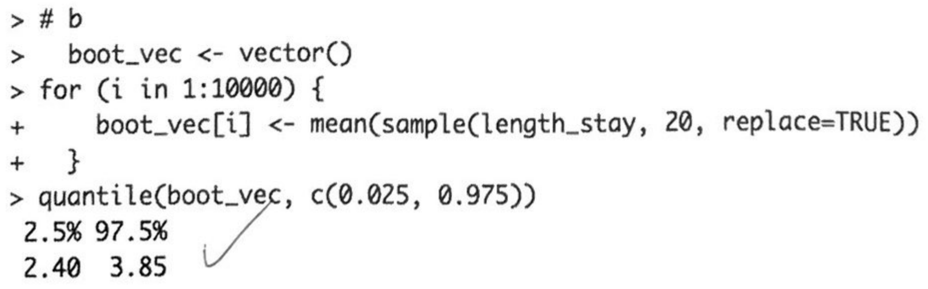
David Levy
10/31/17

Biostatistics 200A
Problem Set 3

1.

a. We can begin calculating a 95% confidence interval by determining a point estimate of the population mean, which we can obtain from the sample mean of the data provided, $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = 3.1 \text{ days}$. We continue by determining the standard error of the means and the $t_{0.025,19}$ and $t_{0.975,19}$ critical values, and constructing our confidence interval as follows: $\bar{X} \pm t_{0.975,19} \bullet \frac{s}{\sqrt{n}} = 3.1 \pm 0.7728 = (2.3272, 3.8728)$. Any time we use a t-distribution, we are operating under the assumption that the target population is normal. Given our previous knowledge of the length of hospital stays, we should observe that it is unlikely for the approach we have employed to be the best method, primarily because our data is right skewed. This can be confirmed by comparing the median and mean of the data, noting that the median is less than the mean. It would probably have been more accurate for us to attempt a logarithmic transformation of the data first.

b. A 95% confidence interval of (2.40, 3.85) was constructed using the bootstrapping method shown below. The results are comparable to those that we attained in part (a).



```
> # b
>   boot_vec <- vector()
> for (i in 1:10000) {
+     boot_vec[i] <- mean(sample(length_stay, 20, replace=TRUE))
+   }
> quantile(boot_vec, c(0.025, 0.975))
 2.5% 97.5%
 2.40  3.85
```

c.

i. A 95% confidence interval using logarithmically transformed data was calculated in much the same way as the confidence interval presented in part (a). The data transformations were treated as unique vectors, and the CI calculation was of the form $\bar{X} \pm t_{0.975,19} \bullet \frac{s}{\sqrt{n}} = 0.9743 \pm 0.2842 = (0.6901, 1.2586)$.

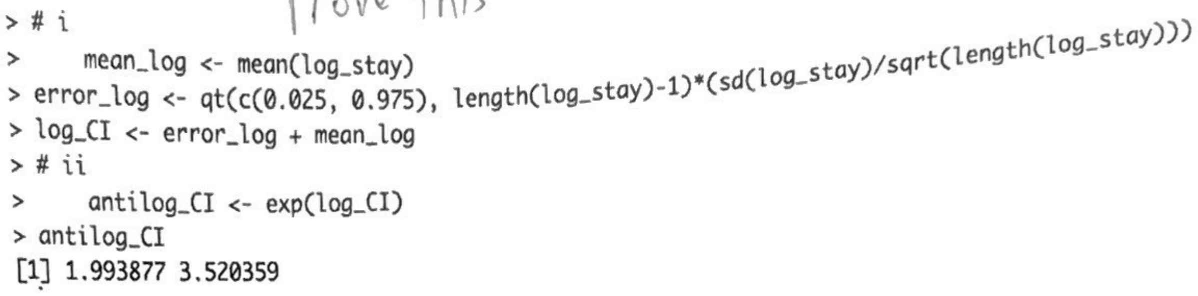
ii. Taking the antilogs of the endpoints of the CI calculated above yields a new confidence interval of (1.9939, 3.5204), which is a confidence

88/90

interval for the median length of stay if the data is lognormally distributed. This is because if the data is lognormally distributed, it is right-skewed, which will affect primarily the mean, but will for the most part allow the median to remain unchanged from what it would be in a normal distribution. The transformation allows us to "pull the mean in" closer to the value of the median, and thus the transformed interval is a confidence interval for the mean of hospital stay length, whereas the antilogs of those endpoints give us a confidence interval for the median of hospital stay length.

-2

Prove this



```
> # i
>     mean_log <- mean(log_stay)
> error_log <- qt(c(0.025, 0.975), length(log_stay)-1)*(sd(log_stay)/sqrt(length(log_stay)))
> log_CI <- error_log + mean_log
> # ii
>     antilog_CI <- exp(log_CI)
> antilog_CI
[1] 1.993877 3.520359
```

2. $n = 24; \bar{X} = 220; s = 35; \mu_0 = 230$
   a. We will use a confidence level of $\alpha = 0.05$ to test our hypothesis that the cholesterol level in the macrobiotic diet group is significantly different than that of the general population. Our null and alternative hypotheses take the form: H$_0$: $\mu = 230$; H$_1$: $\mu \neq 230$. Our n is small in this case, so we should use the Student's t-test. We must assume normality of the underlying distribution under the null hypothesis in order to rely on the following statistic: $t_s = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{220-230}{\frac{35}{\sqrt{24}}} = -1.3997$. When we compare our t-statistic to the critical values of $t_{0.025,23} = -2.0687$ and $t_{0.975,23} = 2.0687$, we see that $t_{0.025,23} < t_s < t_{0.975,23}$ and we cannot reject the null hypothesis. Determining the p-value is a matter of calculating $2P(X < t_s) = 0.175$. We can interpret this value to mean that the likelihood (under the null hypothesis) of obtaining a sample mean as or more extreme than the one we did ($\bar{X} = 220$) is 0.175. This does not meet our 0.05 confidence level requirement, and therefore we cannot reject the null hypothesis. In other words, the evidence that we have collected in this study is insufficient to claim that those with primarily macrobiotic diets have significantly different cholesterol levels than the general population. ✓

   b. A 95% confidence interval calculation is similar to those done in the previous problem. $\bar{X} \pm t_{0.975,23} \bullet \frac{s}{\sqrt{n}} = 220 \pm 14.7792 = (205.2208, 234.7792)$. We ✓

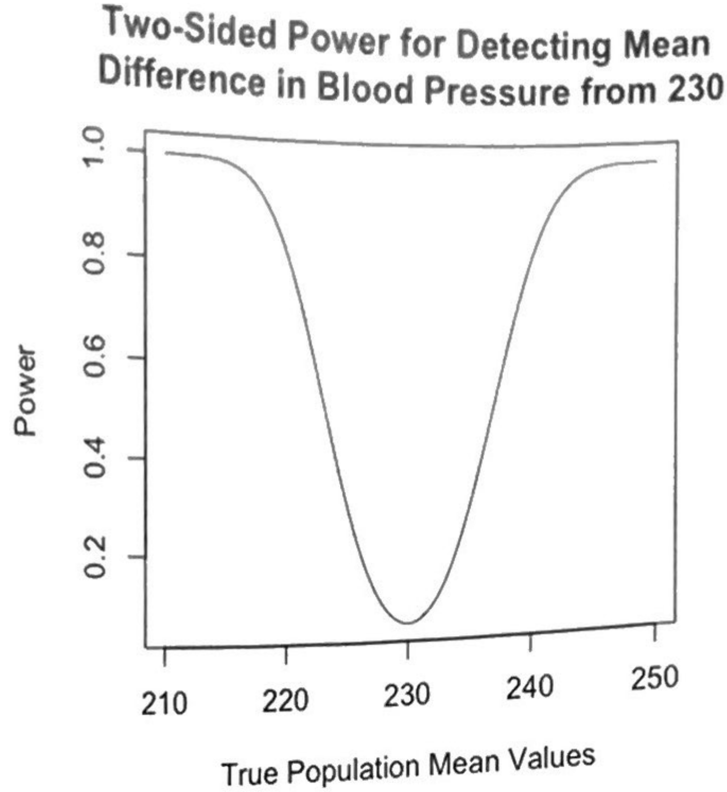
should note that this CI contains $\mu_0 = 230$, which supports the conclusion we made in part (a).

3. $n = 100; \mu_1 = 227; s = 35$

   a. Two-sided power is calculated using $P\left(Z < \frac{\mu_0-\mu_1}{\frac{s}{\sqrt{n}}} - Z_{1-\frac{\alpha}{2}} \middle| \mu = \mu_1\right) + P\left(Z > \frac{\mu_0-\mu_1}{\frac{s}{\sqrt{n}}} + Z_{1-\frac{\alpha}{2}} \middle| \mu = \mu_1\right) = P\left(Z < \frac{230-227}{\frac{35}{\sqrt{100}}} - 1.96\right) + P\left(Z > \frac{230-227}{\frac{35}{\sqrt{100}}} + 1.96\right) = P(Z < -1.1029) + P(Z > 2.8171) = 0.1375$. One-sided power for a left-tailed test is calculated using $P\left(Z < \frac{\mu_0-\mu_1}{\frac{s}{\sqrt{n}}} - Z_{1-\frac{\alpha}{2}} \middle| \mu = \mu_1\right) = P\left(Z < \frac{230-227}{\frac{35}{\sqrt{100}}} - 1.96\right) = 0.1350$. These two power calculations are close, but do not yield equivalent results. Clearly, in this case, the right tail term is not negligible, as it does change our two power values when we evaluate to 4 decimal places. Both of these calculations give us a general idea of how likely we are to avoid a type II error.

Good

   b. We have been asked to perform 2 calculations. One when $\mu_1 = 225$ and one when $\mu_1 = 220$. We already described the mathematics above, and thus we make the following calculations here: In the first case, power $= P\left(Z < \frac{230-225}{\frac{35}{\sqrt{100}}} - 1.96\right) + P\left(Z > \frac{230-225}{\frac{35}{\sqrt{100}}} + 1.96\right) = P(Z < -0.5314) + P(Z > 3.3886) = 0.2979$. ✓ In the latter case, power $= P\left(Z < \frac{230-220}{\frac{35}{\sqrt{100}}} - 1.96\right) + P\left(Z > \frac{230-220}{\frac{35}{\sqrt{100}}} + 1.96\right) = P(Z < 0.8971) + P(Z > 4.1871) = 0.8152$. ✓ The power graph is included below.



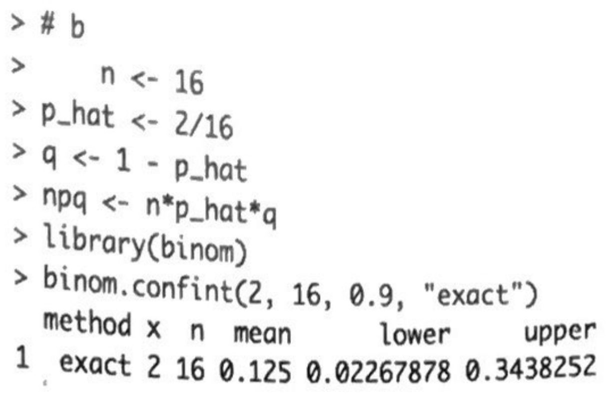
4.

a. These data are binomially distributed. We can estimate incidence by observing that in this sample of $n = 800$ children, $X = 100$ confirmed cases of malaria, giving us $\hat{p} = 0.125$. Importantly, $n\hat{p}(1-\hat{p}) = 87.5 > 5$, which allows us to employ the normal approximation to the binomial distribution. This allows us to construct our 90% confidence interval using $\hat{p} \pm Z_{1-\frac{\alpha}{2}} \bullet$ ✓

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.125 \pm 0.0192 = (0.1058, 0.1442).$$

Good

b. This is an interesting question. Now, $n = 16$ children, and $X = 2$ confirmed cases of malaria, again giving us an estimate of $\hat{p} = 0.125$. However, in this instance, $n\hat{p}(1-\hat{p}) = 1.75 < 5$, so the normal approximation to the binomial distribution is not reasonable. Instead, we have to do an exact calculation involving combination terms and exponentials of $\hat{p}$ and $(1-\hat{p})$. These calculations would be tedious, so we instead employ the "binom" library in R to obtain lower and upper bounds to satisfy $P(B_l < p < B_u) = 0.9$. The R output and 90% confidence interval (labeled "lower" and "upper") are included below.

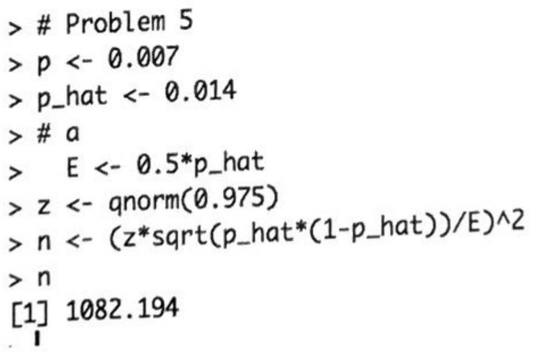


```
> # b
>       n <- 16
> p_hat <- 2/16
> q <- 1 - p_hat
> npq <- n*p_hat*q
> library(binom)
> binom.confint(2, 16, 0.9, "exact")
  method x  n  mean       lower     upper
1  exact 2 16 0.125 0.02267878 0.3438252
```

Good

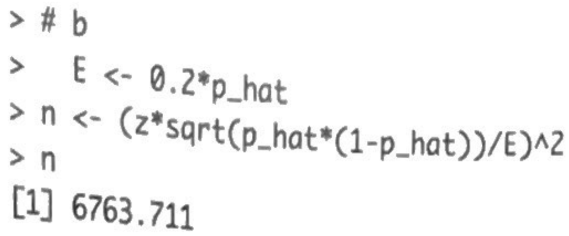
5. We are given $p = 0.007$ incidence of skin cancer over 5 years for American women, and asked to calculate sample sizes based on the hypothesis that $\hat{p} = 0.014$.

a. Using $\epsilon = \frac{\hat{p}}{2} = \frac{0.014}{2}$ and $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, we can make obtain and make use of the formula $n \geq \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{p}(1-\hat{p})}}{\epsilon}\right)^2 = \left(\frac{1.96 \cdot \sqrt{0.014(1-0.014)}}{0.007}\right)^2$. The output of the calculation in R is included below. Because we cannot recommend using fractions of subjects, we round up and state that the minimum sample size for a study with the qualities that the researcher desires should be $n = 1083$. ✓



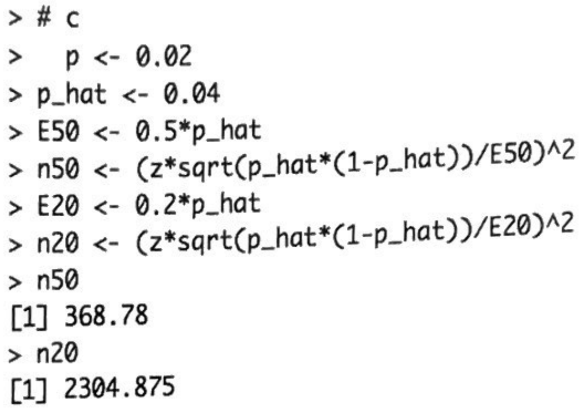
```
> # Problem 5
> p <- 0.007
> p_hat <- 0.014
> # a
>    E <- 0.5*p_hat
> z <- qnorm(0.975)
> n <- (z*sqrt(p_hat*(1-p_hat))/E)^2
> n
[1] 1082.194
```

b. The calculation is exactly the same, only $\epsilon = \frac{\hat{p}}{5} = \frac{0.014}{5} = 0.0028$. The output of the R calculation is shown below. For this case, we recommend a sample size of $n = 6764$. ✓



```
> # b
>   E <- 0.2*p_hat
> n <- (z*sqrt(p_hat*(1-p_hat))/E)^2
> n
[1] 6763.711
```

c. The logic used to solve this problem is exactly the same as in parts (a) and (b). The only difference is that $p = 0.02$ is the incidence of skin cancer over 5 years for American women in our target population, and we are asked to calculate sample sizes based on the hypothesis that $\hat{p} = 0.04$. In the case that we desire a 95% error bound that is about 50% of the true value, we recommend a sample size of $n = 369$, and in the case that we desire a 95% error bound that is about 20% of the true value, we recommend a sample size of $n = 2305$.

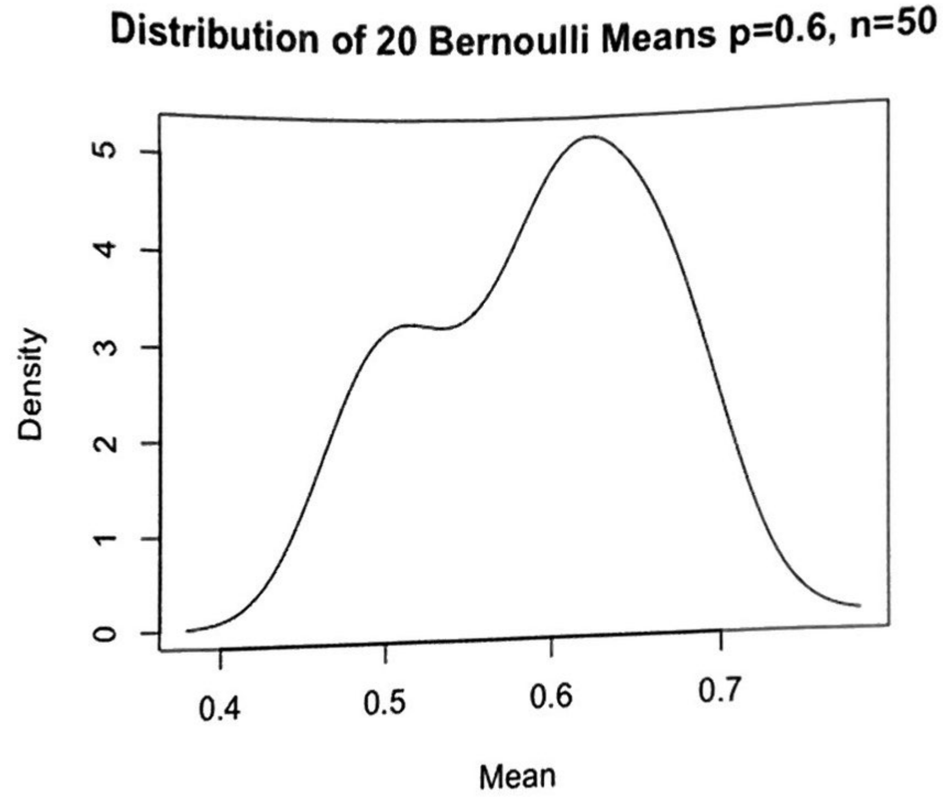


```
> # c
>   p <- 0.02
> p_hat <- 0.04
> E50 <- 0.5*p_hat
> n50 <- (z*sqrt(p_hat*(1-p_hat))/E50)^2
> E20 <- 0.2*p_hat
> n20 <- (z*sqrt(p_hat*(1-p_hat))/E20)^2
> n50
[1] 368.78
> n20
[1] 2304.875
```

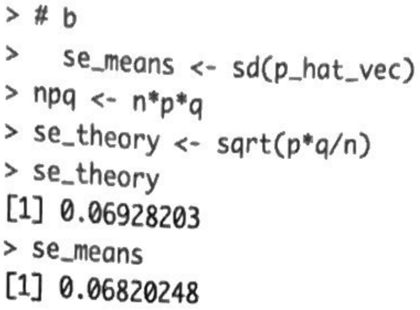
6.

a. The R input and output for this part of the problem is shown below. These means do not look obviously normally distributed. They do look close, however, and when we note that $npq = 12 > 5$, we note that the Central Limit Theorem should apply, and that as we conduct more trials, the density plot shown below will approach normal.

```
> # Problem 6
>   # a
>   p <- 0.6
> q <- 1-p
> n <- 50
> p_hat_vec <- vector()
> for (i in 1:20){
+     vec <- c(rbinom(n, 1, p))
+     p_hat_vec[i] <- (sum(vec == 1)/length(vec))
+   }
> plot(density(p_hat_vec), main="Distribution of 20 Bernoulli Means p=0.6, n=50",
+         xlab="Mean", ylab="Density")
```

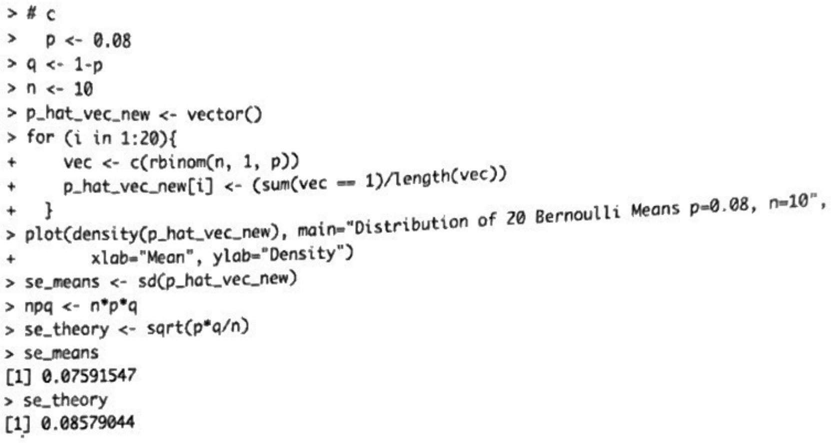


b. The R output for sample standard deviation is shown below. We can compare this to the theoretical value of $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(0.4)}{50}} = 0.0693$, and see that the two numbers are very close.

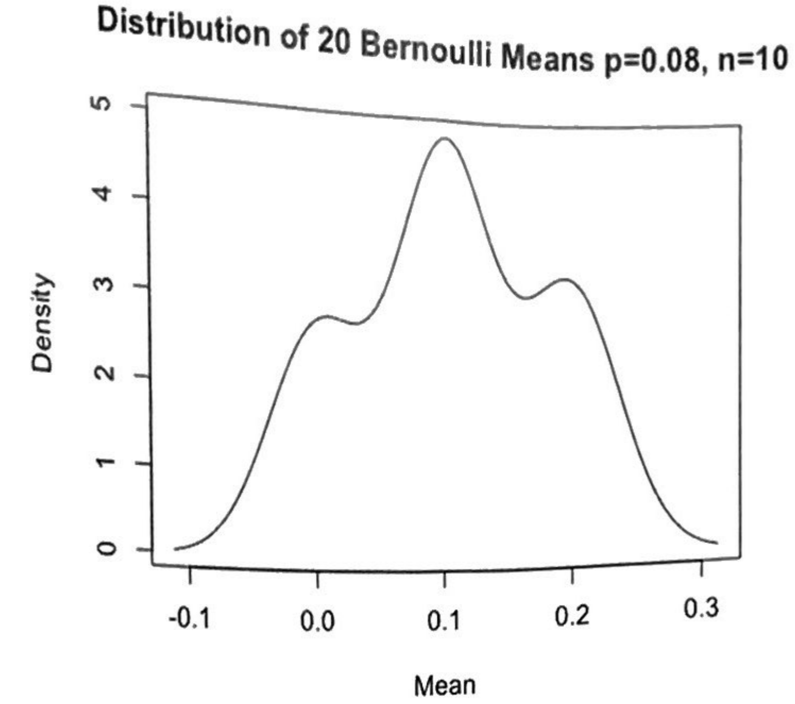


```
> # b
>   se_means <- sd(p_hat_vec)
> npq <- n*p*q
> se_theory <- sqrt(p*q/n)
> se_theory
[1] 0.06928203
> se_means
[1] 0.06820248
```

c. The same process was repeated using $n = 10$ and $p = 0.08$. Importantly, $npq = 0.736 < 5$, which tells us that the Central Limit Theorem does not apply. And when we inspect the density plot visually, we can state confidently that the data is farther from normal than was the data generated in the previous parts of the problem. The sample standard deviations are still close to the theoretical values as indicated below.

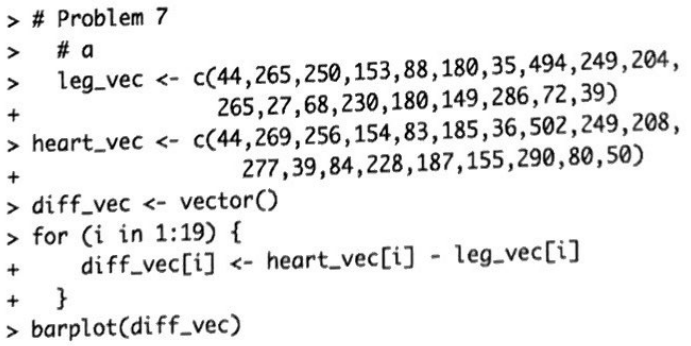


```
> # c
>   p <- 0.08
> q <- 1-p
> n <- 10
> p_hat_vec_new <- vector()
> for (i in 1:20){
+     vec <- c(rbinom(n, 1, p))
+     p_hat_vec_new[i] <- (sum(vec == 1)/length(vec))
+   }
> plot(density(p_hat_vec_new), main="Distribution of 20 Bernoulli Means p=0.08, n=10",
+        xlab="Mean", ylab="Density")
> se_means <- sd(p_hat_vec_new)
> npq <- n*p*q
> se_theory <- sqrt(p*q/n)
> se_means
[1] 0.07591547
> se_theory
[1] 0.08579044
```

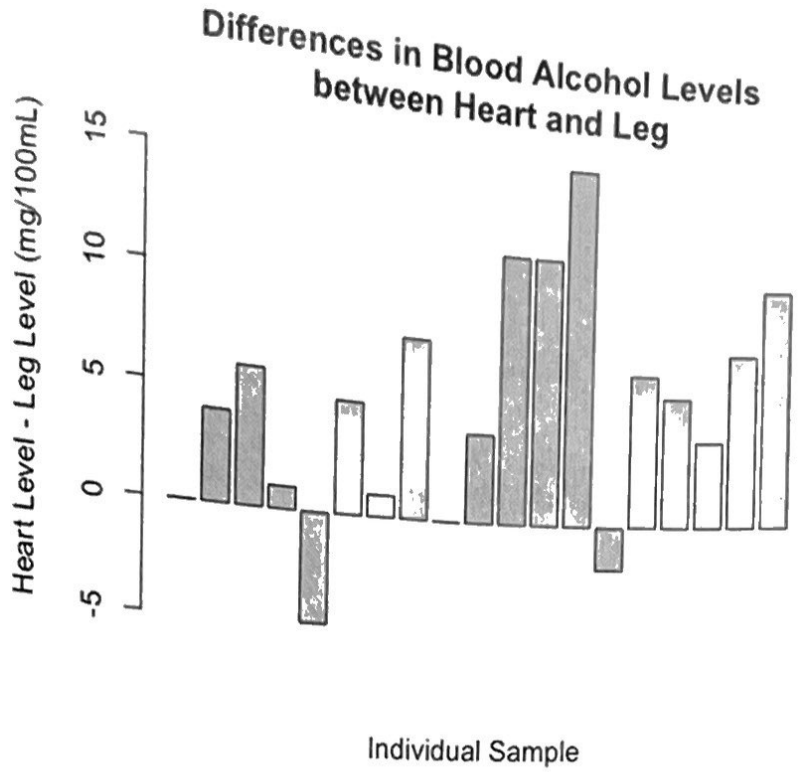


7. Note that the sample with missing data was excluded from all analyses.

   a. For the most part, the heart blood alcohol level seems to have been greater than the leg blood alcohol level. This clearly was not the case for every sample, as shown in the bar plot below.

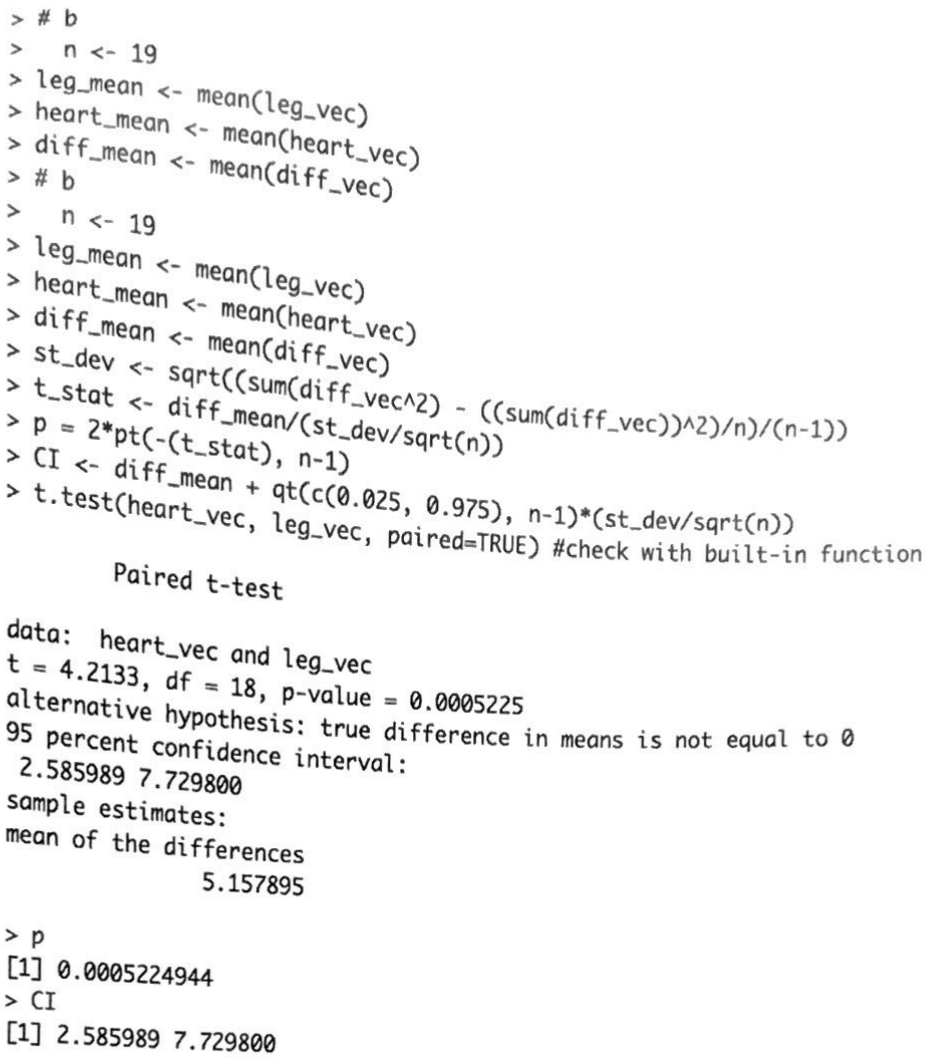


```
> # Problem 7
>   # a
>   leg_vec <- c(44,265,250,153,88,180,35,494,249,204,
+                265,27,68,230,180,149,286,72,39)
> heart_vec <- c(44,269,256,154,83,185,36,502,249,208,
+                277,39,84,228,187,155,290,80,50)
> diff_vec <- vector()
> for (i in 1:19) {
+     diff_vec[i] <- heart_vec[i] - leg_vec[i]
+   }
> barplot(diff_vec)
```



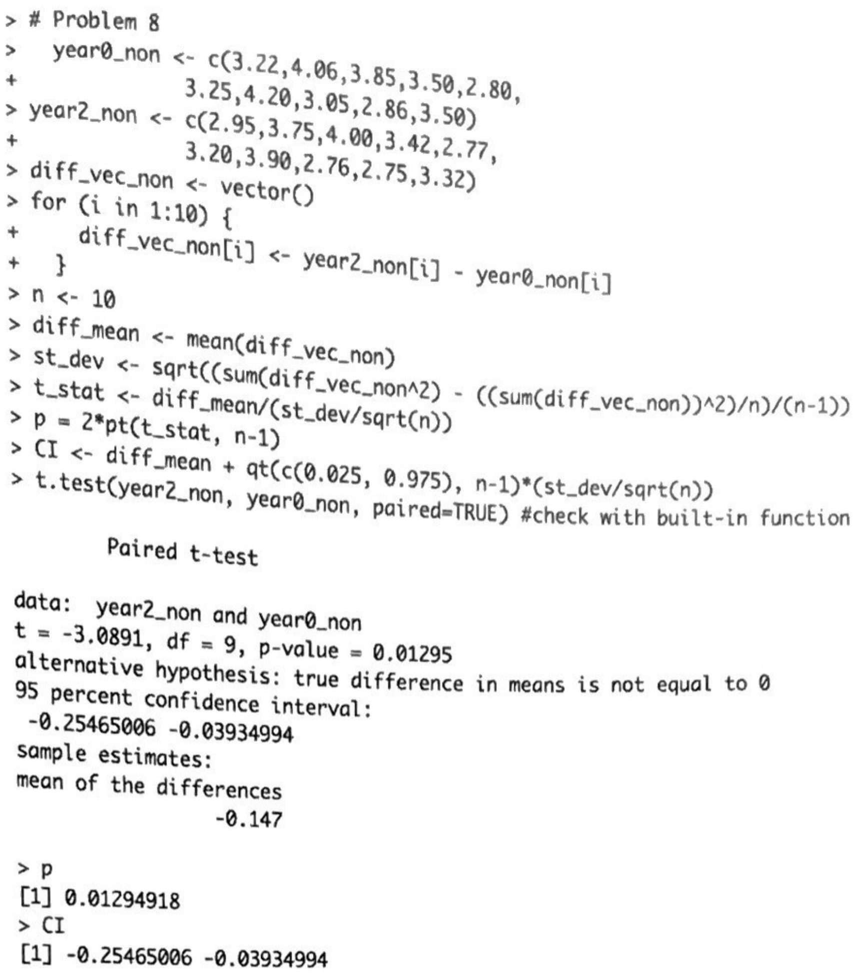
b. The R results shown below are the results of a paired t-test, with a p-value of 0.0005225 and confidence interval (2.5860, 7.7300). The sample mean of the differences is 5.1579 mg/100mL. The results of this test indicate that at the $\alpha = 0.05$ level, the differences between the leg samples and heart samples are highly statistically significant. Our point estimation for the mean difference between leg and heart samples is 5.1579 mg/100mL, and we are 95% confident that the true mean lies in the interval (2.5860, 7.7300).

✓



```
> # b
>   n <- 19
> leg_mean <- mean(leg_vec)
> heart_mean <- mean(heart_vec)
> diff_mean <- mean(diff_vec)
> # b
>   n <- 19
> leg_mean <- mean(leg_vec)
> heart_mean <- mean(heart_vec)
> diff_mean <- mean(diff_vec)
> st_dev <- sqrt((sum(diff_vec^2) - ((sum(diff_vec))^2)/n)/(n-1))
> t_stat <- diff_mean/(st_dev/sqrt(n))
> p = 2*pt(-(t_stat), n-1)
> CI <- diff_mean + qt(c(0.025, 0.975), n-1)*(st_dev/sqrt(n))
> t.test(heart_vec, leg_vec, paired=TRUE) #check with built-in function

	Paired t-test

data:  heart_vec and leg_vec
t = 4.2133, df = 18, p-value = 0.0005225
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.585989 7.729800
sample estimates:
mean of the differences
               5.157895

> p
[1] 0.0005224944
> CI
[1] 2.585989 7.729800
```

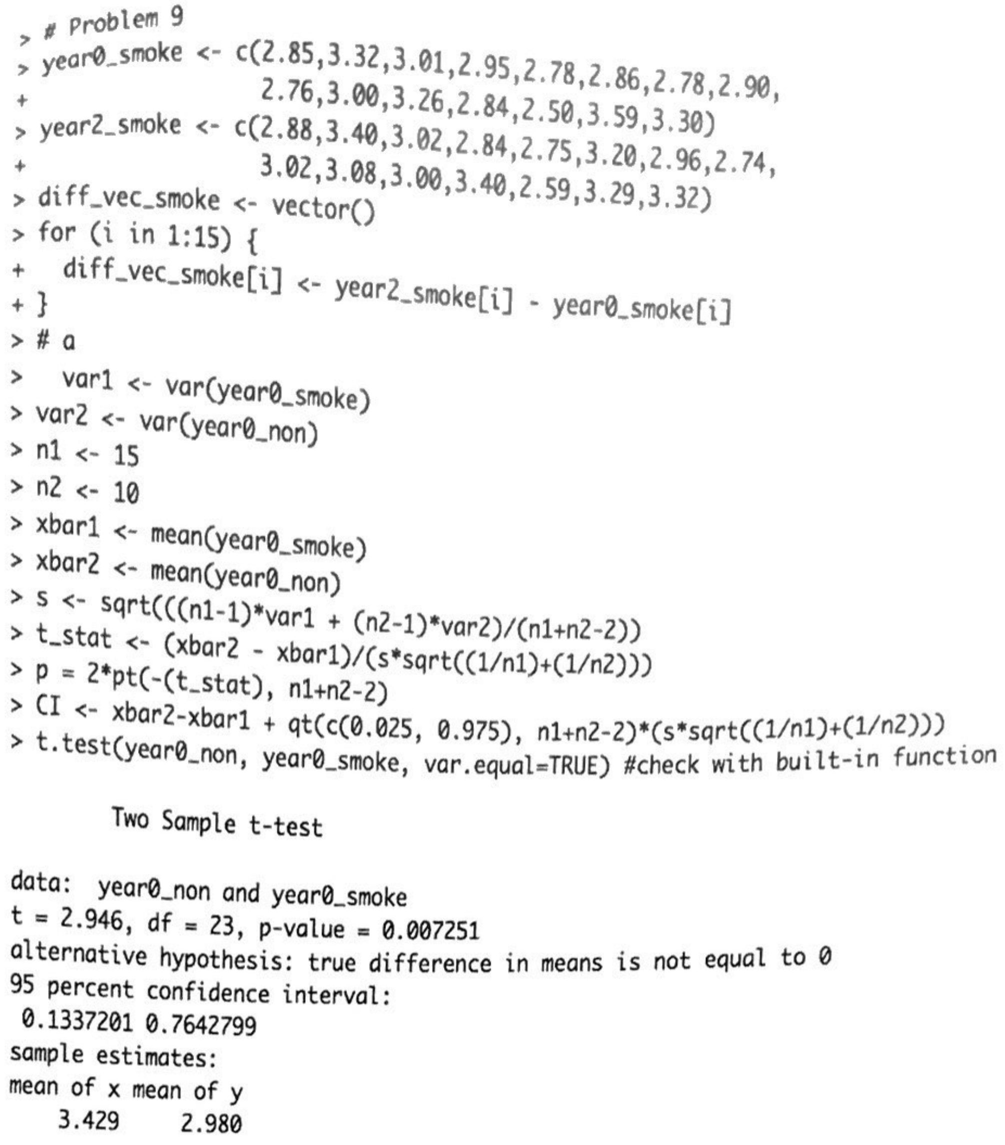
8. FEV data from the same individual compared between two points in time warrants a paired t-test. Specifically, we can begin with the hypotheses H$_0$: $\mu_0 = \mu_1$ and H$_1$: $\mu_0 \neq \mu_1$. A paired t-test was used to determine at the $\alpha = 0.05$ level that there was a statistically significant difference in FEV levels over the 2 years. The p-value for this t-test was p = 0.0129, and therefore we can reject the null hypothesis. Additionally, we can be 95% confident that the true mean of the FEV level differences between year 0 and year 2 lies in the interval (-0.2547, -0.0393). Our point estimate for the mean difference is -0.147. In general, this will mean that the FEV level at year 0 is greater than the FEV level at year 2.



```
> # Problem 8
>   year0_non <- c(3.22,4.06,3.85,3.50,2.80,
+                  3.25,4.20,3.05,2.86,3.50)
> year2_non <- c(2.95,3.75,4.00,3.42,2.77,
+                  3.20,3.90,2.76,2.75,3.32)
> diff_vec_non <- vector()
> for (i in 1:10) {
+     diff_vec_non[i] <- year2_non[i] - year0_non[i]
+   }
> n <- 10
> diff_mean <- mean(diff_vec_non)
> st_dev <- sqrt((sum(diff_vec_non^2) - ((sum(diff_vec_non))^2)/n)/(n-1))
> t_stat <- diff_mean/(st_dev/sqrt(n))
> p = 2*pt(t_stat, n-1)
> CI <- diff_mean + qt(c(0.025, 0.975), n-1)*(st_dev/sqrt(n))
> t.test(year2_non, year0_non, paired=TRUE) #check with built-in function

        Paired t-test

data:  year2_non and year0_non
t = -3.0891, df = 9, p-value = 0.01295
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.25465006 -0.03934994
sample estimates:
mean of the differences
                 -0.147

> p
[1] 0.01294918
> CI
[1] -0.25465006 -0.03934994
```

9.

a. We wish to compare the smokers and non-smokers at baseline. Thus, the appropriate null hypothesis is H$_0$: $\mu_{non} = \mu_{smokers}$ and H$_1$: $\mu_{non} \neq \mu_{smokers}$. We are looking at two independent groups, with sample sizes $n_{non} = 10$ and $n_{smokers} = 15$ respectively. A two-sample t-test is warranted, and the calculations are shown below. The results are a p-value 0f 0.007251, which allows us to reject the null hypothesis at the $\alpha = 0.05$ level. This p-value gives us the probability of obtaining as or more extreme as the one we did under the null hypothesis. The 95% confidence interval for the difference in means between the two samples is included in the calculation below, as are point estimations of the means for each group, with x representing the non-smokers and y representing the smokers.

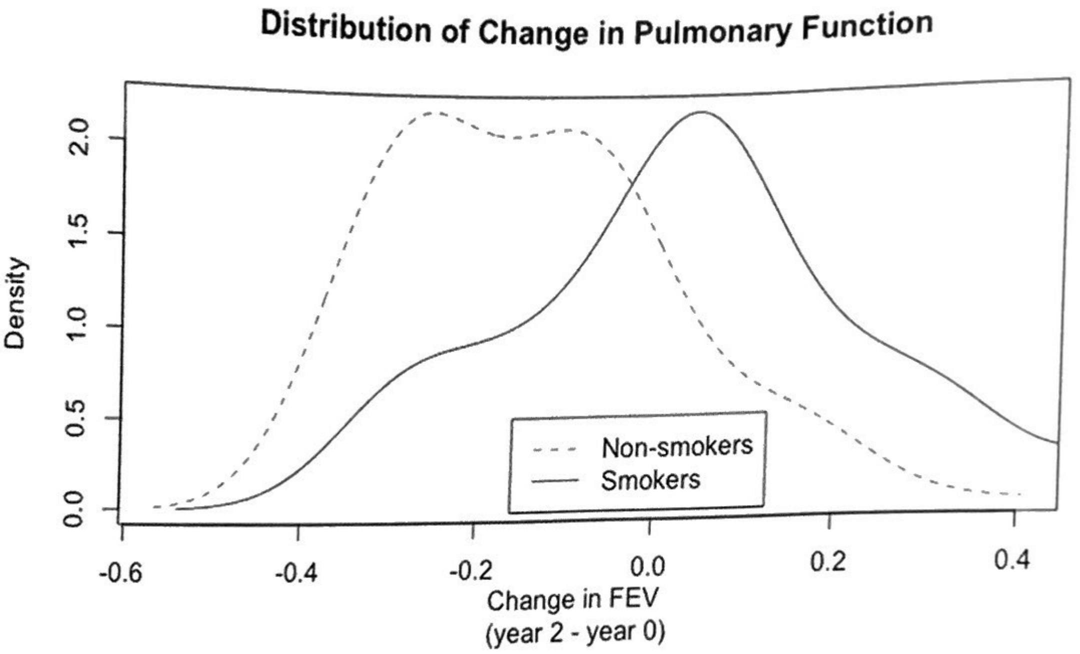


```
> # Problem 9
> year0_smoke <- c(2.85,3.32,3.01,2.95,2.78,2.86,2.78,2.90,
+                   2.76,3.00,3.26,2.84,2.50,3.59,3.30)
> year2_smoke <- c(2.88,3.40,3.02,2.84,2.75,3.20,2.96,2.74,
+                   3.02,3.08,3.00,3.40,2.59,3.29,3.32)
> diff_vec_smoke <- vector()
> for (i in 1:15) {
+   diff_vec_smoke[i] <- year2_smoke[i] - year0_smoke[i]
+ }
> # a
>   var1 <- var(year0_smoke)
> var2 <- var(year0_non)
> n1 <- 15
> n2 <- 10
> xbar1 <- mean(year0_smoke)
> xbar2 <- mean(year0_non)
> s <- sqrt(((n1-1)*var1 + (n2-1)*var2)/(n1+n2-2))
> t_stat <- (xbar2 - xbar1)/(s*sqrt((1/n1)+(1/n2)))
> p = 2*pt(-(t_stat), n1+n2-2)
> CI <- xbar2-xbar1 + qt(c(0.025, 0.975), n1+n2-2)*(s*sqrt((1/n1)+(1/n2)))
> t.test(year0_non, year0_smoke, var.equal=TRUE) #check with built-in function

        Two Sample t-test

data:  year0_non and year0_smoke
t = 2.946, df = 23, p-value = 0.007251
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1337201 0.7642799
sample estimates:
mean of x mean of y
    3.429     2.980
```

b. Another two sample t-test can be used to determine whether there is a significant difference between the changes in FEV measurements from year 0 to year 2 within each group. Here, the hypotheses can be stated H$_0$: $\mu_{diff,non} = \mu_{diff,smokers}$ and H$_1$: $\mu_{diff,non} \neq \mu_{diff,smokers}$. The R calculations and results are included below. The p-value is 0.02188, and so at the $\alpha = 0.05$ level we can reject the null hypothesis. A p-value of 0.02188 gives us the probability that we would get a result as or more extreme as we did under the null hypothesis. A confidence interval for the difference between the sample means is included, as are point estimates for the mean FEV difference between year 0 and year 2 within each group. Interestingly, the results imply that the non-smokers saw an average decrease in FEV of 0.1470 over two years, whereas the smokers saw an average increase in FEV of 0.0527 over two years.

```
> # b
>   xbar1 <- mean(diff_vec_smoke)
> xbar2 <- mean(diff_vec_non)
> var1 <- var(diff_vec_smoke)
> var2 <- var(diff_vec_non)
> s <- sqrt(((n1-1)*var1 + (n2-1)*var2)/(n1+n2-2))
> t_stat <- (xbar1 - xbar2)/(s*sqrt((1/n1)+(1/n2)))
> p = 2*pt(-(t_stat), n1+n2-2)
> CI <- xbar2-xbar1 + qt(c(0.025, 0.975), n1+n2-2)*(s*sqrt((1/n1)+(1/n2)))
> t.test(diff_vec_non, diff_vec_smoke, var.equal=TRUE)

        Two Sample t-test

data:  diff_vec_non and diff_vec_smoke
t = -2.4589, df = 23, p-value = 0.02188
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.36764246 -0.03169087
sample estimates:
  mean of x   mean of y
-0.14700000  0.05266667
```

c.



d. Transforming the data will allow the underlying distribution to become more normal, and thus in general it will be preferable to report p-values for transformed data. In this case, as shown below, the difference in p-value is almost negligible (0.0223 in the transformed calculation versus 0.02188 in the original calculation), and does not change the fact that we will reject the null hypothesis at the $\alpha = 0.05$ level. In the interest of accuracy, I would still report the log-transformed p-value.

```
> # d
>   log_non0 <- log(year0_non)
> log_non2 <- log(year2_non)
> log_smoke0 <- log(year0_smoke)
> log_smoke2 <- log(year2_smoke)
> diff_log_non <- vector()
> diff_log_smoke <- vector()
> for (i in 1:10) {
+     diff_log_non[i] <- log_non2[i] - log_non0[i]
+   }
> for (i in 1:15) {
+     diff_log_smoke[i] <- log_smoke2[i] - log_smoke0[i]
+   }
> xbar1 <- mean(diff_log_smoke)
> xbar2 <- mean(diff_log_non)
> var1 <- var(diff_log_smoke)
> var2 <- var(diff_log_non)
> s <- sqrt(((n1-1)*var1 + (n2-1)*var2)/(n1+n2-2))
> t_stat <- (xbar1 - xbar2)/(s*sqrt((1/n1)+(1/n2)))
> p = 2*pt(-(t_stat), n1+n2-2)
> CI <- xbar2-xbar1 + qt(c(0.025, 0.975), n1+n2-2)*(s*sqrt((1/n1)+(1/n2)))
> t.test(diff_log_non, diff_log_smoke, var.equal=TRUE)

        Two Sample t-test

data:  diff_log_non and diff_log_smoke
t = -2.4504, df = 23, p-value = 0.0223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.115661414 -0.009769908
sample estimates:
  mean of x   mean of y 
-0.04442186  0.01829380 
```