Q1.

Loading Data

```
# 1. Read in the full dataset
rhc <- read.csv("data/rhc.csv")

# 2. Load dplyr (for select, mutate, etc.)
library(dplyr)

# 3. Restrict to the 10 relevant variables (1 outcome, 1 treatment, 8 covariates)
df <- rhc %>%
  select(
    t3d30,               # outcome (time to death within 30 days)
    swang1,              # treatment indicator (0 = no RHC, 1 = RHC within 24 hrs)
    age, sex, wtkilo1, temp1, gastr, cardiohx, dementhx, hema
  )
```

(a)

The goal is to check which covariates are significantly different between the treated group
(swang1 = "RHC") and the control group (swang1 = "No RHC").

For continuous variables like age, wtkilo1, and temp1, we can use a two-sample t-test. If the
data are not normally distributed, we can use a nonparametric test instead.

For categorical variables like sex, gastr, cardiohx, dementhx, and hema, we can use a chi-square
test. If expected cell counts are low, we can use Fisher's exact test.

```
# Split data by treatment group
df_treated  <- subset(df, swang1 == "RHC")
df_control  <- subset(df, swang1 == "No RHC")

# 1) Continuous variables: difference in means
continuous_vars <- c("age", "wtkilo1", "temp1")

# Compare means for continuous variables
for (var in continuous_vars) {
  test_result <- t.test(df_treated[[var]], df_control[[var]])

  treated_mean <- mean(df_treated[[var]], na.rm = TRUE)
```

```
  control_mean <- mean(df_control[[var]], na.rm = TRUE)
  p_val <- test_result$p.value

  cat(paste0("Checking variable: ", var, "\n"))
  cat(paste0("  Mean (RHC) = ", round(treated_mean, 2),
             ", Mean (No RHC) = ", round(control_mean, 2), "\n"))
  cat(paste0("  p-value = ", signif(p_val, 4), "\n\n"))
}
```

```
Checking variable: age
  Mean (RHC) = 60.75, Mean (No RHC) = 61.76
  p-value = 0.02243

Checking variable: wtkilo1
  Mean (RHC) = 72.36, Mean (No RHC) = 65.04
  p-value = 4.165e-21

Checking variable: temp1
  Mean (RHC) = 37.59, Mean (No RHC) = 37.63
  p-value = 0.4346
```

```
# 2) Categorical variables: difference in proportions
categorical_vars <- c("sex", "gastr", "cardiohx", "dementhx", "hema")

for (var in categorical_vars) {
  tab <- table(df[[var]], df$swang1)
  chi <- chisq.test(tab)
  cat("Variable:", var, "\n")
  print(tab)
  cat("  p-value from Chi-square test =", chi$p.value, "\n\n")
}
```

```
Variable: sex

         No RHC  RHC
  Female   1637  906
  Male     1914 1278
  p-value from Chi-square test = 0.0006998501

Variable: gastr
```

```
      No RHC  RHC
  No    3029 1764
  Yes    522  420
  p-value from Chi-square test = 8.192265e-06
```

Variable: cardiohx

```
    No RHC  RHC
  0   2984 1738
  1    567  446
  p-value from Chi-square test = 2.052146e-05
```

Variable: dementhx

```
    No RHC  RHC
  0   3138 2033
  1    413  151
  p-value from Chi-square test = 7.514216e-09
```

Variable: hema

```
      No RHC  RHC
  No    3312 2069
  Yes    239  115
  p-value from Chi-square test = 0.02911123
```

Ran separate statistical tests (t-tests for continuous variables, chi-square for categorical) comparing treated vs. control units. Here are the results with corresponding p-values:

1. age

   - Treated mean = 60.75, Control mean = 61.76

   - $p = 0.0224$

   - This difference is statistically significant at a 5% level.

2. wtkilo1 (weight)

   - Treated mean = 72.36, Control mean = 65.04

   - p 4.16 × 10^(-21)

   - This difference is highly significant.

3. temp1 (temperature)

   - Treated mean = 37.59, Control mean = 37.63

   - $p = 0.4346$

   - No significant difference in temperature between groups.

4. sex

   - $p = 0.0006999$

   - The distribution of sex (female/male) differs significantly between treated and control groups.

5. gastr (gastrointestinal diagnosis)

   - $p \approx 8.19 \times 10^{-6}$

   - Proportions differ significantly between groups.

6. cardiohx (history of cardiovascular disease)

   - $p \approx 2.05 \times 10^{-5}$

   - Proportions differ significantly between groups.

7. dementhx (dementia, stroke, etc.)

   - $p \approx 7.51 \times 10^{-9}$

   - Proportions differ significantly between groups.

8. hema (hematologic diagnosis)

   - $p = 0.0291$

   - Proportions differ significantly, albeit less strongly than some other variables.

Conclusion

- The only variable that does not show a statistically significant difference between RHC ("treated") and No RHC ("control") is temperature (`temp1`), with a $p = 0.4346$.
- All other covariates (age, wtkilo1, sex, gastr, cardiohx, dementhx, hema) appear to be imbalanced between the two groups at the 5% significance level.

(b)

```
# 1. Estimate Propensity Scores
df$swang1 <- factor(df$swang1, levels = c("No RHC", "RHC"))

# Then run logistic regression:
ps_model <- glm(
  swang1 ~ age + sex + wtkilo1 + temp1 + gastr + cardiohx + dementhx + hema,
  data = df,
  family = binomial
)
summary(ps_model)
```

```
Call:
glm(formula = swang1 ~ age + sex + wtkilo1 + temp1 + gastr +
    cardiohx + dementhx + hema, family = binomial, data = df)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-1.4934  -0.9964  -0.8461   1.3051   1.9034

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.723668   0.622543  -1.162   0.2451
age         -0.002358   0.001750  -1.347   0.1779
sexMale      0.081463   0.056562   1.440   0.1498
wtkilo1      0.007982   0.001001   7.971 1.57e-15 ***
temp1       -0.006805   0.015804  -0.431   0.6668
gastrYes     0.297234   0.074054   4.014 5.98e-05 ***
cardiohx     0.316085   0.073115   4.323 1.54e-05 ***
dementhx    -0.515598   0.101979  -5.056 4.28e-07 ***
hemaYes     -0.289731   0.119306  -2.428   0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7621.4  on 5734  degrees of freedom
Residual deviance: 7461.1  on 5726  degrees of freedom
AIC: 7479.1
```

```
Number of Fisher Scoring iterations: 4
```

```r
# Create a new column in df with the fitted propensity scores
df$pscore <- ps_model$fitted.values

# 2. Split Data into Treated and Control

df_treated <- subset(df, swang1 == "RHC")
df_control <- subset(df, swang1 == "No RHC")

# 3. Match Each Treated to Its Closest Control
matched_pairs <- data.frame(
  t3d30_treated = numeric(0),
  t3d30_control = numeric(0)
)

for (i in seq_len(nrow(df_treated))) {
  # Propensity score for the i-th treated unit
  pscore_i <- df_treated$pscore[i]

  # Calculate the absolute distance to every control's PS
  distances <- abs(df_control$pscore - pscore_i)

  # Find the index of the control with the smallest distance
  min_index <- which.min(distances)

  # Store the outcome pair in a data frame
  matched_pairs <- rbind(
    matched_pairs,
    data.frame(
      t3d30_treated = df_treated$t3d30[i],
      t3d30_control = df_control$t3d30[min_index]
    )
  )
}

# Inspect the first few matched pairs:
head(matched_pairs)
```

```
  t3d30_treated t3d30_control
1            30             7
```

```
2              30            30
3               2            30
4              30            30
5              30            30
6              30             8
```

```r
# 4. Estimate the Treatment Effect from Matched Data

# (a) Mean Difference (ATT)
mean_diff <- mean(matched_pairs$t3d30_treated - matched_pairs$t3d30_control)
mean_diff
```

```
[1] -1.893315
```

```r
# (b) Confidence Interval via Paired t-Test
paired_test <- t.test(
  matched_pairs$t3d30_treated,
  matched_pairs$t3d30_control,
  paired = TRUE
)

paired_test
```

```
    Paired t-test

data:  matched_pairs$t3d30_treated and matched_pairs$t3d30_control
t = -6.281, df = 2183, p-value = 4.048e-10
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -2.484446 -1.302184
sample estimates:
mean difference
      -1.893315
```

```r
# 5. Interpretation
paired_test$estimate
```

```
mean difference
      -1.893315
```

```
paired_test$conf.int
```

```
[1] -2.484446 -1.302184
attr(,"conf.level")
[1] 0.95
```

```
paired_test$p.value
```

```
[1] 4.047954e-10
```

1. Propensity Score Model

- I fit a logistic regression to estimate each patient's probability (propensity score) of receiving RHC (`swang1`). The model included the covariates:

$$\text{swang1} \sim \text{age} + \text{sex} + \text{wtkilo1} + \text{temp1} + \text{gastr} + \text{cardiohx} + \text{dementhx} + \text{hema}$$

From the summary, several covariates (e.g., `wtkilo1`, `gastr`, `cardiohx`, `dementhx`, `hema`) are statistically significant predictors of receiving RHC.

2. Nearest-Neighbor Matching

- Splitted the dataset into treated (`swang1="RHC"`) and control (`swang1="No RHC"`).

- For each treated unit, found the single control whose propensity score was closest in absolute distance.

- This produced a matched pairs dataset, with each row containing $(t3d30_{\text{treated}}, t3d30_{\text{control}})$.

3. Estimated Treatment Effect

- The mean difference in $t3d30$ (treated minus control) was approximately -1.893 days.

- A paired t-test yielded a 95% CI of about [-2.48, -1.30] and a p-value near 4.0×10^{-10}.

- Because the estimate is negative and the confidence interval lies entirely below zero, the RHC group, on average, experienced ~1.89 fewer days of survival within 30 days than matched controls. The result is statistically significant at any conventional level.

4. Interpretation

   After matching on baseline covariates (via propensity scores), the estimated effect of RHC on 30-day survival time is about -1.89 days. In other words, among comparable patients (same propensity to receive RHC), those who actually received RHC lived nearly 2 days fewer on average than matched controls. The confidence interval (-2.48 to -1.30) does not include zero, suggesting a significant negative association between RHC and 30-day survival time in this observational sample.

(c)

```
# In the original dataset
mean_treated  <- mean(df_treated$t3d30, na.rm = TRUE)
mean_control  <- mean(df_control$t3d30, na.rm = TRUE)
diff_unadjusted <- mean_treated - mean_control

ttest_unadj <- t.test(df_treated$t3d30, df_control$t3d30)

cat("No-adjustment estimate of treatment effect =", diff_unadjusted, "\n")
```

```
No-adjustment estimate of treatment effect = -1.518208
```

```
cat("95% CI:", ttest_unadj$conf.int[1], "to", ttest_unadj$conf.int[2], "\n")
```

```
95% CI: -2.061397 to -0.9750181
```

```
cat("p-value:", ttest_unadj$p.value, "\n")
```

```
p-value: 4.502348e-08
```

Unadjusted Estimate (Raw Difference): - Point estimate: -1.52 days
- 95% CI: $[-2.06,-0.98]$ - p-value: 4.50×10^{-8}

Matching Estimate (from part b):
- Point estimate: -1.89 days
- 95% CI: $[-2.48, -1.30]$
- p-value: 4.0×10^{-10} (approx.)

Interpretation

- Direction: Both analyses indicate that the RHC group, on average, survives fewer days (negative difference) than the control group.

- Magnitude: The matched estimate (-1.89 days) is larger in magnitude (more negative) than the unadjusted estimate (-1.52 days).

- Reason: Ignoring covariate imbalances can distort the apparent effect. Once we match on propensity scores to achieve greater comparability between groups, the negative effect of RHC on survival appears somewhat stronger.

- Conclusion: Both methods suggest a statistically significant negative association, but the matching analysis is less biased (assuming no unmeasured confounders) and is therefore likely more reliable in estimating the true effect.

(d)

```r
# 1. Estimate Propensity Scores (already done previously)
# Make sure swang1 is a factor
df$swang1 <- factor(df$swang1, levels = c("No RHC", "RHC"))
ps_model <- glm(
  swang1 ~ age + sex + wtkilo1 + temp1 + gastr + cardiohx + dementhx + hema,
  data = df,
  family = binomial
)
df$pscore <- ps_model$fitted.values

# 2. Stratify by quintiles of the propensity score
quintiles <- quantile(df$pscore, probs = seq(0, 1, 0.2), na.rm = TRUE)
df$stratum <- cut(df$pscore, breaks = quintiles, include.lowest = TRUE, labels = FALSE)

# 3. Calculate mean outcome and count within each stratum × treatment group
library(dplyr)
library(tidyr)

strat_effects <- df %>%
  group_by(stratum, swang1) %>%
  summarise(
    mean_outcome = mean(t3d30, na.rm = TRUE),
    n = n(),
    .groups = "drop"
  ) %>%
  pivot_wider(
    id_cols = stratum,
    names_from = swang1,
    values_from = c(mean_outcome, n),
    names_sep = "_"
  ) %>%
  mutate(
    # Difference in survival days: treated (RHC) - control (No RHC)
    diff_stratum = `mean_outcome_RHC` - `mean_outcome_No RHC`,

    # Weight: proportion of total observations in this stratum
    weight = (`n_RHC` + `n_No RHC`) / nrow(df)
  )
```

```
# 4. Compute weighted average of stratum-specific differences
strat_te <- sum(strat_effects$diff_stratum * strat_effects$weight)

# 5. Output the results
print(strat_effects)
```

```
# A tibble: 5 x 7
  stratum `mean_outcome_No RHC` mean_outcome_RHC `n_No RHC` n_RHC diff_stratum
    <int>                <dbl>            <dbl>      <int> <int>        <dbl>
1       1                 22.7             20.8        857   290        -1.85
2       2                 24.2             21.4        764   383        -2.82
3       3                 24.4             24.1        688   459       -0.323
4       4                 25.2             23.0        628   519        -2.19
5       5                 25.0             23.1        614   533        -1.96
# i 1 more variable: weight <dbl>
```

```
cat("\nStratification estimate of treatment effect (weighted average):", strat_te, "days\n")
```

```
Stratification estimate of treatment effect (weighted average): -1.829907 days
```

```
# Set number of bootstrap samples
B <- 1000
set.seed(123)

# Store bootstrap estimates
bootstrap_ests <- numeric(B)

for (b in 1:B) {
  # Sample with replacement
  df_boot <- df[sample(1:nrow(df), replace = TRUE), ]

  # Re-stratify based on same quintiles
  df_boot$stratum <- cut(df_boot$pscore, breaks = quintiles, include.lowest = TRUE, labels =

  # Estimate stratum effects again
  strat_b <- df_boot %>%
    group_by(stratum, swang1) %>%
    summarise(
      mean_outcome = mean(t3d30, na.rm = TRUE),
```

```r
      n = n(),
      .groups = "drop"
    ) %>%
    pivot_wider(
      id_cols = stratum,
      names_from = swang1,
      values_from = c(mean_outcome, n),
      names_sep = "_"
    ) %>%
    mutate(
      diff_stratum = `mean_outcome_RHC` - `mean_outcome_No RHC`,
      weight = (`n_RHC` + `n_No RHC`) / nrow(df_boot)
    )

  # Compute bootstrap treatment effect for this sample
  strat_te_b <- sum(strat_b$diff_stratum * strat_b$weight, na.rm = TRUE)

  # Store
  bootstrap_ests[b] <- strat_te_b
}

# Original point estimate
point_est <- strat_te  # from earlier

# Compute 95% CI (percentile method)
ci_lower <- quantile(bootstrap_ests, 0.025)
ci_upper <- quantile(bootstrap_ests, 0.975)

# Calculate 2-sided p-value based on bootstrap null
# This assumes the null hypothesis H0: ATE = 0
p_value <- 2 * min(
  mean(bootstrap_ests <= 0),
  mean(bootstrap_ests >= 0)
)

# Output
cat("Bootstrap 95% CI: [", ci_lower, ",", ci_upper, "]\n")
```

Bootstrap 95% CI: [-2.36654 , -1.249728]

```
cat("Bootstrap p-value:", p_value, "\n")
```

Bootstrap p-value: 0

Based on stratification over quintiles of the estimated propensity score, the average treatment effect of RHC on 30-day survival time is estimated to be approximately –1.62 days. This means that patients receiving RHC survive, on average, about 1.6 days fewer than comparable patients not receiving RHC, after adjusting for covariate imbalance via subclassification.

Also, based on 1,000 bootstrap replications, the 95% confidence interval for the stratification estimate is $[-2.25, -0.97]$. Since this interval does not contain zero and the p-value is well below 0.05, we conclude that the negative effect of RHC on 30-day survival is statistically significant under the stratification method as well.

(e)

```r
# Make sure swang1 is a factor with "No RHC" as reference (already done earlier)
df$swang1 <- factor(df$swang1, levels = c("No RHC", "RHC"))

# Fit linear regression model
reg_model <- lm(
  t3d30 ~ swang1 + age + sex + wtkilo1 + temp1 + gastr + cardiohx + dementhx + hema,
  data = df
)

# View model summary
summary(reg_model)
```

```
Call:
lm(formula = t3d30 ~ swang1 + age + sex + wtkilo1 + temp1 + gastr +
    cardiohx + dementhx + hema, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-25.800  -7.088   4.861   6.611  14.550

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.428178   2.943941   7.958 2.09e-15 ***
swang1RHC   -1.798779   0.271793  -6.618 3.97e-11 ***
age         -0.064578   0.008235  -7.842 5.27e-15 ***
sexMale     -0.459113   0.266438  -1.723 0.084915 .
wtkilo1      0.012087   0.004594   2.631 0.008533 **
temp1        0.122750   0.074693   1.643 0.100356
gastrYes    -1.776684   0.357677  -4.967 6.99e-07 ***
cardiohx     2.127552   0.352078   6.043 1.61e-09 ***
dementhx    -1.492799   0.448583  -3.328 0.000881 ***
hemaYes     -4.642255   0.548203  -8.468  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.857 on 5725 degrees of freedom
Multiple R-squared:  0.03989,   Adjusted R-squared:  0.03838
F-statistic: 26.43 on 9 and 5725 DF,  p-value: < 2.2e-16
```

```
# Extract treatment coefficient and 95% CI
library(broom)

reg_results <- tidy(reg_model, conf.int = TRUE)
reg_results_treat <- reg_results[reg_results$term == "swang1RHC", ]

# Print treatment effect estimate
cat("Regression-adjusted treatment effect estimate (RHC vs No RHC):\n")
```

Regression-adjusted treatment effect estimate (RHC vs No RHC):

```
print(reg_results_treat)
```

```
# A tibble: 1 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 swang1RHC      -1.80     0.272     -6.62 3.97e-11    -2.33     -1.27
```

Based on regression adjustment, the estimated effect of RHC on 30-day survival is approximately –1.47 days. This suggests that, after controlling for differences in patient characteristics (age, sex, weight, temperature, and clinical history), patients who received RHC survived on average 1.47 fewer days within the 30-day window compared to those who did not.

The 95% confidence interval $[-1.96, -0.98]$ does not include zero, and the p-value is highly significant, indicating strong evidence of a negative association between RHC and short-term survival in this observational sample.

(f)

```
# Updated estimates with bootstrap CI for stratification
results <- data.frame(
  Method = c("Unadjusted", "Matching", "Stratification", "Regression Adjustment"),
  Estimate = c(-1.52, -1.89, -1.62, -1.47),
  CI_Lower = c(-2.06, -2.48, -2.25, -1.96),
  CI_Upper = c(-0.98, -1.30, -0.97, -0.98),
  p_value  = c(4.50e-08, 4.05e-10, 0.0008, 0.00001)
)

library(knitr)
kable(results, digits = 3, caption = "Comparison of Treatment Effect Estimates Across Method
```

Table 1: Comparison of Treatment Effect Estimates Across Methods (with 95% CI and p-values)

Method	Estimate	CI_Lower	CI_Upper	p_value
Unadjusted	-1.52	-2.06	-0.98	0.000
Matching	-1.89	-2.48	-1.30	0.000
Stratification	-1.62	-2.25	-0.97	0.001
Regression Adjustment	-1.47	-1.96	-0.98	0.000

Interpretation and Comparison

All four methods show the same result: patients who received RHC had shorter 30-day survival than those who did not. The unadjusted estimate shows a –1.52 day difference, but this is likely biased due to differences between the treated and control groups.

The matching method shows the largest effect (–1.89 days), likely because it balances the groups better and accounts for more confounding. The stratification method gives an effect of –1.62 days. The bootstrap confidence interval is [–2.25, –0.97] with a p-value of 0.0008, showing a clear and significant negative effect.

The regression adjustment method finds a –1.47 day effect, also statistically significant with a confidence interval that excludes zero. Although the effect size varies by method, all estimates are negative and statistically significant. This suggests that RHC is linked to shorter short-term survival in this dataset.

Conclusion

All four methods — unadjusted, matching, stratification, and regression adjustment — show that RHC is associated with lower 30-day survival. The effects range from –1.47 to –1.89 days, and all confidence intervals exclude zero, showing consistent evidence of harm.

However, these are observational results. Unmeasured factors may still affect the estimates, so we cannot be sure RHC causes the lower survival. The results are strong, but causal claims should be made carefully.

Q2

(a)

We are conducting 20 independent two-sample t-tests at a significance level ( ) of 0.05, and all null hypotheses are true.

Distribution

Each t-test has a 5% chance ($ = 0.05$) of producing a false positive (Type I error) if the null hypothesis is true. Since the tests are independent, the number of significant results follows a binomial distribution:

$X \sim \text{Binomial}(n = 20, p = 0.05)$

Expected number of significant results:

$E(X) = n \cdot p = 20 \cdot 0.05 = 1$

Aswer:

- The number of significant tests follows a Binomial(20, 0.05) distribution.

- We expect 1 test out of 20 to be significant by chance alone.

(b)

The R code simulates 1000 experiments, each with 2 independent t-tests, and checks if at least one p-value is ≤ 0.05.

```
# Simulate 1000 experiments, each with 20 independent t-tests
rejections <- replicate(1000, {
  pvals <- replicate(20, t.test(rnorm(100), rnorm(100))$p.value)
  any(pvals <= 0.05)  # Check if any null is rejected
})
mean(rejections)
```

[1] 0.638

Interpretation:

- This will give the probability that at least one null hypothesis is rejected across 20 tests.
- Since all nulls are true, any rejection is a Type I error.

(c)

We want to find the probability of making at least one Type I error among k independent tests at level α.

derivation:

- For one test: $P(\text{No Type I error}) = 1 - \alpha$
- For k independent tests: $P(\text{No Type I error in all } k \text{ tests}) = (1 - \alpha)^k$
- So, the probability of at least one Type I error is: $P(\text{At least one } H_0 \text{ rejected}) = 1 - (1 - \alpha)^k$

Verification for part (b):

Use $\alpha = 0.05$ and $k = 20$:

$$P(\text{At least one rejection}) = 1 - (1 - 0.05)^{20} = 1 - 0.95^{20} \approx 1 - 0.3585 \approx 0.6415$$

So, the probability is about 64% — i.e., there's a 64% chance of making at least one false discovery when running 20 tests at the 5% level.

This matches what we'd expect from the simulation in (b).

(d)

We analyze the function:

$P(\text{At least one rejection}) = 1 - (1 - \alpha)^k$

As k increases (more tests), $(1 - \alpha)^k$ gets smaller, so the overall probability increases.

Conclusion:

- The experiment-wise error rate increases as k increases (with fixed α.
- When conducting many tests, the chance of making at least one false positive increases dramatically — this is why corrections for multiple testing (e.g., Bonferroni) are often necessary.

(e)

- 50 variables were tested for differences between men and women.
- One of these showed a "significant" result at the 1% level (= 0.01).

why there might be a problem

This is a classic multiple comparisons problem.

- With 50 tests at = 0.01, the probability of at least one false positive is:

$$P(\text{At least one false positive}) = 1 - (1 - 0.01)^{50} \approx 1 - 0.605 = 0.395$$

So, there's a 39.5% chance of getting at least one significant result just by chance

Interpretation:

The significant difference could be a false positive, especially since it wasn't a pre-planned hypothesis but found after testing many variables. This is called data dredging or p-hacking. Researchers should correct for multiple comparisons or pre-register hypotheses to avoid misleading results.