STAT 614 - HW 7

1.

a. We are interested in testing Ho: The population mean pulmonary function (mean ln(FEV)) is the same for smokers and non-smokers. Vs. Ha: The population mean pulmonary function (mean ln(FEV)) for non-smokers is greater.

A test and CI for differences in mean ln(FEV) between the two groups suggests that smokers have average ln(FEV) that is 1.1604750 - 0.8883953 = 0.2720807 log(FEV) units higher in the smokers than non-smokers. The 95% confidence interval has endpoints (rounded to two decimal places) of 0.19 to 0.36 which suggests mean log(FEV) is higher in smokers than non-smokers (p = 2.363×10^{-10} using equal variances assumed for the two-sided test but it is essentially the same if we don't assume equal variances). If we convert to the hypothesized one-sided test, where Ha: the smokers have lower mean ln(FEV) than non-smokers, the p-value > 0.99999.

Note that the confidence interval suggests the median FEV of smokers is 1.2 to 1.4 times the median FEV of non-smokers (with 95% confidence).

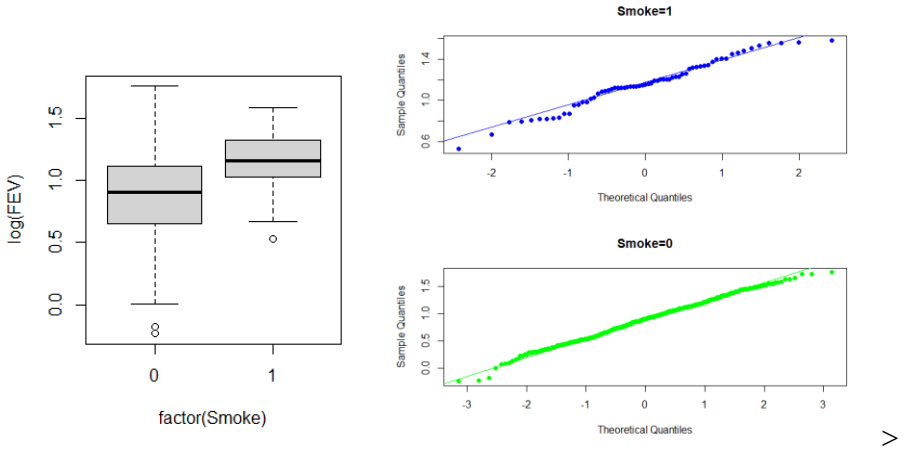
```
> t.test(log(FEV)~factor(Smoke),data=data,var.equal=T)

        Two Sample t-test

data:  log(FEV) by factor(Smoke)
t = -6.4369, df = 652, p-value = 2.364e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3550798 -0.1890815
sample estimates:
mean in group 0 mean in group 1
      0.8883953       1.1604760

> exp(0.3550798)
[1] 1.426294
> exp(0.1890815)
[1] 1.208139
```

Note that using LN(FEV) we see that the normality assumption is reasonably met but the equal variances assumption is suspect, although the sample variances are not *too* difference from each other with $S^2 = 0.110$ for non-smokers and $S^2 = 0.055$ for smokers (the variance of non-smokers is double that of smokers, so it is not *too* bad but possibly of concern. There are several outliers in the boxplots but with the sample sizes so large I am less concerned about them.



>

```
favstats(log(FEV)~factor(Smoke),data=data)
  factor(Smoke)        min        Q1    median       Q3      max      mean        sd   n missing
1             0 -0.2344573 0.6523252 0.9021918 1.114486 1.756650 0.8883953 0.3316671 589       0
2             1  0.5270926 1.0278321 1.1534161 1.322022 1.583505 1.1604760 0.2342048  65       0
```

```
> var(log(data$FEV[data$Smoke==1]))
[1] 0.0548519
> var(log(data$FEV[data$Smoke==0]))
[1] 0.1100031
```

b. This seems weird, smokers have higher median pulmonary function (forced expiratory volume). So - Yes! This is surprising as I would expect lung function to be better in non-smokers, on average, than in smokers. This suggests the opposite! Wassup with that?!

2.

a. The test of no association is equivalent to testing if the slope coefficient is 0: Ho: $\beta_1 = 0$ vs. Ha: $\beta_1 \neq 0$. The test statistics is t = 6.437 with a p-value < 0.001. There is a lot of evidence of an association between smoking status and average FEV.

b. The estimated slope coefficient is 0.272, thus we estimate that the average LN(FEV) is 0.272 higher in the smokers than in the non-smokers. (0.19 to 0.36 with 95% confidence). Note that this suggests the median FEV of smokers is 1.2 to 2.4 times the median FEV of non-smokers

c. These are (and should be) identical results to the two-independent samples t-procedure (assuming equal variances) results from #1! I've highlighted the relevant identical output below.

d. If smoking status had three levels, then we would include **two** indicator variables in the model. (DO NOT treat smoking status with the 0/1/2 values as a quantitative variable in your model! Smoking status is categorical and should be treated as thus!) Taking never smoked as the reference group, I1 defined below indicates past smoker vs. never smoked and I2 indicates current smoker vs. never smoked.

Smoking Status	I1	I2
0 = never smoked	0	0
1 = past smoker	1	0
2 = current smoker	0	1

```
> fit <- lm(log(FEV)~Smoke,data=data)
> #fit <- lm(log(FEV)~Age+Smoke+I(Age^2),data=data)
> summary(fit)

Call:
lm(formula = log(FEV) ~ Smoke, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12285 -0.22803  0.01238  0.21777  0.86825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.88840    0.01333  66.668  < 2e-16 ***
Smoke        0.27208    0.04227   6.437 2.36e-10 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3234 on 652 degrees of freedom
Multiple R-squared:  0.05975,    Adjusted R-squared:  0.05831
F-statistic: 41.43 on 1 and 652 DF,  p-value: 2.364e-10

> anova(fit)
Analysis of Variance Table

Response: log(FEV)
           Df Sum Sq Mean Sq F value    Pr(>F)
Smoke       1  4.334  4.3336  41.434 2.364e-10 ***
Residuals 652 68.192  0.1046
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> confint(fit)
                2.5 %    97.5 %
(Intercept) 0.8622291 0.9145616
Smoke       0.1890815 0.3550798
```