HW 1

1

a. {(A, U), (A, E), (B, U), (B, E), (C, U), (C, E), (D, U), (D, E), (F, U), (F, E)}
b. {(has cancer, young), (has cancer, middle), (has cancer, old), (no cancer, young), (no cancer, middle), (no cancer, old)}
c. {(smoker, recently used), (smoker, used in past), (smoker, never used), (non-smoker, recently used), (non-smoker, used in past), (non-smoker, never used)}
d. {(high intelligence, Border Collie), (high intelligence, German Sheprad), (high intelligence, Dachshund), (medium intelligence, Border Collie), (medium intelligence, German Sheprad), (medium intelligence, Dachshund), (low intelligence, Border Collie), (low intelligence, German Sheprad), (low intelligence, Dachshund)}

2

a. $\mu_{U_1} = 3 + 4 * 4 = 19$ $\sigma_{U_1} = 4 * 8 = 32$
b. $\mu_{U_2} = -10 + 4 * 2 = -2$ $\sigma_{U_2} = 2 * 8 = 16$
c. $\mu_{U_3} = \frac{1}{4} - 4 = -3.75$ $\sigma_{U_3} = 1 * 8 = 8$
d. $\mu_{U_4} = \frac{3}{4} - 4 * \frac{1}{4} = -\frac{1}{4}$ $\sigma_{U_4} = \frac{1}{4} * 8 = 2$

3

a. $E[\bar{Y}] = 20$ $\sigma^2[\bar{Y}] = 25/10 = 2.5$
b. $E[sum(Y_i)] = 20 * 10 = 200$ $\sigma^2[sum(Y_i)] = 25 * 10 = 250$
c. $E[Y^*] = a + E[\bar{Y}] = a + 20$ $\sigma^2[Y^*] = b^2 * \sigma^2[\bar{Y}] = b^2 * 200$
d. $E[Y^*] = 5 - 2 * 200 = -395$ $\sigma^2[Y^*] = 2^2 * 250 = 1000$

4

a. This is a normal distribution because a linear combination of normal distributions is also normal. $E[\bar{Y}_1 - \bar{Y}_2] = \mu_1 - \mu_2$ $\sigma^2[\bar{Y}_1 - \bar{Y}_2] = \frac{\sigma_1^2}{100} + \frac{\sigma_2^2}{100}$
b. This is a normal distribution because a linear combination of normal distributions is also normal. $E[\bar{Y}_1 + \bar{Y}_2] = \mu_1 + \mu_2$ $\sigma^2[\bar{Y}_1 + \bar{Y}_2] = \frac{\sigma_1^2}{100} + \frac{\sigma_2^2}{100}$
c. This is a normal distribution because a linear combination of normal distributions is also normal. $E[\frac{\bar{Y}_1 + \bar{Y}_2}{2} - \bar{Y}_3] = \frac{\mu_1 + \mu_2}{2} - \mu_3$ $\sigma^2[\frac{\bar{Y}_1 + \bar{Y}_2}{2} - \bar{Y}_3] = \frac{\frac{\sigma_1^2}{100} + \frac{\sigma_2^2}{100}}{4} + \frac{\sigma_3^2}{100}$
d. This is a normal distribution because a linear combination of normal distributions is also normal. $E[\bar{Y}_1 + \bar{Y}_3 - 2\bar{Y}_2] = \mu_1 + \mu_3 - 2\mu_2$ $\sigma^2[\bar{Y}_1 + \bar{Y}_3 - 2\bar{Y}_2] = \frac{\sigma_1^2}{100} + \frac{\sigma_3^2}{100} + 4 * \frac{\sigma_2^2}{100}$

5(a)

The null hypothesis is that the mean of smokers is equal to the mean of nonsmokers. The alternate hypothesis is that the two means differ.

5(b)

```
## t_s =  4.40371
```

Sample difference in mean is 4.4037 standard deviations away from the null hypothesis.

5(c)

$p < 0.0005$ The p-value is the area of the tails in the standard distribution where we reject the null hypothesis.

5(d)

We reject the null hypothesis because $p < \alpha = 0.05$.

6(a)

Type I error would be incorrectly rejecting the null hypothesis, where we would fail to state that average systolic blood pressure is the same across smoking statuses when it is.

6(b)

Type II error is to fail to reject the null hypothesis, where we would fail to state that average systolic blood pressure differs between smoking statuses when it is.

6(c)

```
## Confidence interval is [4.479227, 17.22077]
```

6(d)

The confidence interval means that we are 99% certain the difference between the two means is between 4.479227 and 17.22077.

6(e)

The largest difference we expect between the two groups with 99% confidence is 17.22077. This means that we expect the mean of non-smokers to be at most 17.22077 less than the mean of smokers.

7

a. This is a mixed study because there are both observational and experimental variables.
b. The primary variable of interest is test scores.
c. The explanatory variables consists of observational and experimental variables. The high school and section number are the observational variables. The instruction (standard or computer-based) is the experimental variable.
d. {(hs 1, section 1, computerized), (hs 1, section 1, in person), (hs 1, section 2, computerized), (hs 1, section 2, in person), (hs 1, section 3, computerized), (hs 1, section 3, in person), (hs 1, section 4, computerized), (hs 1, section 4, in person), (hs 2, section 1, computerized), (hs 1, section 2, in person), (hs 2, section 2, computerized), (hs 2, section 2, in person), (hs 2, section 3, computerized), (hs 2, section 3, in person), (hs 2, section 4, computerized), (hs 2, section 4, in person), (hs 3, section 1, computerized), (hs 3, section 1, in person), (hs 3, section 2, computerized), (hs 3, section 2, in person), (hs 3, section 3, computerized), (hs 3, section 3, in person), (hs 3, section 4, computerized), (hs 3, section 4, in person)}

8

a. This is a mixed study because there are both observational and experimental variables.
b. The primary variable of interest is number of days required for successful completion.
c. The explanatory variables consists of observational and experimental variables. The prior physical fitness status (below average, average, above average) is the observational variable. The doctor they are paired with (out of three possible doctors) is the experimental variable.
d. {(above average, doctor 1), (average, doctor 1), (below average, doctor 1), (above average, doctor 2), (average, doctor 2), (below average, doctor 2), (above average, doctor 3), (average, doctor 3), (below average, doctor 3)}

I(a)

```
##   stress sbp_mean
## 1     HS 147.0172
## 2     LS 142.3510
## 3     MS 145.1429
```

High Stress had the highest average sbp.

I(b)

```
##   stress   sbp_sd
## 1     HS 28.03991
## 2     LS 28.28667
## 3     MS 27.67478
```

Yes, the standard deviations are approximately equal.

I(c)

```
##   exercise age_mean
## 1        H 41.34320
## 2        L 39.80000
## 3        M 39.33824
```

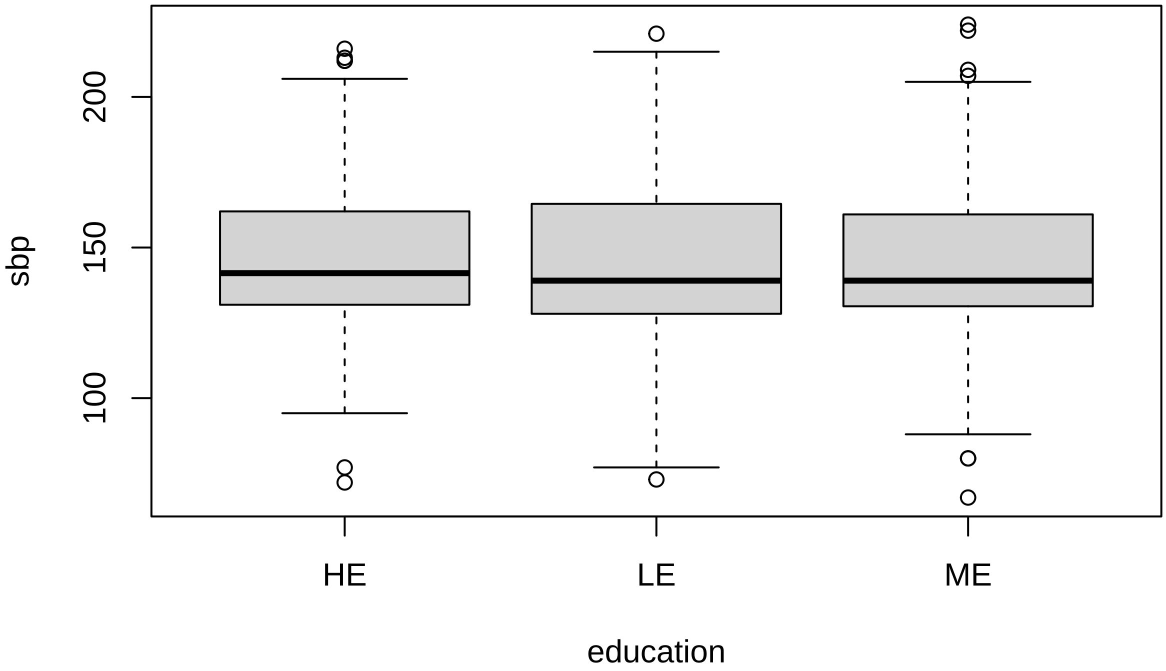
Medium exercise has the lowest average age.

I(d)

```
##   exercise   age_sd
## 1        H 12.33561
## 2        L 14.13980
## 3        M 13.20618
```

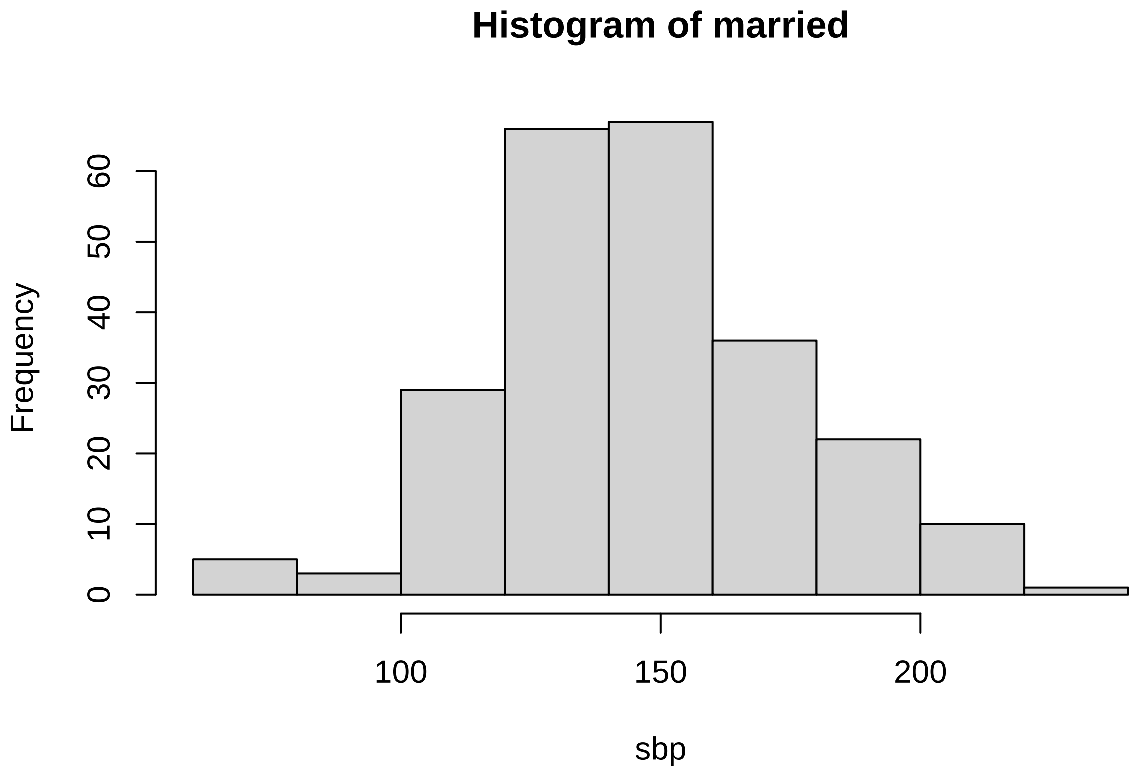
Low exercise group differs most from its mean as it has the highest standard deviation.

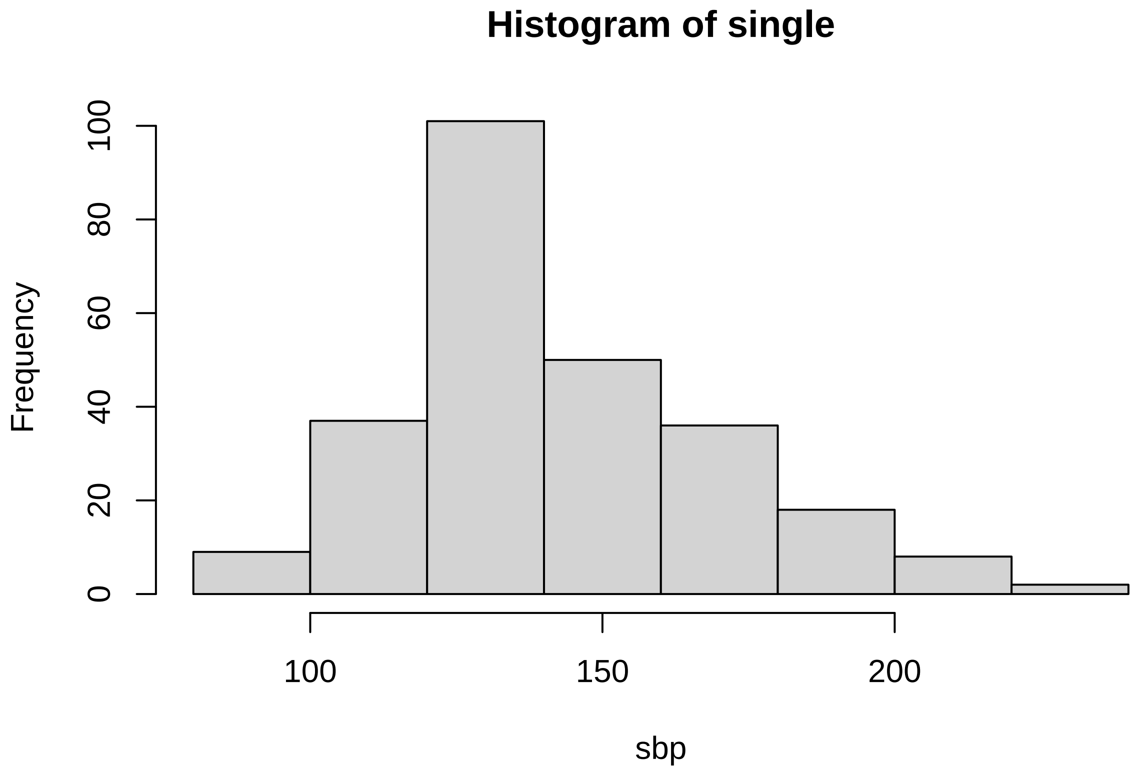
II(a)



There does not seem to be a trend as the boxplots have approximately the same means, lower quartiles, and upper quartiles. The lower education has a slightly lower lower quartile and slightly higher upper quartile than the others.

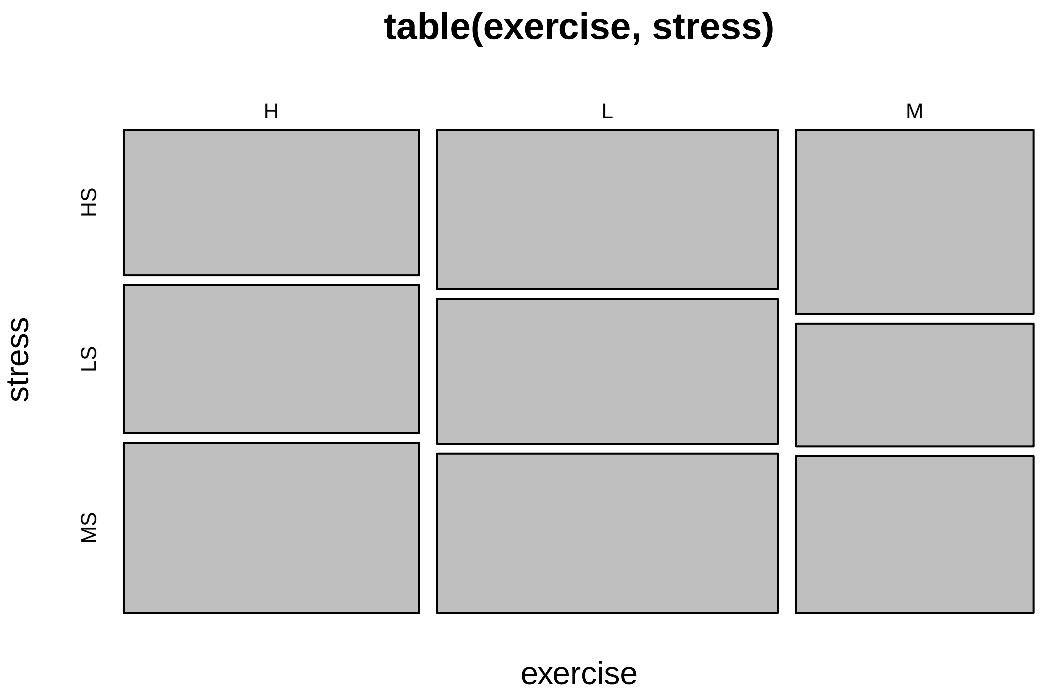
II(b)


Histogram of married
Frequency
0
10
20
30
40
50
60
100
150
200
sbp


Histogram of single
Frequency
0
20
40
60
80
100
100
150
200
sbp

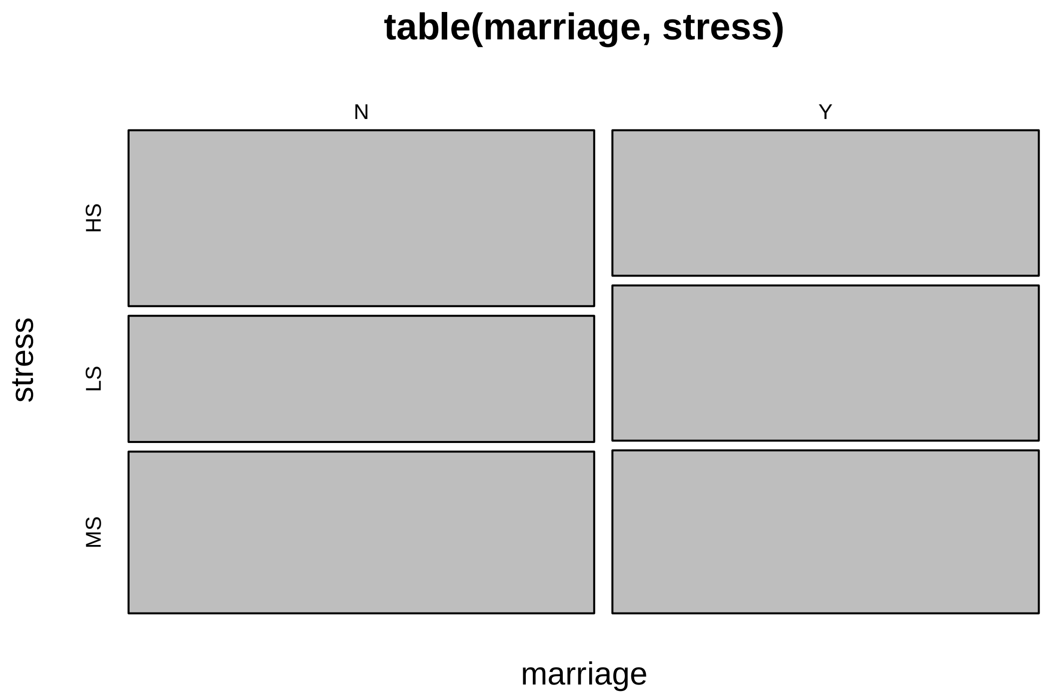
Both groups vary around the same amount since the range of data is around the same.

II(c)



Medium exercise has the highest proportion of high stress subjects.

II(d)



The has marriage group has the highest proportion of low stress subjects.

III(a)

```
## exercise
##   H   L   M
## 169 195 136
```

Low exercise group has the most subjects.

III(b)

```
## high_ed_stress
## HS LS MS
## 57 51 62
```

Middle stress group has the highest age.

III(c)

```
##   stress age_mean
## 1     HS 40.53448
## 2     LS 39.18543
## 3     MS 40.73143
```

Middle stress group has the highest average age.

III(d)

```
##   gender sbp_mean
## 1      F 144.8902
## 2      M 145.0212
```

Females have the lowest average systolic blood pressure.

IV(a)

```
## t_s =  1.352333
```

IV(b)

```
## p =  0.1768947
```

IV(c)

```
## [1] -1.537372  8.326066
## attr(,"conf.level")
## [1] 0.95
```

Confidence interval is $[-1.537372, 8.326066]$.

IV(d)

The confidence interval tells us the range one mean lies within the other on a normal distribution.

IV(e)

The interval shows that one mean can lie above or below the other mean, so we cannot confidently say which group has a higher average sbp.

Appendix

```
# for correct formatting of paper
knitr::opts_chunk$set(echo = FALSE)

# preload all data
data <- read.csv("GSK.csv")
```

```r
stress <- data$stress
sbp <- data$sbp
exercise <- data$exercise
age <- data$age
education <- data$educatn
marriage <- data$married
gender <- data$gender

# 5-b
mean_1 <- 150.03
mean_2 <- 139.18
s_1 <- 27.49
s_2 <- 27.49
n_1 <- 266
n_2 <- 234

s_p_squared <- (s_1**2 * (n_1 - 1) + s_2**2 * (n_2 - 1)) / (n_1 + n_2 - 2)
sqrt_thingy <- sqrt(s_p_squared * (1 / n_1 + 1 / n_2))
t_s <- (mean_1 - mean_2) / sqrt_thingy

cat("t_s = ", t_s)

# 6-c
t <- 2.585718
lower <- (mean_1 - mean_2) - t * sqrt_thingy
upper <- (mean_1 - mean_2) + t * sqrt_thingy

cat("Confidence interval is [", lower, ", ", upper, "]", sep = "")

# I-a
res <- aggregate(sbp ~ stress, data = data, FUN = mean)
colnames(res)[colnames(res) == "sbp"] <- "sbp_mean"

res

# I-b
res <- aggregate(sbp ~ stress, data = data, FUN = sd)
colnames(res)[colnames(res) == "sbp"] <- "sbp_sd"

res

# I-c
res <- aggregate(age ~ exercise, data = data, FUN = mean)
colnames(res)[colnames(res) == "age"] <- "age_mean"

res

# I-d
res <- aggregate(age ~ exercise, data = data, FUN = sd)
colnames(res)[colnames(res) == "age"] <- "age_sd"

res
```

```r
# II-a
boxplot(sbp ~ education, data = data, xlab = "education", ylab = "sbp")

# II-b
married <- data[marriage == "Y", "sbp"]
single <- data[marriage == "N", "sbp"]

hist(married, xlab = "sbp")
hist(single, xlab = "sbp")

# II-c
mosaicplot(table(exercise, stress))

# II-d
mosaicplot(table(marriage, stress))

# III-a
table(exercise)

# III-b
high_ed_stress <- data[education == "HE", "stress"]
table(high_ed_stress)

# III-c
res <- aggregate(age ~ stress, data = data, FUN = mean)
colnames(res)[colnames(res) == "age"] <- "age_mean"
res

# III-d
res <- aggregate(sbp ~ gender, data = data, FUN = mean)
colnames(res)[colnames(res) == "sbp"] <- "sbp_mean"
res

# IV-a
res <- t.test(married, single, data = data)
cat("t_s = ", res$statistic)

# IV-b
cat("p = ", res$p.value)

# IV-c
res$conf.int
```