ST502: Final Project - Part 1

Apostolos Stamenos & Tyler Pollard

4/19/2022

It is plausible that exposure to chemicals in tobacco smoke leads to differential impact in terms of health outcomes. We conducted different version of the two-sample t-test at significance level $\alpha = 0.05$ to formally determine whether or not mean systolic blood pressure differs for smokers and nonsmokers.

Let $Y_{1j}, \ldots, Y_{n_j j}$ be systolic blood pressure measurements from a simple random sample of sample size n_j, where $j = 1$ denotes that the individual was selected from the population of nonsmokers and $j = 2$ denotes that the individual was selected from the population of smokers. For the samples from each population, we assume the parametric model $Y_{1j}, \ldots, Y_{n_j j} \overset{\text{iid}}{\sim} N\left(\mu_j, \sigma_j^2\right)$, where μ_j is the mean systolic blood pressure and σ_j^2 is the unknown variance for population $j \in \{1, 2\}$. We tested the following hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2$$

For the pooled variance t-test, we also make the additional assumption that $\sigma_1^2 = \sigma_2^2$. The two-sample test statistic is: $T = \frac{\left(\overline{Y_1} - \overline{Y_2}\right) - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ is a weighted average of the sample variances S_1^2 and S_2^2. Under H_0, $T \sim t_{n_1+n_2-2}$. We can also construct a confidence interval $CI = \overline{Y_1} - \overline{Y_2} \pm t_{\alpha/2} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Table 1: **Summary of tests**

Test	Point Estimate	SE	df	Test Statistic	p-value	Confidence Interval
Pooled Variance	9.16	3.01	298	3.04	0.0026	(3.23, 15.08)
Satterthwaite						

Discuss results of each test (Tyler)

Both the histograms (Figure 1) and the normal QQ plots (Figure 2) indicate that the data are skewed to the right. The boxplots (Figure 3) indicate the presence of outliers. If the outliers are excluded, the distributions look fairly symmetrical, but we decided to keep the outliers in the analysis. By the Central Limit Theorem, even if the two datasets are not completely normal, their sample means are asymptotically normally distributed. Since the sample sizes for smokers and nonsmokers are sufficiently large, the use of t-tests and confidence intervals is justified by the Central Limit Theorem. The pooled variance and Satterthwaite t-tests differ because of their degrees of freedom and standard errors, so we also had to assess the assumption of equal variances. The boxplots (Figure 3) indicate that the distribution of systolic blood pressure for nonsmokers is more spread out than the distribution of systolic blood pressure for smokers. Based on the boxplots, there is no indication that the true population variances are equal. In addition to visually inspecting the distributions of systolic blood pressure for the two groups, we conducted a formal hypothesis test for equality of variances. We used the median-based extension of the Levene test, as specified in Brown & Forsythe (1974), since this version is more robust to deviations from Normality:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_A : \sigma_1^2 \neq \sigma_2^2$$

With a p-value of 0.045, we reject the null hypothesis of equal variances at the 5% significance level. Thus, we conclude that the t-test with the Satterthwaite approximation is preferred.

Regardless of which test we conduct (pooled vs Satterthwaite) and which method we use (p-value vs confidence interval), we conclude that the mean systolic blood pressure differs for smokers and nonsmokers.

Appendix A: Data Visualizations

Figure 1: **Histograms of systolic blood pressure for smokers and nonsmokers**

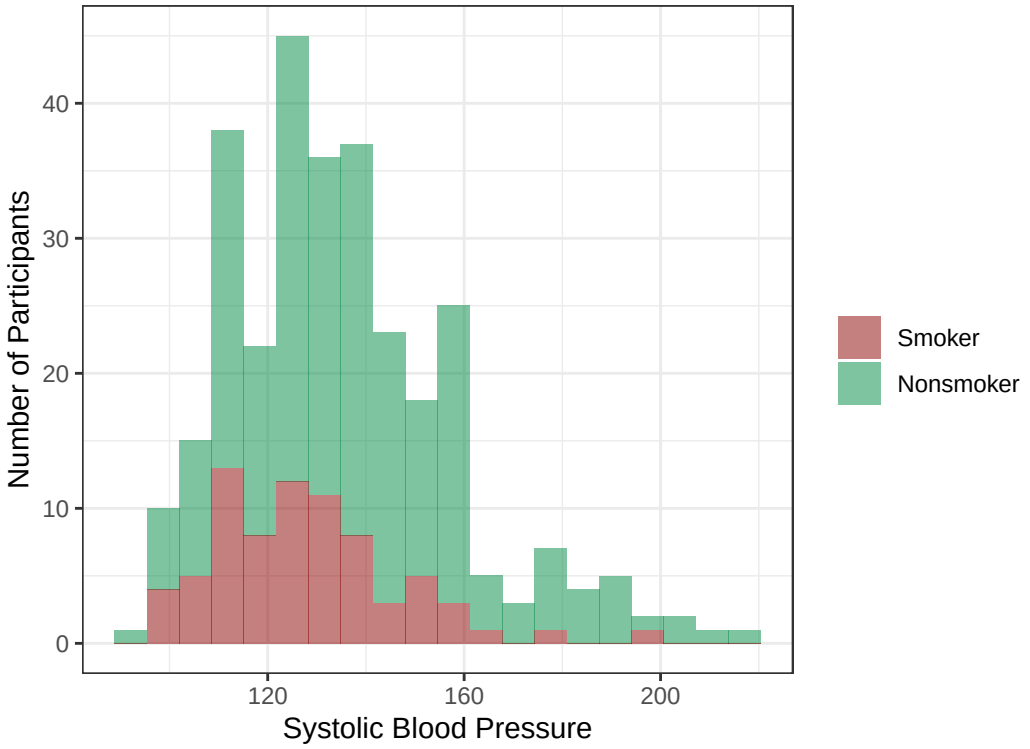


Figure 2: **Normal QQ plots of systolic blood pressure for smokers and nonsmokers**

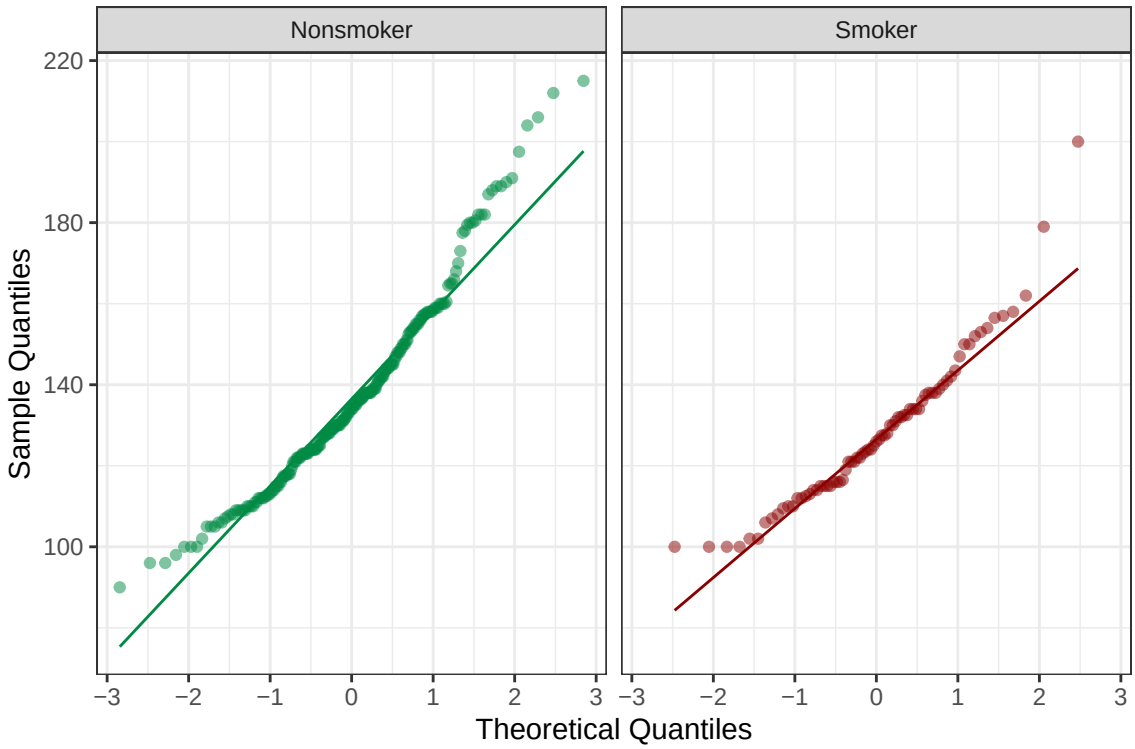
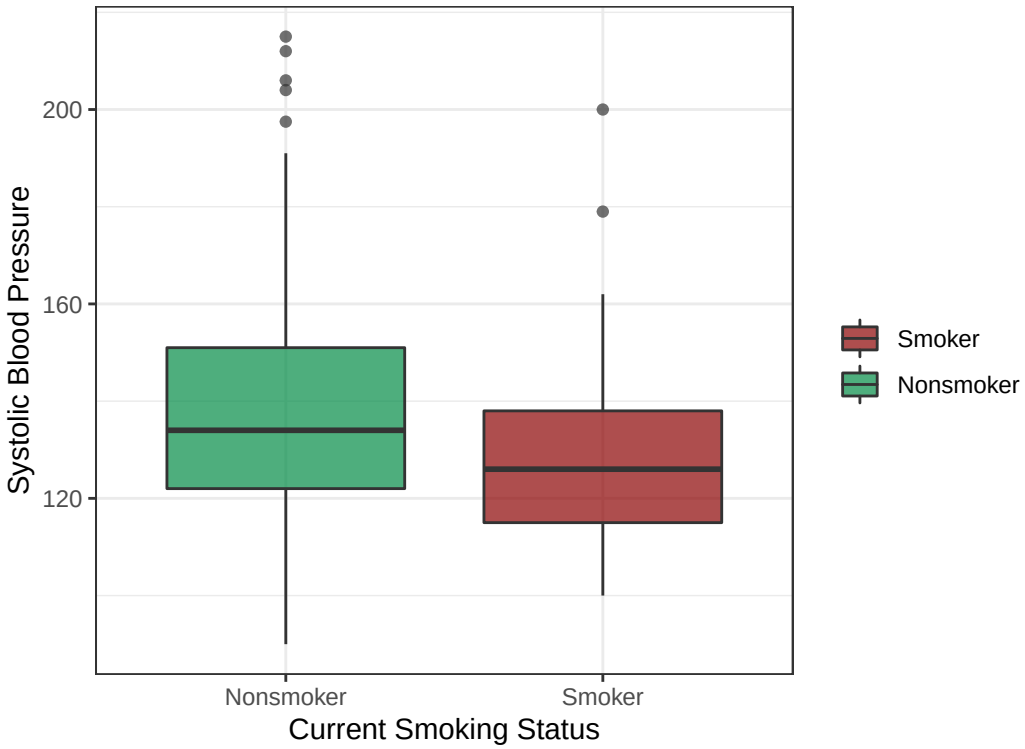


Figure 3: **Boxplots of systolic blood pressure for smokers and nonsmokers**



Appendix B: R Code

Bibliography

Brown, M. B. and Forsythe, A. B. (1974), *Journal of the American Statistical Association*, 69, pp. 364-367