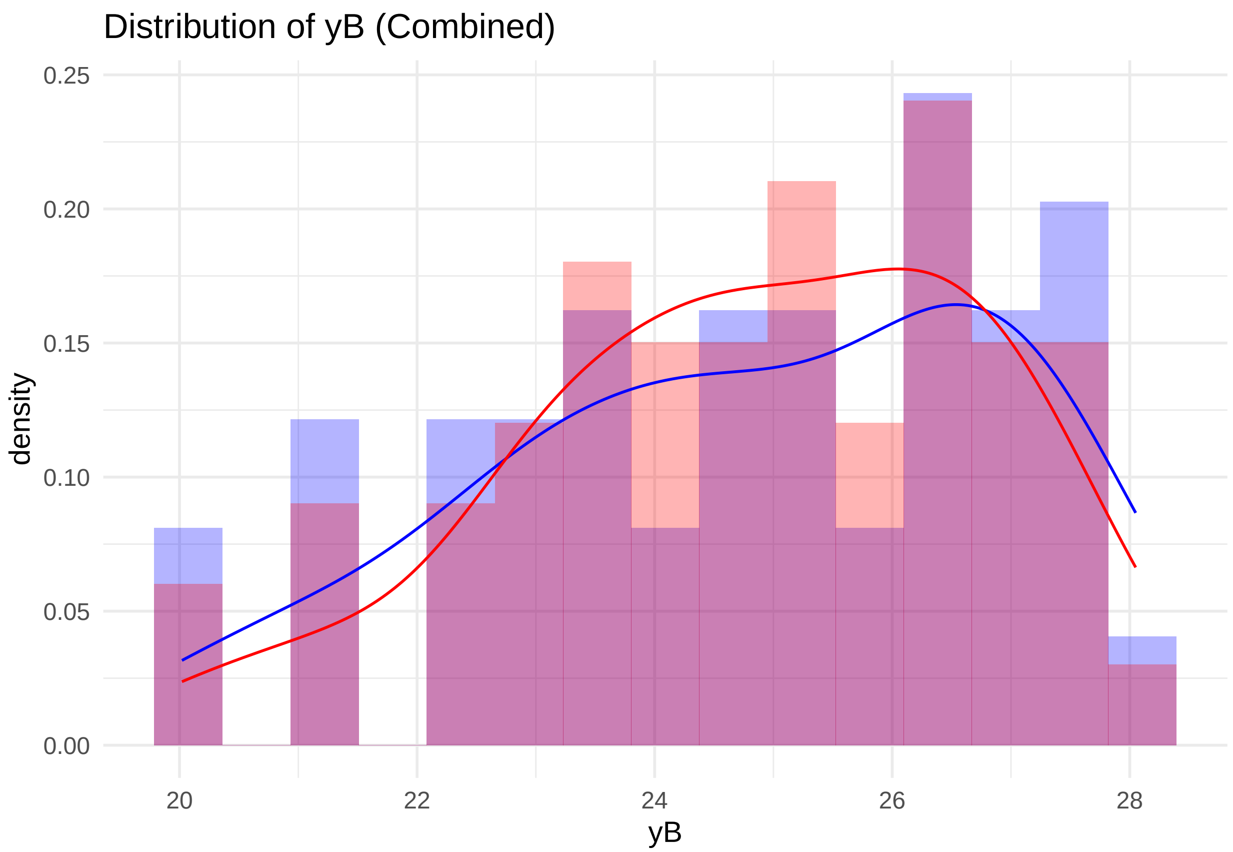
```
plot_distribution_combined("yA")
```

```
## Warning: Removed 17 rows containing non-finite outside the scale range
## (‘stat_bin()‘).

## Warning: Removed 17 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



```
plot_distribution_combined("yB")
```

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_bin()‘).

## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_density()‘).
```



```r
# Covariance comparison
cat("\nOriginal covariance (complete cases):\n")
```

```
##
## Original covariance (complete cases):
```

```r
print(cov(df, use = "complete.obs"))
```

```
##          yA       yB
## yA 5.321267 3.136247
## yB 3.136247 4.864169
```

```r
cat("\nImputed covariance:\n")
```

```
##
## Imputed covariance:
```

```r
print(cov(df_imputed))
```

```
##          yA       yB
## yA 3.576391 2.755524
## yB 2.755524 3.775615
```

```
# Statistical test
cat("\nPaired t-test results:\n")
```

```
##
## Paired t-test results:
```

```
print(t.test(df_imputed$yA, df_imputed$yB, paired = TRUE))
```

```
##
##  Paired t-test
##
## data:  df_imputed$yA and df_imputed$yB
## t = -3.4662, df = 57, p-value = 0.001011
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.9742845 -0.2607698
## sample estimates:
## mean difference
##      -0.6175271
```

“‘