Wine and Obesity Data Analysis

Adrian T. and Olivia K.

2025-12-09

Table of contents

1. Dataset Overview **2**
A. Wine Quality Dataset . . . 2
Sampling Distribution of Density . . . 2
B. Obesity Dataset . . . 3
Sampling Distribution of Weight . . . 5

2. One-Sample T-Test: Wine Alcohol Content **5**
2.1 Research Question and Hypothesis . . . 5
2.2 Assumptions . . . 6
2.3 Test . . . 6
2.4 Conclusion:** . . . 6

3. Bootstrap Approach: Wine pH **7**
3.1 Bootstrap Distribution . . . 7
3.2 QQ Plot for Bootstrap . . . 8
3.3 Bootstrap Confidence Interval . . . 8

4. Analysis of Variance (Wine Dataset) **9**
4.1 F Test . . . 9
4.2 ANOVA Table . . . 9
4.3 AVOVA Assumptions Check . . . 9
4.4 Conclusion . . . 9

5. Multiple Comparisons (Obesity Dataset) **9**
5.1 ANOVA Model and Assumptions . . . 9
5.2 QQ Plot and Residual Analysis: . . . 10
5.3 F-Tests for Factors . . . 10
5.4 Pairwise Comparisons . . . 11

6. References **12**

7. Code Appendix **12**

1. Dataset Overview

A. Wine Quality Dataset

- **Collection Year:** 2009
- **Study Type:** Observational; each wine sample measured independently.
- **Description:** Physiochemical properties for red and white Portuguese "Vinho Verde" wines.

- **Size:** 6497 rows, 12 columns

- **Variables:** `fixed_acidity`, `volatile_acidity`, `citric_acid`, `residual_sugar`, `chlorides`, `free_sulfur_dioxide`, `total_sulfur_dioxide`, `density`, `ph`, `sulphates`, `alcohol`, `quality`, `color`

Table 1: Dataset 1 Variables

Variable Name	Role	Type	Description
fixed_acidity	Feature	Continuous	
volatile_acidity	Feature	Continuous	
citric_acid	Feature	Continuous	
residual_sugar	Feature	Continuous	
chlorides	Feature	Continuous	
free_sulfur_dioxide	Feature	Continuous	
total_sulfur_dioxide	Feature	Continuous	
density	Feature	Continuous	
pH	Feature	Continuous	
sulphates	Feature	Continuous	
alcohol	Feature	Continuous	
quality	Target	Integer	score between 0 and 10
color	Other	Categorical	red or white

Sampling Distribution of Density

The following histogram shows the sampling distribution of the **sample mean of the `density` variable**.

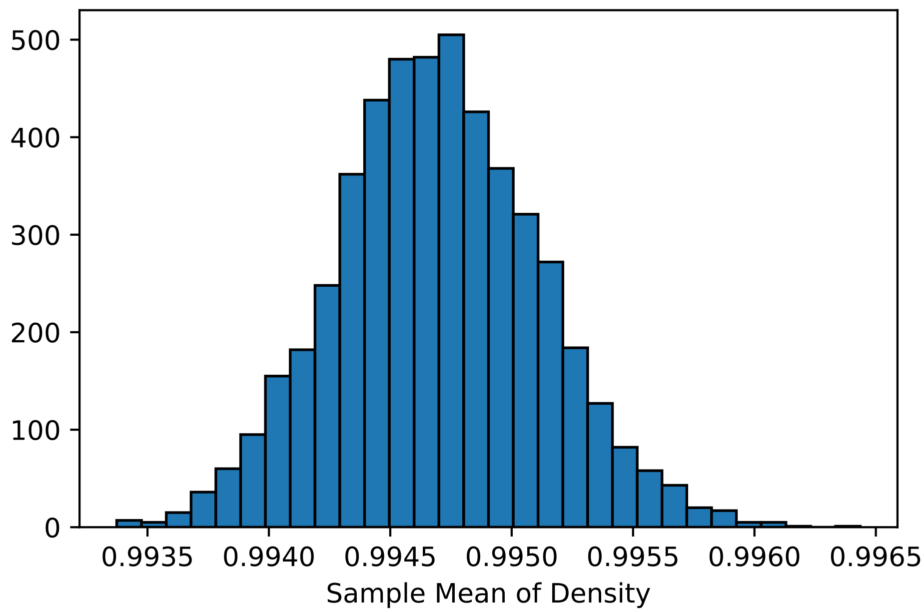

Figure 1: Sampling Distribution of Sample Mean (Density)

Figure 1 shows the sampling distribution of the sample mean of wine densities. The shape of the sampling distribution is approximately normal with little to no skew.

B. Obesity Dataset

- **Collection Year:** 2019
- **Study Type:** Observational; each individual measured independently.
- **Description:** Estimation of obesity levels in individuals from Mexico, Peru, and Colombia, based on eating habits and physical condition.

- **Size:** 2111 rows, 16 columns

- **Variables:** `Gender`, `Age`, `Height`, `Weight`, `Family_history_with_overweight`, `FAVC`, `FCVC`, `NCP`, `CAEC`, `SMOKE`, `CH2O`, `SCC`, `FAF`, `TUE`, `CALC`, `MTRANS`, `NObeyesdad`

Table 2: Dataset 2 Variables

Variable Name	Role	Type	Description	Units
Gender	Feature	Categorical	Biological sex	-

Variable Name	Role	Type	Description	Units
Age	Feature	Continuous	Age in years	Years
Height	Feature	Continuous	Height	Meters
Weight	Feature	Continuous	Weight	Kg
Family_history_with_overweight	Feature	Binary	Family history of overweight	-
FAVC	Feature	Binary	Frequent high-calorie food intake	-
FCVC	Feature	Integer	Frequency of vegetable consumption	-
NCP	Feature	Continuous	Number of main meals per day	Count
CAEC	Feature	Categorical	Snacking between meals	-
SMOKE	Feature	Binary	Smokes or not	-
CH20	Feature	Binary	Monitors calorie intake	-
SCC	Feature	Continuous	Physical activity frequency	Hours/Week
FAF	Feature	Continuous	Daily screen/technology use	Hours
TUE	Feature	Continuous	Daily electronics use	Hours
CALC	Feature	Categorical	Alcohol consumption	-
MTRANS	Feature	Categorical	Mode of transportation	-
NObeyesdad	Target	Categorical	Obesity classification	-

Sampling Distribution of Weight

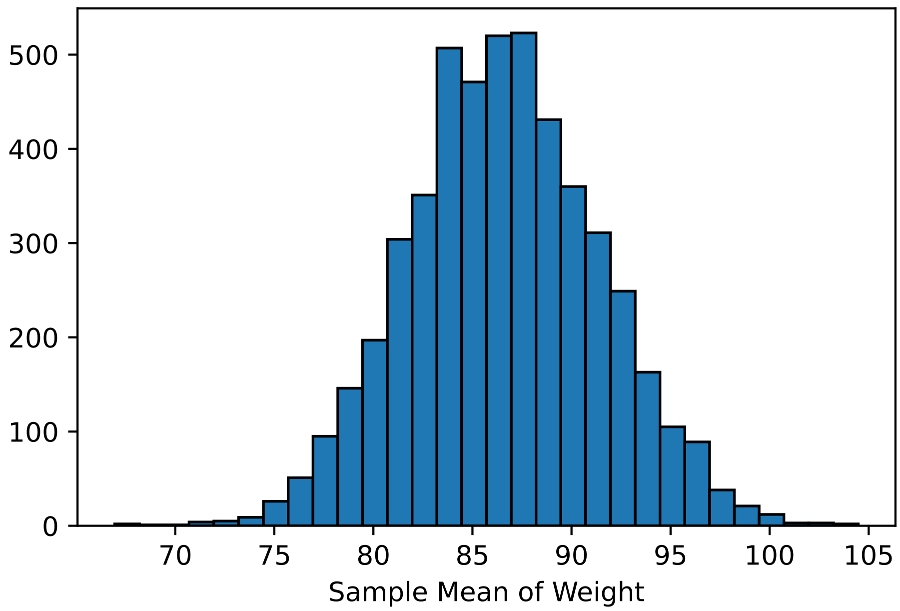

Figure 2: Sampling Distribution of Sample Mean (Weight)

Figure 2 shows the sampling distribution of the sample mean of individual weights.

---

2. One-Sample T-Test: Wine Alcohol Content

2.1 Research Question and Hypothesis

Research Question: Is the average alcohol content of red and white wine greater than 10.5%?

Hypotheses: The null hypothesis states that the population mean alcohol content is equal to 10.5%, while the alternate hypothesis claims the population mean alcohol content of red and white wine is greater than 10.5%.

2.2 Assumptions

2.3 Test

Results:
- Sample mean: 10.4918
- t-statistic: -0.5541
- Critical t-value (= 0.05): 1.6451
- One-tailed p-value: 0.7102
- 95% Confidence Interval: (10.4628, 10.5208)

2.4 Conclusion:**

Since the critical t-value was greater than alpha, we fail to reject the null hypothesis. There is insufficient evidence that the mean alcohol content exceeds 10.5%.

---

3. Bootstrap Approach: Wine pH

3.1 Bootstrap Distribution

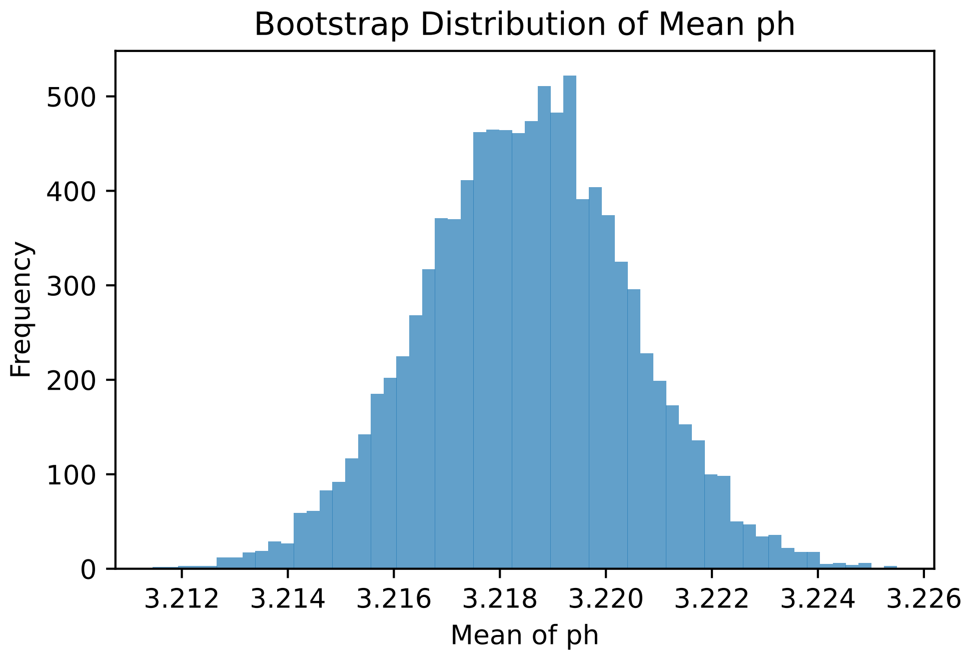

Figure 3: Bootstrap Distribution of Wine pH

3.2 QQ Plot for Bootstrap

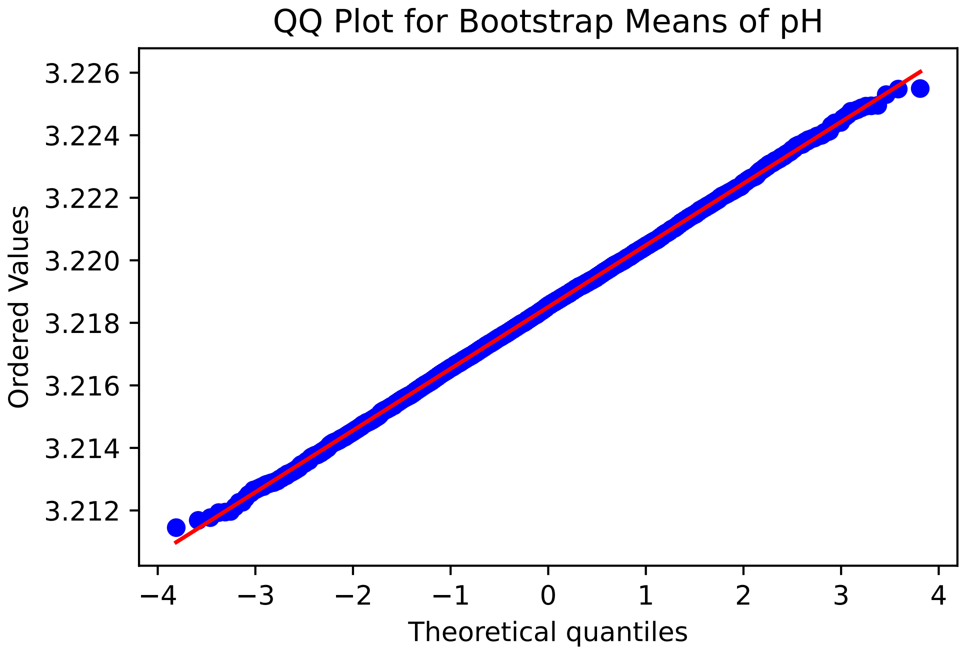

Figure 4: QQ Plot of Bootstrap Means (pH)

3.3 Bootstrap Confidence Interval

```
(3.2146020470986607, 3.2223411574572878)
```

Conclusion: Using an alpha of 0.05, We are 95% confident that the true population mean pH lies between 3.215 and 3.222.

---

4. Analysis of Variance (Wine Dataset)

4.1 F Test

4.2 ANOVA Table

4.3 AVOVA Assumptions Check

4.4 Conclusion

---

5. Multiple Comparisons (Obesity Dataset)

5.1 ANOVA Model and Assumptions

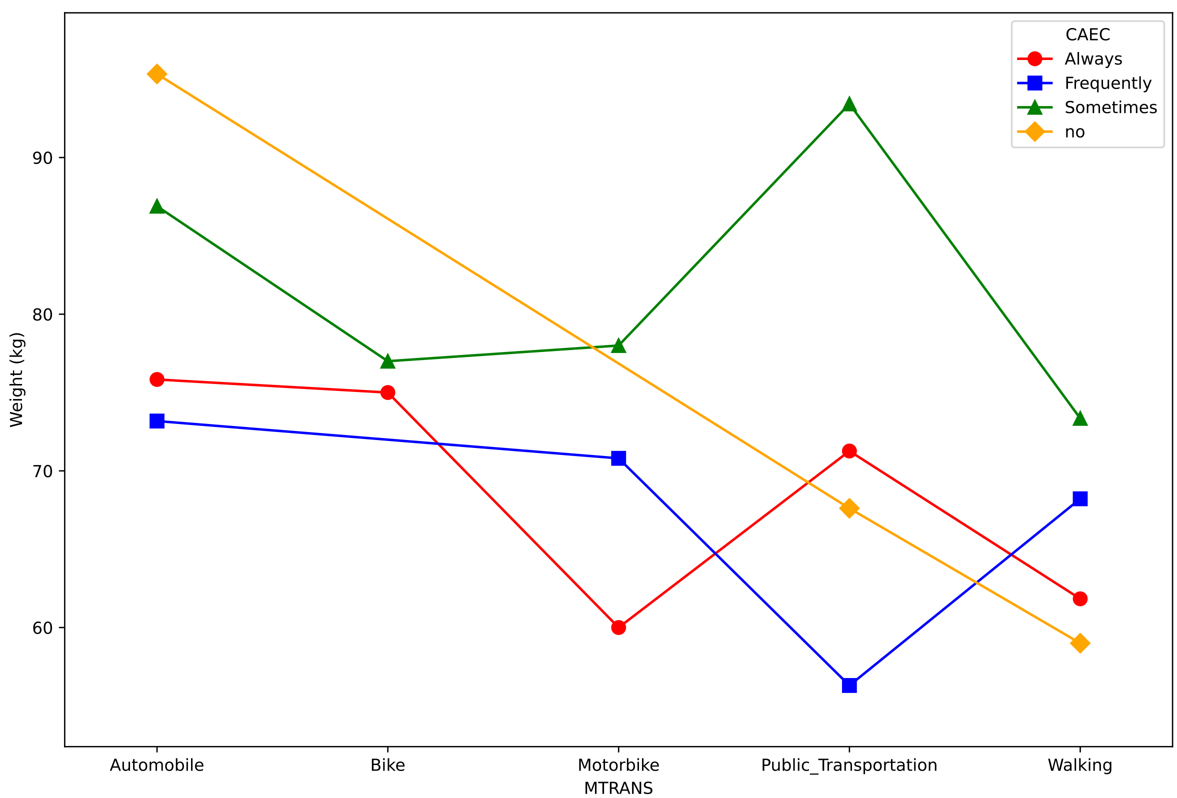

Figure 5: Two-way ANOVA with interaction: Weight ~ MTRANS * CAEC

- **Interpretation:** Based on Figure 5, Weight differs across the different MTRANS (Method of Transportation) categories. The lines for different CAEC overlap, suggesting a relationship exists between MTRANS and CAEC levels.

5.2 QQ Plot and Residual Analysis:

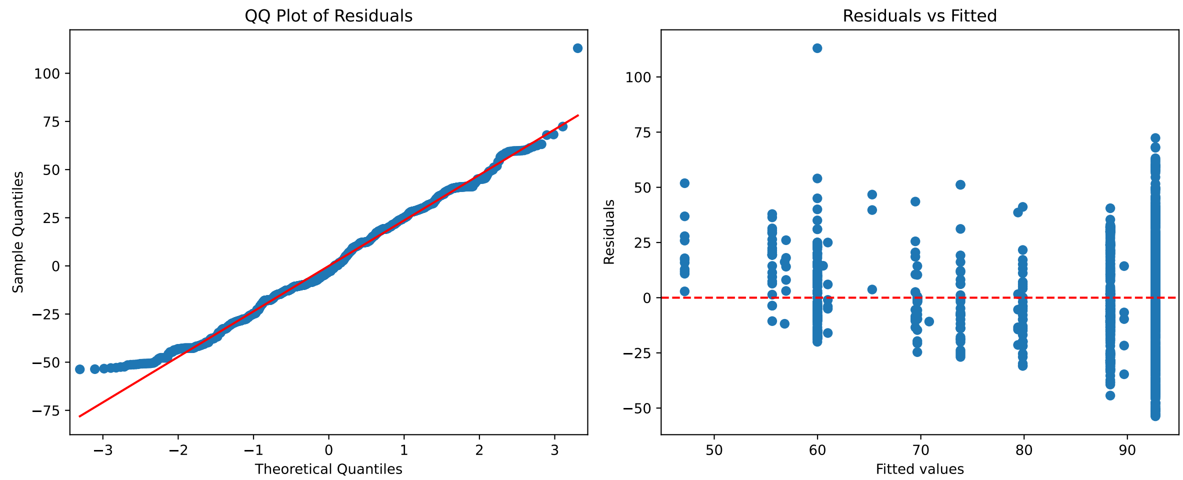

Figure 6: QQ Plot and Residual Plot

- QQ plot shows approximate normality as the data falls around the line.
- Residual plot shows no clear pattern.
- **Conclusion:** ANOVA assumptions satisfied (normality, independence, equal variance).

5.3 F-Tests for Factors

Table 3: ANOVA results for MTRANS and CAEC factors

Factor	F-statistic	F-critical	p-value	Conclusion
MTRANS	6.85	2.376	1.78×10^{-5}	Reject H : At least one group mean differs
CAEC	149.69	2.609	1.11×10^{-16}	Reject H : At least one group mean differs

Interpretation: Both MTRANS and CAEC significantly affect Weight; differences between group means are unlikely due to chance.

5.4 Pairwise Comparisons

Table 4: Tukey HSD Pairwise Comparison for MTRANS

	group1	group2	meandiff	p-adj	lower	upper	reject
0	Automobile	Bike	-9.1933	0.8866	-36.2772	17.8906	False
1	Automobile	Motorbike	-12.8167	0.4894	-34.5151	8.8817	False
2	Automobile	Public_Transportation	1.5791	0.7845	-2.1981	5.3563	False
3	Automobile	Walking	-15.3115	0.0003	-25.3800	-5.2430	True
4	Bike	Motorbike	-3.6234	0.9985	-38.0069	30.7601	False
5	Bike	Public_Transportation	10.7724	0.8109	-16.1659	37.7107	False
6	Bike	Walking	-6.1182	0.9772	-34.6275	22.3911	False
7	Motorbike	Public_Transportation	14.3958	0.3584	-7.1206	35.9122	False
8	Motorbike	Walking	-2.4948	0.9984	-25.9482	20.9586	False
9	Public_Transportation	Walking	-16.8906	0.0000	-26.5606	-7.2206	True

Conclusion: From Figure 5, the Tukey HSD pairwise comparisons revealed a few significant differences between pairs of transportation methods. There were differences between Automobile vs. Walking and also Public Transportation vs. Walking. The mean difference in Weight were 15.3 and 16.9 respectivly. These results suggest that walking is assosiated with a lower average weight.

Table 5: Tukey HSD Pairwise Comparison for CAEC

	group1	group2	meandiff	p-adj	lower	upper	reject
0	Always	Frequently	-12.2049	0.0041	-21.4825	-2.9273	True
1	Always	Sometimes	20.2698	0.0000	11.7416	28.7980	True
2	Always	no	-2.1881	0.9659	-14.1876	9.8115	False
3	Frequently	Sometimes	32.4747	0.0000	28.2813	36.6681	True
4	Frequently	no	10.0168	0.0322	0.5911	19.4425	True
5	Sometimes	no	-22.4579	0.0000	-31.1469	-13.7688	True

Conclusion:

6. References

Cortez, Paulo, et al. "Wine Quality." UCI Machine Learning Repository, 2009, https://doi.org/10.24432/C56S3T.

"Estimation of Obesity Levels Based On Eating Habits and Physical Condition ." UCI Machine Learning Repository, 2019, https://doi.org/10.24432/C5H31Z.

—'

7. Code Appendix