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: $T = \frac{\left(\overline{Y_1} - \overline{Y_2}\right) - 0}{S_p\sqrt{(1/n_1)+(1/n_2)}}$ 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 $CI = \overline{Y_1} - \overline{Y_2} \pm t_{n_1+n_2-2,\alpha/2} \cdot S_p\sqrt{(1/n_1) + (1/n_2)}$.

For the Satterthwaite t-test, we assume that $\sigma_1^2 \neq \sigma_2^2$. The two-sample test statistic is: $T = \frac{\left(\overline{Y_1} - \overline{Y_2}\right) - 0}{\sqrt{\left(S_1^2/n_1\right)+\left(S_2^2/n_2\right)}}$. Under H_0, $T \sim t_\nu$ where the degrees of freedom, $\nu = \frac{\left(\left(s_1^2/n_1\right)+\left(s_2^2/n_2\right)\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1)+\left(s_2^2/n_2\right)^2/(n_2-1)}$ rounded down to the nearest integer. We can also construct a confidence interval $CI = \overline{Y_1} - \overline{Y_2} \pm t_{\nu,\alpha/2} \cdot \sqrt{(S_1^2/n_1) + (S_2^2/n_2)}$.

Table 1: **Summary Statistics of two Samples**

Group	Sample Size	Sample Mean	Sample Variance
Nonsmokers	$n_1 = 225$	$\overline{y_1} = 137.22$	$s_1^2 = 562.14$
Smokers	$n_2 = 75$	$\overline{y_2} = 128.07$	$s_2^2 = 352.21$

The p-value for each test was calculated using the corresponding above equations, Table 1, and p-value $= 2P_{H_0}(T > |t_{obs}|)$ where T is a random variable from the t distribution with the corresponding degrees of freedom for each test. The calculated p-value for each test in Table 1 is less than the alpha level of 0.05 meaning we have significant evidence to reject the null hypothesis of equal means. Both of the calculated 95% confidence intervals about the difference of means for each test in Table 1 do not contain 0 meaning we have significant evidence to reject the null hypothesis of equal means. 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.

Table 2: **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	9.16	2.68	158	3.41	0.0008	(3.86, 14.46)

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.

Appendix A: Data Visualizations

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

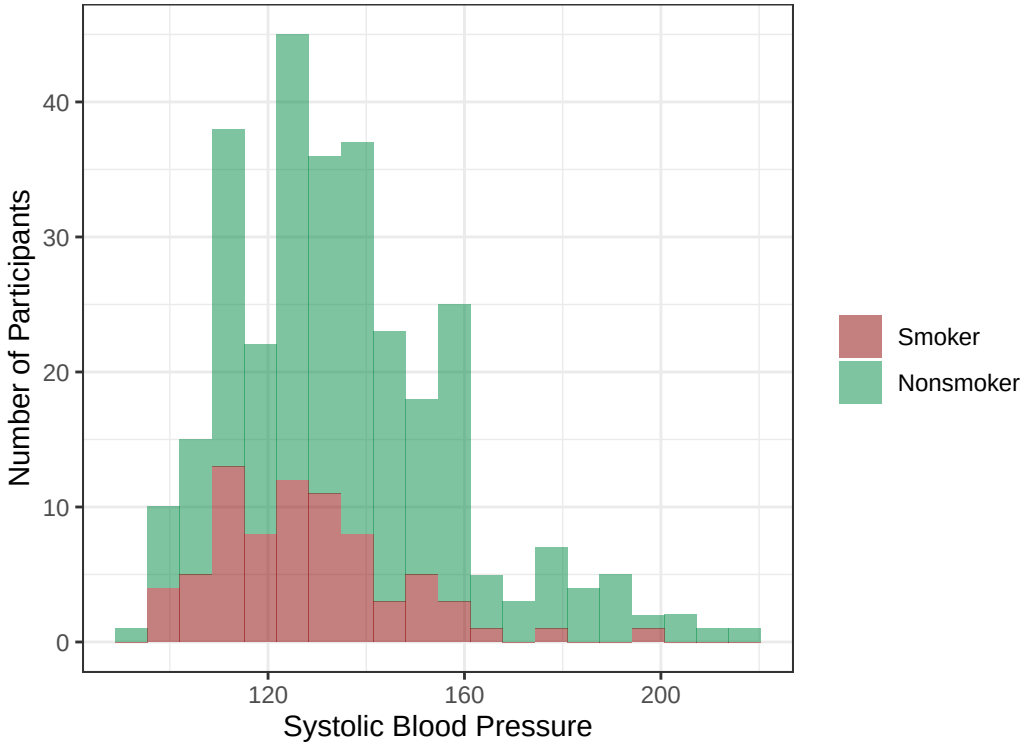


Figure 2: **Normal QQ plots of systolic blood pressure for smokers and nonsmokers**

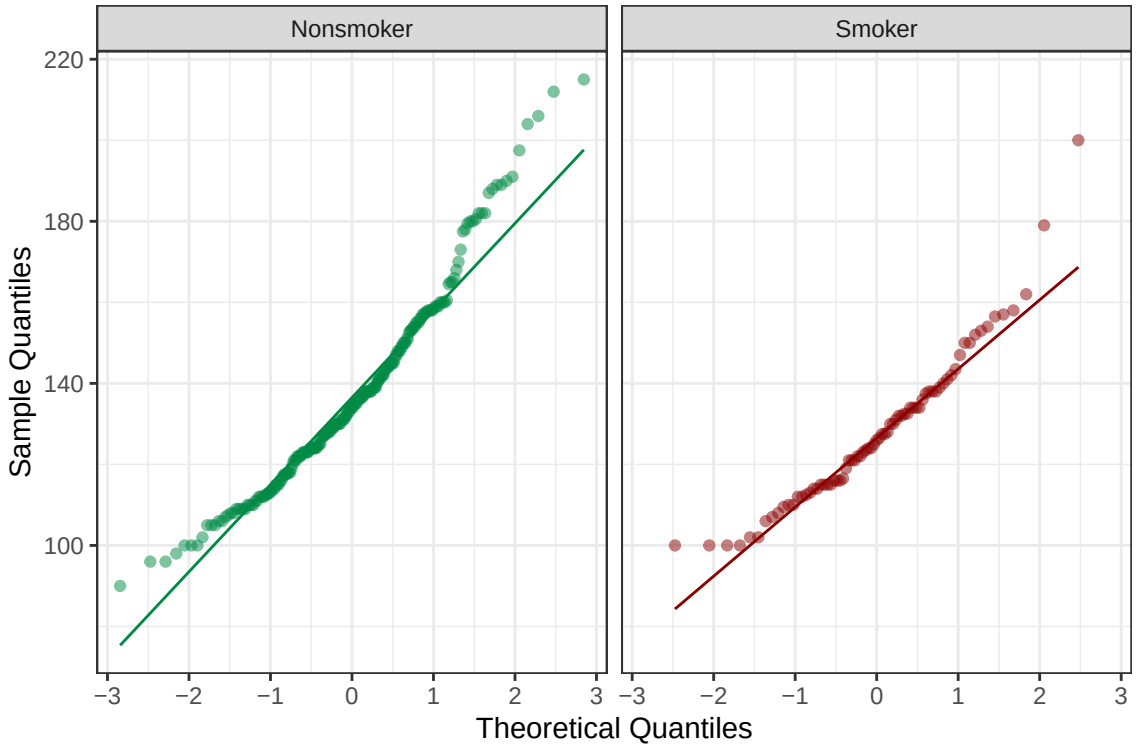
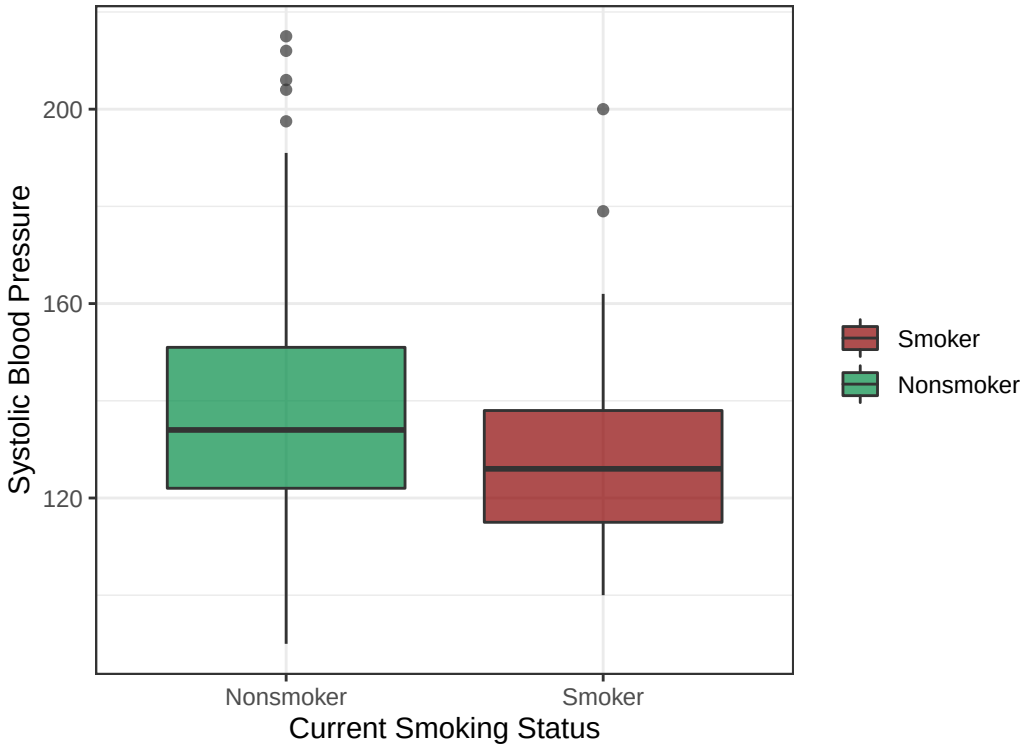


Figure 3: **Boxplots of systolic blood pressure for smokers and nonsmokers**



Appendix B: R Code

```r
## Load required libraries
library(tidyverse)
library(data.table)
library(MESS)
library(car)

## ========================================== PART 1 ==========================================
## Read in data
# This dataset represents two independent samples of systolic blood pressure (sysBP) for
# smokers (currentSmoker=1) and nonsmokers (currentSmoker=0) in the Framingham heart study
data <- fread("framingham_data.csv")

## Filter into smoker and nonsmoker data frames
data <- data %>%
  mutate(currentSmoker = if_else(currentSmoker == 1, 'Smoker', 'Nonsmoker'))
smoker_df <- data %>% filter(currentSmoker == "Smoker")
nonsmoker_df <- data %>% filter(currentSmoker == "Nonsmoker")

## Set alpha level for all tests
alpha <- 0.05

## Shared statistics
# Calculate number of data points in each group
n_nonsmoker <- length(nonsmoker_df$sysBP)
n_smoker <- length(smoker_df$sysBP)

# Calculate sample means
samp_mean_nonsmoker <- sum(nonsmoker_df$sysBP)/n_nonsmoker
samp_mean_smoker <- sum(smoker_df$sysBP)/n_smoker

# Calculate sample variances
samp_var_nonsmoker <- sum((nonsmoker_df$sysBP - samp_mean_nonsmoker)^2)/(n_nonsmoker - 1)
samp_var_smoker <- sum((smoker_df$sysBP - samp_mean_smoker)^2)/(n_smoker - 1)

## -------------------- Pooled Two Sample t-test p-value method ----------------------------
# Calculate the degrees of freedom
df <- n_nonsmoker + n_smoker - 2

# Calculate the point estimate for difference of sample means
diff <- samp_mean_nonsmoker - samp_mean_smoker
# Set the value of true difference of population means to 0 under the null hypothesis
D_0 <- 0

# Calculate pooled sample variance
samp_var_pooled <- ((n_nonsmoker-1)*samp_var_nonsmoker+(n_smoker-1)*samp_var_smoker)/df

# Calculate the standard error
```

```r
se_pooled <- (sqrt(samp_var_pooled)*sqrt(1/n_nonsmoker+1/n_smoker))
# Calculate the observed t statistic
T_pooled <- (diff - D_0)/se_pooled
# Calculate the p-value using observed t statistic and degrees of freedom
p_val_pooled <- 2*pt(abs(T_pooled), df = df, lower.tail = FALSE)

## -------------------- Pooled Two Sample t-test confidence interval method -----------------
# Calculate the 0.025 and 0.975 t quantiles for pooled degrees of freedom
t_quants <- qt(c(alpha/2, 1-alpha/2), df)
# Calculate confidence interval using difference of sample means, t quantiles, and standard error
CI_pooled <- diff+se_pooled*t_quants

## -------------- Satterthwaite Approximation Two Sample t-test p-value method --------------
# Calculate degrees of freedom for t test using Satterthwaite approximation
nu <- (samp_var_nonsmoker/n_nonsmoker + samp_var_smoker/n_smoker)^2/(
  (samp_var_nonsmoker/n_nonsmoker)^2/(n_nonsmoker - 1) + (samp_var_smoker/n_smoker)^2/(n_smoker - 1)
)
# Round down the degrees of freedom to nearest integer
nu <- floor(nu)

# Calculate observed t test statistic
t_obs_satterthwaite <- (samp_mean_nonsmoker - samp_mean_smoker)/sqrt(samp_var_nonsmoker/n_nonsmoker + samp_va

# Calculate the p-value using observed t statistic and degrees of freedom
p_value_satterthwaite <- 2*pt(abs(t_obs_satterthwaite), df = nu, lower.tail = FALSE)

## ---------- Satterthwaite Approximation Two Sample t-test confidence interval method ------
# Calculate the standard error
se_satterthwaite <- sqrt(samp_var_nonsmoker/(n_nonsmoker) + samp_var_smoker/(n_smoker))
# Calculate confidence interval using difference of sample means, 0.025 and 0.975 t quantiles, and standard e
ci_satterthwaite <- c((samp_mean_nonsmoker - samp_mean_smoker) - qt(1-(alpha/2), df = nu)*se_satterthwaite,
                      (samp_mean_nonsmoker - samp_mean_smoker) + qt(1-(alpha/2), df = nu)*se_satterthwaite)

## ------------------------- Checking normal assumption ----------------------------------
## Combined visualizations
# Set colors for each data set
col_nonsmoker <- 'springgreen4'
col_smoker <- 'darkred'

# Histogram
ggplot(data = data) +
  geom_histogram(aes(x = sysBP, fill = currentSmoker), alpha = 0.5, bins = 20) +
  scale_fill_manual(name = '', values = c('Smoker' = col_smoker, 'Nonsmoker' = col_nonsmoker)) +
  labs(x = "Systolic Blood Pressure", y = "Number of Participants") +
  theme_bw()

# QQ Plots
ggplot(data = data, aes(sample = sysBP, color = currentSmoker)) +
  geom_qq(alpha = 0.5) +
```

```r
  geom_qq_line() +
  facet_grid(cols = vars(currentSmoker)) +
  scale_colour_manual(name = '', values = c('Smoker' = col_smoker, 'Nonsmoker' = col_nonsmoker)) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_bw() +
  theme(legend.position = "none")

# Boxplots
ggplot(data = data, aes(x = currentSmoker, y = sysBP, fill = currentSmoker)) +
  scale_fill_manual(name = '', values = c('Smoker' = col_smoker, 'Nonsmoker' = col_nonsmoker)) +
  labs(x = "Current Smoking Status", y = "Systolic Blood Pressure") +
  geom_boxplot(alpha = 0.7) +
  theme_bw()

## Check equal variance assumption using Levene Test for medians
leveneTest(sysBP ~ as.factor(currentSmoker), data = data, center = "median")
```

Bibliography

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