STAT30040 Homework 4

Sarah Adilijiang

Problem 1

(a) Pooled two-sample t-test

```
calcium = read.table("calcium2.txt", col.names=c("oxalate","flame"))
x = calcium$oxalate;    y = calcium$flame

# sample means
x_bar = mean(x);    y_bar = mean(y)

# pooled sample variance
n=nrow(calcium)
s_x = sqrt(sum((x-x_bar)^2)/(n-1))  #or s_x = sd(x)
s_y = sqrt(sum((y-y_bar)^2)/(n-1))  #or s_y=sd(y)
s_p = sqrt((s_x^2 + s_y^2)/2)

# test statistic T
T_stat = (x_bar-y_bar)/(s_p*sqrt(2/n)); T_stat
```

```
## [1] 0.1902449
```

```
# p-value
p_val = 2*pt(T_stat, 2*(n-1), lower.tail=FALSE); p_val
```

```
## [1] 0.8492822
```

```
# check the results
t.test(x,y,mu=0, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 0.19024, df = 234, p-value = 0.8493
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3330060  0.4041924
## sample estimates:
## mean of x mean of y
##  2.387542  2.351949
```

Answer:

$H_0 : \mu_X = \mu_Y \ vs \ H_a : \mu_X \neq \mu_Y$

Two samples Xi's and Yi's are independent and we are assuming equal variances ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$).

So $s_p^2 = \frac{(n-1)s_X^2+(n-1)s_Y^2}{n+n-2} = \frac{s_X^2+s_Y^2}{2}$

Under H_0, approximately we have: $T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{n}}} \sim t(n+n-2) = t(2n-2)$, where $n = 118$

So under H_0: $T = \frac{\bar{X}-\bar{Y}}{s_p\sqrt{\frac{2}{n}}} \approx 0.1902$

Thus p-value: $p = 2 \times P(t(2n-2) > 0.1902) \approx 0.8493$

Since p-value is large, we do not reject null hypothesis. Therefore, there is no significance evidence that the two methods are different from each other.

(b) Paired t-test

```r
# sample mean of D
D_bar = x_bar - y_bar

# sample variance of D
D = x-y
s_D = sqrt(sum((D-D_bar)^2)/(n-1))   #or s_D = sd(x-y)

# test statistic T
T_stat = D_bar/(s_D/sqrt(n)); T_stat
```

```
## [1] 4.172354
```

```r
# p-value
p_val = 2*pt(T_stat, n-1, lower.tail=FALSE); p_val
```

```
## [1] 5.818985e-05
```

```r
# check the results
t.test(x,y,mu=0, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 4.1724, df = 117, p-value = 5.819e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01869856 0.05248788
## sample estimates:
## mean of the differences
##              0.03559322
```

Answer:

$H_0 : \mu_X = \mu_Y \ vs \ H_a : \mu_X \neq \mu_Y$

Set $D_i = X_i - Y_i$, so $\bar{D} = \bar{X} - \bar{Y}$, and $s_D^2 = \frac{\sum (D_i - \bar{D})^2}{n-1}$

Under H_0, approximately we have: $T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \sim t(n-1)$, where $n = 118$

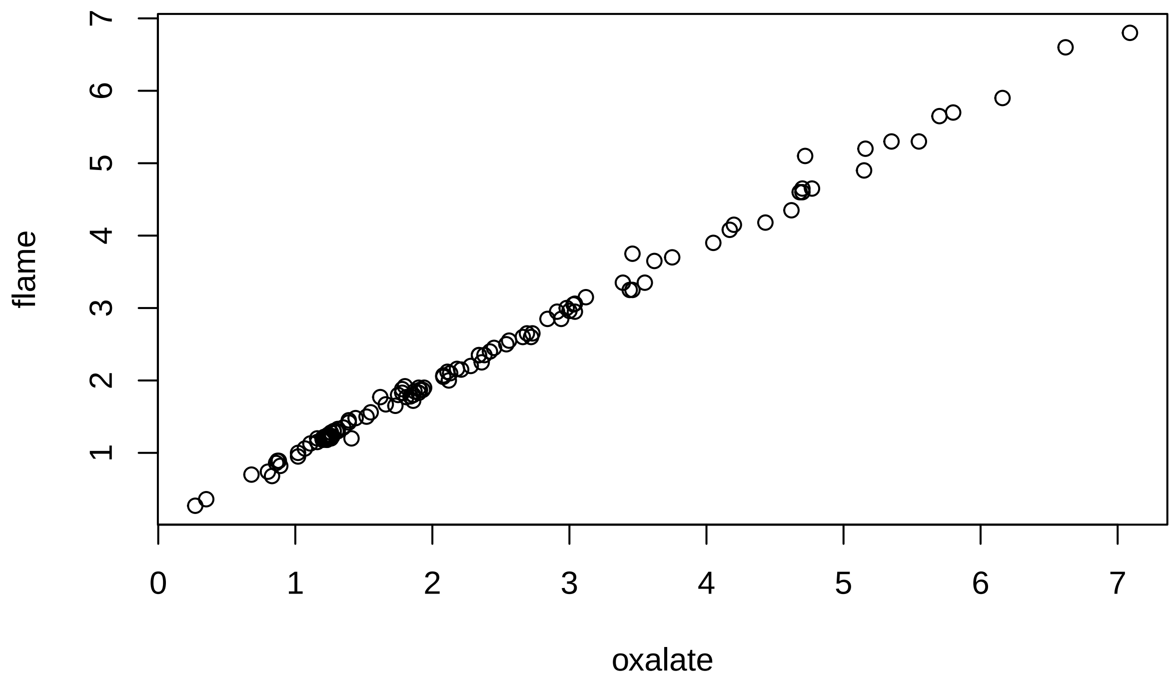
So under H_0: $T = \frac{\bar{D}}{s_D/\sqrt{n}} \approx 4.1724$

Thus p-value: $p = 2 \times P(t(n-1) > 4.1724) \approx 5.8190 \times 10^{-5}$

Since p-value is very small, we reject the null hypothesis. Therefore, there is statistically significant evidence that the two methods are different from each other.

(c) sample correlation coefficient

```r
# scatterplot
plot(x, y, xlab="oxalate", ylab="flame")
```




```
# sample correlation coefficient
r = sum((x-x_bar)*(y-y_bar)) / ((n-1)*s_x*s_y); r
```

```
## [1] 0.9981424
```

```
# check the results
cor(x,y)
```

```
## [1] 0.9981424
```

```
# 95% CI for true correlation coefficient
z = qnorm(0.975, 0, 1, lower.tail=TRUE)
sd = sqrt(1/(n-3))
w = 0.5*log((1+r)/(1-r))
L = w - z*sd
R = w + z*sd
c((exp(2*L)-1)/(exp(2*L)+1), (exp(2*R)-1)/(exp(2*R)+1))
```

```
## [1] 0.9973237 0.9987108
```

Answer:

Sample correlation coefficient: $\hat{\rho} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)s_X s_Y} \approx 0.9981$

When n is large, approximately we have: $\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} \sim N(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3})$

So a $1-\alpha = 95\%$ CI for $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$ is: $(\frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}, \frac{1}{2}\log\frac{1+\hat{\rho}}{1-\hat{\rho}} + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}})$. Let's set it equals to (L, R)

Thus a $1-\alpha = 95\%$ CI for true correlation coefficient ρ is: $(\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2R}-1}{e^{2R}+1}) \approx (0.9973, 0.9987)$

(d)

Answer:

The sample correlation coefficient in question (c) is very close to 1, indicating that the two samples are highly correlated in this example, thus the paired t-test is more appropriate for this situation. In the pooled two-sample t-test, the samples are assumed to be independent, which are actually violated by the data here.

Problem 2

(a) Pooled two-sample t-test

```
ozone = read.csv("ozonerats.csv")
x = ozone$control;    y = ozone$treat

# pooled two-sample t-test (equal variances)
t.test(x,y,mu=0, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 2.4919, df = 43, p-value = 0.01664
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.177186 20.656806
## sample estimates:
## mean of x mean of y
##  22.42609  11.00909
```

Answer:

Since p-value $= 0.01664 < 0.05$, we reject the null hypothesis at 5% significance level. Therefore, there is statistically significant evidence that ozone does have effects on the weight gains of rats.

(b) Welch two-sample t-test

```
# Welch two-sample t-test (unequal variances)
t.test(x,y,mu=0, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 2.4629, df = 32.918, p-value = 0.01918
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.985043 20.848949
## sample estimates:
## mean of x mean of y
##  22.42609  11.00909
```

Answer:

Since p-value $= 0.01918 < 0.05$, we reject the null hypothesis at 5% significance level. Therefore, there is statistically significant evidence that ozone does have effects on the weight gains of rats.

(c)

```
# sample variances
s_x = sd(x); s_x^2
```

```
## [1] 116.1384
```

```
s_y = sd(na.omit(y)); s_y^2
```

```
## [1] 361.6504
```

```
# F-test for comparing variances
var.test(x,y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.32113, num df = 22, denom df = 21, p-value = 0.01072
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1341539 0.7619765
## sample estimates:
## ratio of variances
##          0.3211344
```

Answer:

The sample variance of treatment group is much larger that of control group, and in the F-test for comparing two variances, the p-value $= 0.01072 < 0.05$, so we reject the null hypothesis that the variances are equal. Therefore, the data here violates the assumption that the variance are equal in the pooled two-sample t-test, thus the two-sample t-test with unequal variance is more appropriate for this situation.

(d) Paired t-test

```
# paired t-test
t.test(x,y,mu=0, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 2.4056, df = 21, p-value = 0.02544
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.521593 20.932953
## sample estimates:
## mean of the differences
##                11.22727
```

Answer:

Since p-value $= 0.02544 < 0.05$, we reject the null hypothesis at 5% significance level. Therefore, there is statistically significant evidence that ozone does have effects on the weight gains of rats.

(e)

```
# sample correlation coefficient
cor(na.omit(ozone)$control, na.omit(ozone)$treat)
```

```
## [1] 0.007799049
```

Answer:

The sample correlation coefficient is very close to 0, indicating that the two samples are not correlated in this example, thus the two-sample t-test assuming independence is more appropriate for this situation. In the paired t-test, the samples are assumed to be dependent, which are actually violated by the data here.