

# Wine and Obesity Data Analysis

Adrian T. and Olivia K.

2025-12-09

## 1. Dataset Overview

### 1.1 Wine Quality Dataset

- **Collection Year:** 2009
- **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
- **Study Type:** Observational; each wine sample measured independently.

### Sampling Distribution of Density

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

### 1.2 Obesity Dataset

- **Collection Year:** 2019
- **Description:** Estimation of obesity levels in individuals from Mexico, Peru, and Colombia, based on eating habits and physical condition.

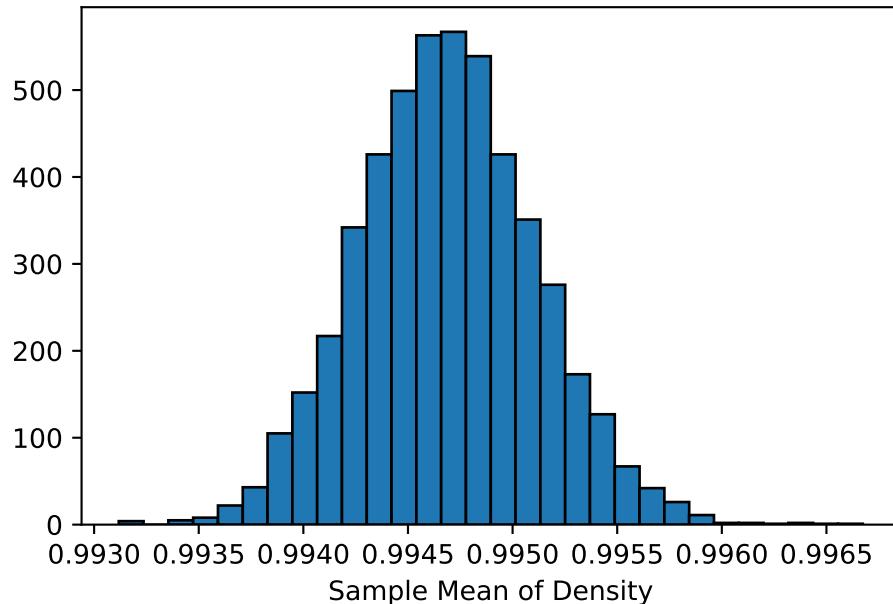


Figure 1: Sampling Distribution of Sample Mean (Density)

- **Size:** 2111 rows, 16 columns
  - **Variables:** Gender, Age, Height, Weight, Family\_history\_with\_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH20, SCC, FAF, TUE, CALC, MTRANS, NObeyesdad
  - **Study Type:** Observational; each individual measured independently.
- 

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

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

**Hypotheses:**

- ( $H_0: \text{mean} = 10.5$ )
- ( $H_a: \text{mean} > 10.5$ )

**Results:**

- Sample mean: 10.4918
- t-statistic: -0.5541

- Critical t-value ( $\alpha = 0.05$ ): 1.6451
- One-tailed p-value: 0.7102
- 95% Confidence Interval: (10.4628, 10.5208)

**Conclusion:** Fail to reject ( $H_0$ ). There is insufficient evidence that the mean alcohol content exceeds 10.5%.

---

### 3. Bootstrap Approach: Wine pH

#### Bootstrap Distribution

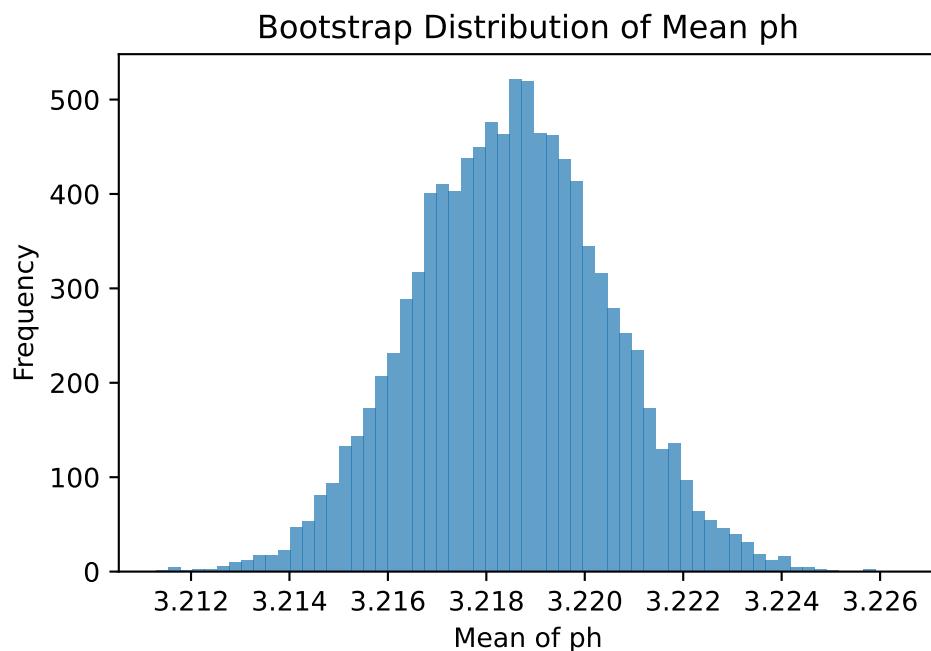


Figure 2: Bootstrap Distribution of Wine pH

#### QQ Plot for Bootstrap

#### 95% Bootstrap Confidence Interval

(3.2146682314914576, 3.222393489302755)

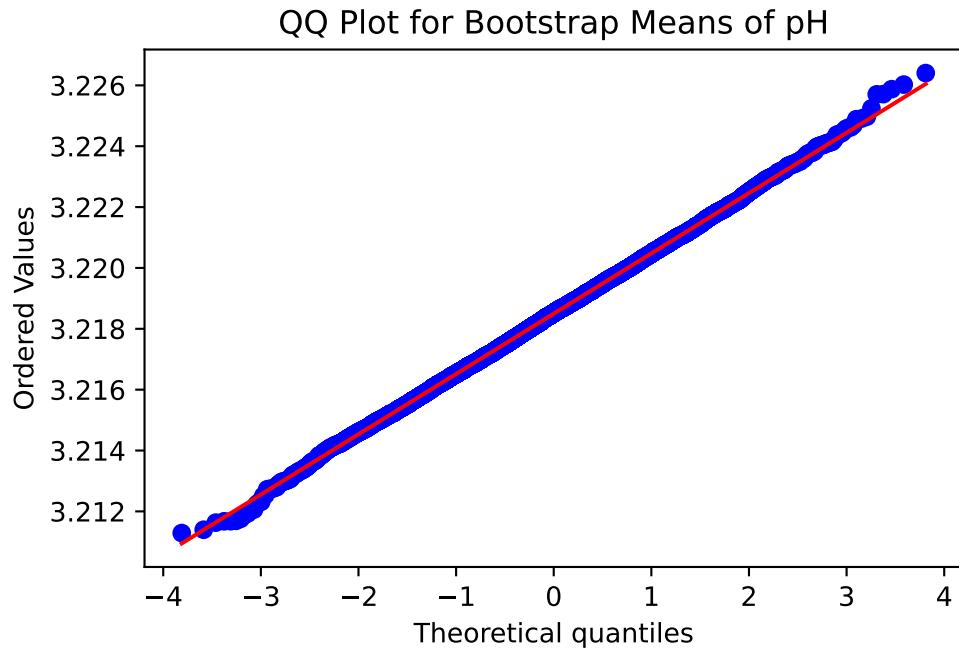


Figure 3: QQ Plot of Bootstrap Means (pH)

**Conclusion:** 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

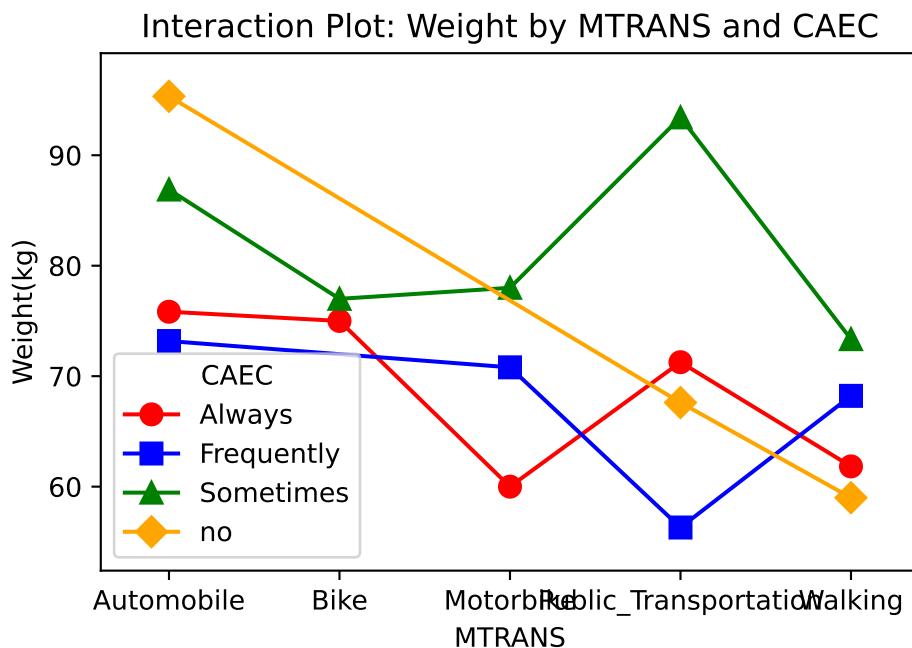
---

## 5. Multiple Comparisons (Obesity Dataset)

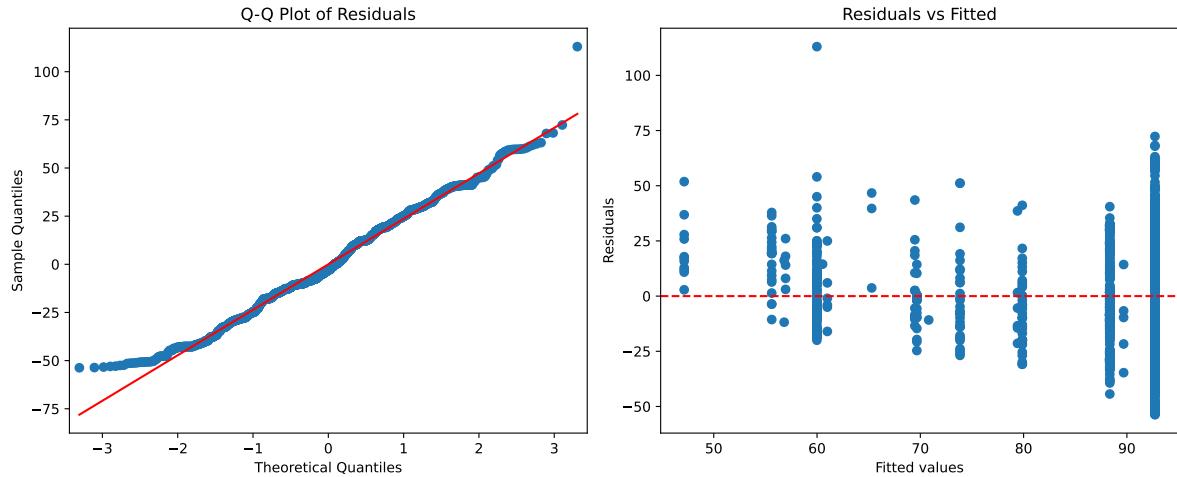
### 5.1 ANOVA Model and Assumptions

- Two-way ANOVA with interaction:  $\text{Weight} \sim \text{MTRANS} * \text{CAEC}$ .

<Figure size 3600x2400 with 0 Axes>



## 5.2 QQ Plot and Residual Analysis:



- QQ plot shows approximate normality.
- Residual plot shows no clear pattern.
- **Conclusion:** ANOVA assumptions satisfied (normality, independence, equal variance).

## 5.3 F-Tests for Factors

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

	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

Figure 4: Tukey HSD Pairwise Comparison for MTRANS

	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

Figure 5: Tukey HSD Pairwise Comparison for CAEC

## **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>.

---

## **Code Appendix**