David Levy
11/28/17

Biostatistics 200A
Problem Set 5

1.

We will use a 2×2 contingency table to evaluate the homogeneity of binomial proportions for the withdrawal rates. The columns in the tables that follow represent "successes" and "failures," where successes correspond with individual withdrawals. The rows represent the two distinct dentifrice groups.

Observed table:

	S	F	T
A	163	260	423
D	119	289	408
T	282	549	831

So,

$$O_{11} = 163$$
$$O_{12} = 260$$
$$O_{21} = 119$$
$$O_{22} = 289$$

Expected table:

$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$, where $n_1\hat{p}_1 = 163, n_2\hat{p}_2 = 119$, so that $\hat{p} = \frac{163 + 119}{423+408} = \frac{282}{831} = 0.339$.
Thus,

$$E_{11} = \frac{423 \times 282}{831} \approx 143.5$$
$$E_{12} = \frac{423 \times 549}{831} \approx 279.5$$
$$E_{21} = \frac{408 \times 282}{831} \approx 138.5$$
$$E_{22} = \frac{408 \times 549}{831} \approx 269.5$$

	S	F	T
A	143.5	279.5	423
D	138.5	269.5	408
T	282	549	831

In essence, we are testing the hypothesis $H_0: p_1 = p_2 = p$ vs. $H_1: p_1 \neq p_2$, so a Pearson's χ^2 statistic is the most appropriate tool at our disposal. We will use the equation for a continuity-corrected statistic:

$$\begin{aligned}\chi^2 &= \frac{(|O_{11}-E_{11}|-0.5)^2}{E_{11}} + \frac{(|O_{12}-E_{12}|-0.5)^2}{E_{12}} + \frac{(|O_{21}-E_{21}|-0.5)^2}{E_{21}} + \frac{(|O_{22}-E_{22}|-0.5)^2}{E_{22}} \\ &= \frac{(|163-143.5|-0.5)^2}{143.5} + \frac{(|260-279.5|-0.5)^2}{279.5} + \frac{(|119-138.5|-0.5)^2}{138.5} + \frac{(|289-269.5|-0.5)^2}{269.5} \\ &= 7.7156.\end{aligned}$$

Now the approximate two-sided p-value should be given by the area to the right of χ^2 under a ${\chi_1}^2$ distribution. The R output for a manual Pearson's χ^2 is shown below, and gives a p-value of $p = 0.005472$.

```
> chi_stat <- ((((abs(o_11-e_11)-0.5)^2)/e_11) + (((abs(o_12-e_12)-0.5)^2)/e_12) + (((abs(o_21-e_21)-0.5)^2)/e_21)
+ (((abs(o_22-e_22)-0.5)^2)/e_22))
> p_val <- pchisq(chi_stat, 1, lower.tail=FALSE)
> chi_stat
[1] 7.716545
> p_val
[1] 0.005471702
```

The R output for a pre-coded Pearson's χ^2 test confirms the calculations performed above. The output is shown below:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 7.7165, df = 1, p-value = 0.005472
```

It may, however, be the case that the continuity correction is inappropriate in this circumstance, because we are not using a continuous model for these data, and the original data present are discrete counts. Thus, we could observe the results of a non-corrected test shown below:

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 8.129, df = 1, p-value = 0.004356
```

Alternatively, we can assume that the normal approximation to the binomial distribution is valid because the samples are sufficiently large. Therefore, it is reasonable to base our significance test on the difference between the sample proportions. Thus, we can define our hypotheses: $H_0: \hat{p}_1 - \hat{p}_2 = 0$; $H_1: \hat{p}_1 - \hat{p}_2 \neq 0$, and calculate a z-statistic with continuity correction under H_0:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$p_1 = \frac{163}{423} = 0.3853,$$

$$p_2 = \frac{119}{408} = 0.2917,$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{163 + 119}{831} = 0.3394,$$

and

$$\hat{q} = 1 - \hat{p} = 0.6606$$

These calculations are confirmed by the following R output:

```
> p1_hat <- o_11/n1
> p2_hat <- o_21/n2
> p1_hat
[1] 0.3853428
> p2_hat
[1] 0.2916667
```

```
> p_hat <- (o_11 + o_21)/gr_tot
> q_hat <- 1-p_hat
> p_hat
[1] 0.3393502
> q_hat
[1] 0.6606498
```

Thus, our z-statistic is calculated as follows:

$$Z_S = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|0.3853 - 0.2917| - \left(\frac{1}{2(423)} + \frac{1}{2(408)}\right)}{\sqrt{(0.3394)(0.6606)\left(\frac{1}{423} + \frac{1}{408}\right)}} = 2.7779$$

This calculation is confirmed by the R output below:

```
> z_stat <- (abs(p1_hat-p2_hat)-((1/(2*n1))+(1/(2*n2))))/sqrt(p_hat*q_hat*((1/n1)+(1/n2)))
> z_stat
[1] 2.777867
```

We can see that the p-value is calculated by finding $p = 2P(X > 2.779) = 0.005472$, which matches the p-value found using Pearson's χ^2 test. In either case, we safely reject the null hypothesis at the 95% confidence level, and state that there is strong evidence to suggest that the proportions $\hat{p}_1$ and $\hat{p}_2$ are significantly different, and that it is very likely that dentifrice A is more disliked than dentifrice D, if that is in fact what is causing subjects to withdraw from the study. Good interpretation

A point estimation for the difference withdrawal rates can be calculated as follows:

$$\hat{p}_1 - \hat{p}_2 = 0.3853 - 0.2917 = 0.09368$$

And confirmed by the R output below:

```
> diff_est <- p1_hat-p2_hat
> diff_est
[1] 0.09367612
```

A 95% confidence interval is constructed as follows:

$$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm (Z_{1-\frac{\alpha}{2}})\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (0.0293, 0.1581)$$

```
> p_hat <- (o_11 + o_21)/gr_tot
> q_hat <- 1-p_hat
> p_hat
[1] 0.3393502
> q_hat
[1] 0.6606498
```

Thus, our z-statistic is calculated as follows:

$$Z_S = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|0.3853 - 0.2917| - \left(\frac{1}{2(423)} + \frac{1}{2(408)}\right)}{\sqrt{(0.3394)(0.6606)\left(\frac{1}{423} + \frac{1}{408}\right)}} = 2.7779$$

This calculation is confirmed by the R output below:

```
> z_stat <- (abs(p1_hat-p2_hat)-((1/(2*n1))+(1/(2*n2))))/sqrt(p_hat*q_hat*((1/n1)+(1/n2)))
> z_stat
[1] 2.777867
```

We can see that the p-value is calculated by finding $p = 2P(X > 2.779) = 0.005472$, which matches the p-value found using Pearson's χ^2 test. In either case, we safely reject the null hypothesis at the 95% confidence level, and state that there is strong evidence to suggest that the proportions $\hat{p}_1$ and $\hat{p}_2$ are significantly different, and that it is very likely that dentifrice A is more disliked than dentifrice D, if that is in fact what is causing subjects to withdraw from the study. Good interpretation

A point estimation for the difference withdrawal rates can be calculated as follows:

$$\hat{p}_1 - \hat{p}_2 = 0.3853 - 0.2917 = 0.09368$$

And confirmed by the R output below:

```
> diff_est <- p1_hat-p2_hat
> diff_est
[1] 0.09367612
```

A 95% confidence interval is constructed as follows:

$$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm (Z_{1-\frac{\alpha}{2}}) \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (0.0293, 0.1581)$$

And confirmed by the R output below:

```
> CI <- c((p1_hat-p2_hat) - qnorm(0.975)*sqrt(p_hat*q_hat*((1/n1)+(1/n2))),
+         (p1_hat-p2_hat) + qnorm(0.975)*sqrt(p_hat*q_hat*((1/n1)+(1/n2))))
> CI
[1] 0.02928025 0.15807200
```

2.

a.

The expected value for table 1 are:

$$E_{11} = \frac{22 \times 12}{50} = 5.28$$
$$E_{12} = \frac{22 \times 38}{50} = 16.72$$
$$E_{21} = \frac{28 \times 12}{50} = 6.72$$
$$E_{22} = \frac{28 \times 38}{50} = 21.28$$

Since all expected values are greater than 5, it should be reasonable for the authors to rely on the accuracy of the Pearson's χ^2. The only real difference here is that the authors used a Yates continuity correction. However, there are no continuity issues to correct for here, so the more appropriate χ^2 calculation is shown in the R output below. Because the χ^2 statistic is larger in this instance, and the p-value smaller, there is no real change in the qualitative result of rejecting the null hypothesis that the binomial proportions of the two groups are equal. There is, at the $\alpha = 0.01$ significance level, an association between having worked on the days in question and contracting the illness. Thus, a Fisher's exact test, while possibly informative about additional details, is not necessary.

```
> chisq.test(contingency_table, correct=FALSE)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 9.914, df = 1, p-value = 0.00164
```

b.

I am not entirely certain why the authors would have used a student's t-test with 40 degrees of freedom here. Even if the data was going to be analyzed in such a way, a normal approximation would have sufficed. However, a much more appropriate analysis would have been generating contingency tables for all the foods that the authors wished to analyze, and calculating individual χ^2 or Fisher's exact test p-values for each, as deemed appropriate by examining the expected values for each contingency table.

c.

We can generate a contingency table for only the data pertaining to salad consumption:

Observed table:

	Ill	Well	T
Salad	25	8	33
No salad	3	6	9
T	28	14	42

Expected table:

	Ill	Well	T
Salad	22	11	33
No salad	6	3	9
T	28	14	42

Because 25% of the expected valued in the expected table are less than 5, the most appropriate way to proceed is by performing a Fisher's exact test. We will rely on R to perform this calculation:

```
        Fisher's Exact Test for Count Data

data:  ctab
p-value = 0.04066
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9978847 45.4474943
sample estimates:
odds ratio
  5.930094
```

With a p-value of 0.04066, we can state that the results are significant at the $\alpha = 0.05$ significance level, but we cannot agree with the authors, who state that the association between eating salad and contracting the illness is significant at the $\alpha = 0.01$ level. At that level, we would fail to reject the null hypothesis that there is no difference in the binomial proportions.

3.

The following contingency tables are both accurate representations of the same data. Corresponding cells between the two tables contain identities that will be used extensively throughout the proof below.

	Success	**Failure**	
X	a	b	n_1
Y	c	d	n_2
	S	F	N

	Success	Failure	
X	$n_1\hat{p}_1$	$n_1 - n_1\hat{p}_1$	n_1
Y	$n_2\hat{p}_2$	$n_2 - n_2\hat{p}_2$	n_2
	$n_1\hat{p}_1$	$n_2\hat{p}_2$	$n_1 + n_2$

We can begin by observing that a Pearson's χ^2 statistic has the general formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

We can put the χ^2 statistic in terms of quantities from the first table as such:

$$(O-E) = a - \frac{n_1 S}{N} = a - \frac{(a+b)(a+c)}{N} = \frac{aN - (a+b)(a+c)}{N}$$
$$= \frac{a(a+b+c+d) - (a+b)(a+c)}{N} = \frac{ad-bc}{N}$$

So,

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(ad-bc)^2}{N^2}\sum \frac{1}{E}$$
$$= \left(\frac{(ad-bc)^2}{N^2}\right)\left(\frac{1}{n_1(S/N)} + \frac{1}{n_2(S/N)} + \frac{1}{n_1(F/N)} + \frac{1}{n_2(F/N)}\right)$$
$$= \frac{N(ad-bc)^2}{n_1 n_2 SF}$$

Now, we can begin to show that if we square the $\frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)}$ statistic, we obtain the χ^2 statistic. First, we observe that:

where

$$\frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

and

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{a + c}{N},$$

Thus,

$$\hat{q} = 1 - \hat{p} = 1 - \frac{a + c}{N} = \frac{N}{N} - \frac{a + c}{N} = \frac{b + d}{N}$$

$$\left(\frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)}\right)^2 = \left(\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}\left(\frac{n_1 + n_2}{n_1 n_2}\right)} = \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}(n_1 + n_2)}$$

$$= \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{a + c}{N}\right)\left(\frac{b + d}{N}\right)(N)} = \frac{\left(\frac{(n_1 n_2)^2}{n_1 n_2}\right)(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{a + c}{N}\right)\left(\frac{b + d}{N}\right)(N)} = \frac{(n_1 n_2)^2(\hat{p}_1 - \hat{p}_2)^2}{n_1 n_2\left(\frac{SF}{N}\right)}$$

$$= \frac{N(n_1 n_2)^2(\hat{p}_1 - \hat{p}_2)^2}{n_1 n_2 SF} = \frac{N(n_1 n_2(\hat{p}_1 - \hat{p}_2))^2}{n_1 n_2 SF}$$

Importantly, we can also observe that:

$$\begin{aligned} ad - bc = {} & (n_1\hat{p}_1)(n_2 - n_2\hat{p}_2) - (n_2\hat{p}_2)(n_1 - n_1\hat{p}_1) \\ & = (n_1 n_2\hat{p}_1 - n_1 n_2\hat{p}_1\hat{p}_2) - (n_1 n_2\hat{p}_2 - n_1 n_2\hat{p}_1\hat{p}_2) = n_1 n_2\hat{p}_1 - n_1 n_2\hat{p}_2 \\ & = n_1 n_2(\hat{p}_1 - \hat{p}_2) \end{aligned}$$

And therefore, we can conclude our proof by stating:

$$\left(\frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)}\right)^2 = \frac{N(n_1 n_2(\hat{p}_1 - \hat{p}_2))^2}{n_1 n_2 SF} = \frac{N(ad - bc)^2}{n_1 n_2 SF} = \chi^2$$

4.

Observed table:

	D	A	T
Tx	2	13	15
C	4	15	19
T	6	28	34

a.

Enumeration of all tables w/ same margins:

	D	A	T
Tx	0	15	15
C	6	13	19
T	6	28	34

$$P(0,15,6,13) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,0!\,15!\,6!\,13!} = 0.02017$$

	D	A	T
Tx	1	14	15
C	5	14	19
T	6	28	34

$$P(1,14,5,14) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,1!\,14!\,5!\,14!} = 0.12970$$

	D	A	T
Tx	2	13	15
C	4	15	19
T	6	28	34

$$P(2,13,4,15) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,2!\,13!\,4!\,15!} = 0.30261$$

	D	A	T
Tx	3	12	15
C	3	16	19
T	6	28	34

$$P(3,12,3,16) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,3!\,12!\,3!\,16!} = 0.32783$$

	D	A	T
Tx	4	11	15
C	2	17	19
T	6	28	34

$$P(4,11,2,17) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,4!\,11!\,2!\,17!} = 0.17356$$

	D	A	T

Tx	5	10	15
C	1	18	19
T	6	28	34

$$P(5,10,1,18) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,5!\,10!\,1!\,18!} = 0.04242$$

	D	A	T
Tx	6	9	15
C	0	19	19
T	6	28	34

$$P(6,9,0,19) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} = \frac{15!\,19!\,6!\,28!}{34!\,6!\,9!\,0!\,19!} = 0.00372$$

The probability of observing a table as or more extreme than the one we observed is given either by obtaining a one-sided p-values from the tables in green and multiplying by two, or by summing all probabilities less than or equal to 0.30261. These methods give the following results, respectively:

```
> 0.02017+0.12970+0.30261
[1] 0.45248
> 0.45248*2
[1] 0.90496
```

so that $p = 0.90496$

```
> 0.02017+0.12970+0.30261+0.17356+0.04242+0.00372
[1] 0.67218
```

so that $p = 0.67218$

In either case, we fail to reject the null hypothesis that the mortality rates are significantly different between treatment and control groups. The calculation is confirmed by the following R output for Fisher's exact test:

```
        Fisher's Exact Test for Count Data

data:  ctab
p-value = 0.6722
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.04590206 4.89390008
sample estimates:
odds ratio
  0.586089
```

b.

For a normal approximation:

$$\hat{p}_1 = \frac{2}{15} = 0.1333$$

$$\hat{p}_2 = \frac{4}{19} = 0.2105$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{2+4}{34} = 0.1765$$

$$\hat{q} = 1 - \hat{p} = 0.8235$$

$$Z_S = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|0.1333 - 0.2105| - \left(\frac{1}{2(15)} + \frac{1}{2(19)}\right)}{\sqrt{(0.1765)(0.8235)\left(\frac{1}{15} + \frac{1}{19}\right)}} = 0.1332$$

The p-value would then be calculated:

```
> z_stat <- (abs(p1_hat-p2_hat)-((1/(2*n1))+(1/(2*n2))))/sqrt(p_hat*q_hat*((1/n1)+(1/n2)))
> p_val <- 2*pnorm(z_stat, lower.tail=FALSE)
> p_val
[1] 0.8940041
```

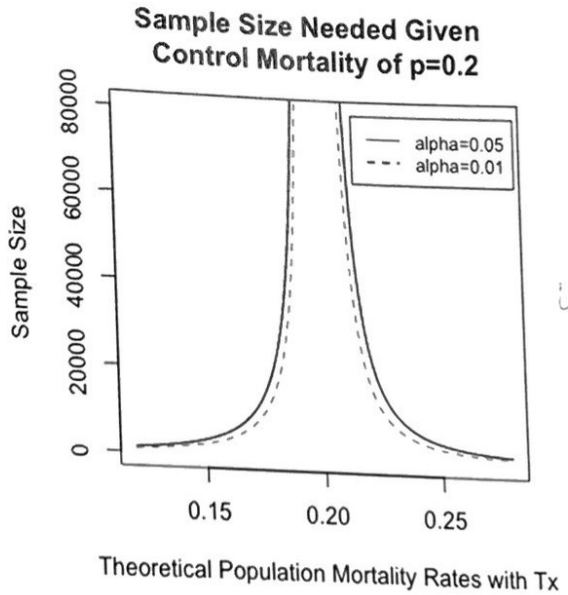
For a Pearson's χ^2 test, we see the following results:

```
> contingency_table <- matrix(c(2,13,4,15), ncol=2, byrow=TRUE)
> chisq.test(contingency_table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.017753, df = 1, p-value = 0.894

Warning message:
In chisq.test(contingency_table) :
  Chi-squared approximation may be incorrect
```

In any case, the p-value we obtain does not allow us to reject the null hypothesis of equivalent mortality rates between the treatment and control groups.

c.

```
> power_fun <- function(p1, p2, powr, alpha){
+   x <- (qnorm(1-(alpha/2))+qnorm(powr))^2
+   y <- (p1*(1-p1))+(p2*(1-p2))
+   z <- (p1-p2)^2
+   (x*y)/z
+ }
> power_vec5 <- vector()
> p2_vec <- seq(0.12, 0.28, by=0.001)
> for(i in 1:length(p2_vec)){
+   power_vec5[i] <- 2*power_fun(0.2, p2_vec[i], 0.9, 0.05)
+ }
> power_vec1 <- vector()
> for(i in 1:length(p2_vec)){
+   power_vec1[i] <- 2*power_fun(0.2, p2_vec[i], 0.9, 0.01)
+ }
> plot(power_vec1~p2_vec,type="l", col="blue",
+      main = "Sample Size Needed Given \n Control Mortality of p=0.2", ylim = c(0,80000),
+      xlab = "Theoretical Population Mortality Rates with Tx", ylab = "Sample Size")
> lines(power_vec5~p2_vec, type="l", lty=2, col="forestgreen")
> legend(0.215, 80000, c("alpha=0.05", "alpha=0.01"),
+        col = c("forestgreen", "blue"), lty = c(1, 2), cex = 0.8)
```

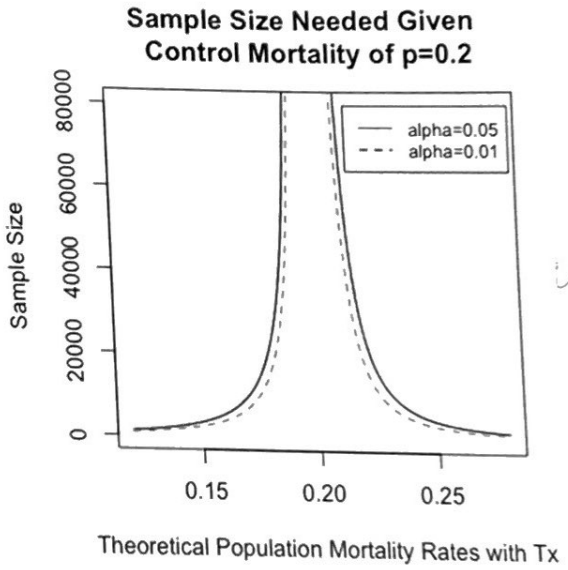


```
> contingency_table <- matrix(c(2,13,4,15), ncol=2, byrow=TRUE)
> chisq.test(contingency_table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.017753, df = 1, p-value = 0.894

Warning message:
In chisq.test(contingency_table) :
  Chi-squared approximation may be incorrect
```

In any case, the p-value we obtain does not allow us to reject the null hypothesis of equivalent mortality rates between the treatment and control groups.

c.

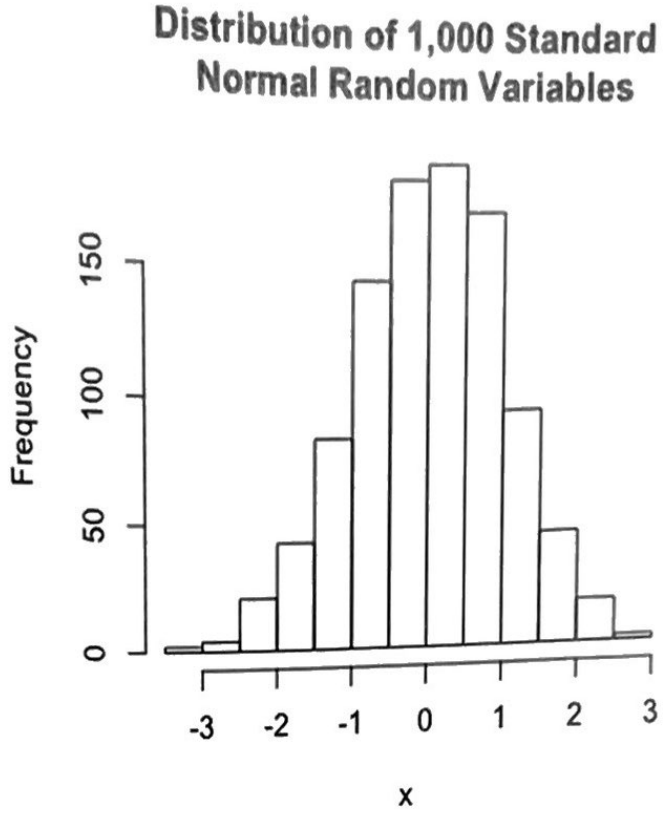
```
> power_fun <- function(p1, p2, powr, alpha){
+   x <- (qnorm(1-(alpha/2))+qnorm(powr))^2
+   y <- (p1*(1-p1))+(p2*(1-p2))
+   z <- (p1-p2)^2
+   (x*y)/z
+ }
> power_vec5 <- vector()
> p2_vec <- seq(0.12, 0.28, by=0.001)
> for(i in 1:length(p2_vec)){
+   power_vec5[i] <- 2*power_fun(0.2, p2_vec[i], 0.9, 0.05)
+ }
> power_vec1 <- vector()
> for(i in 1:length(p2_vec)){
+   power_vec1[i] <- 2*power_fun(0.2, p2_vec[i], 0.9, 0.01)
+ }
> plot(power_vec1~p2_vec,type="l", col="blue",
+      main = "Sample Size Needed Given \n Control Mortality of p=0.2", ylim = c(0,80000),
+      xlab = "Theoretical Population Mortality Rates with Tx", ylab = "Sample Size")
> lines(power_vec5~p2_vec, type="l", lty=2, col="forestgreen")
> legend(0.215, 80000, c("alpha=0.05", "alpha=0.01"),
+        col = c("forestgreen", "blue"), lty = c(1, 2), cex = 0.8)
```



5.

a.

```
x <- rnorm(1000, 0, 1)
hist(x, main="Distribution of 1,000 Standard \n Normal Random Variables")
```

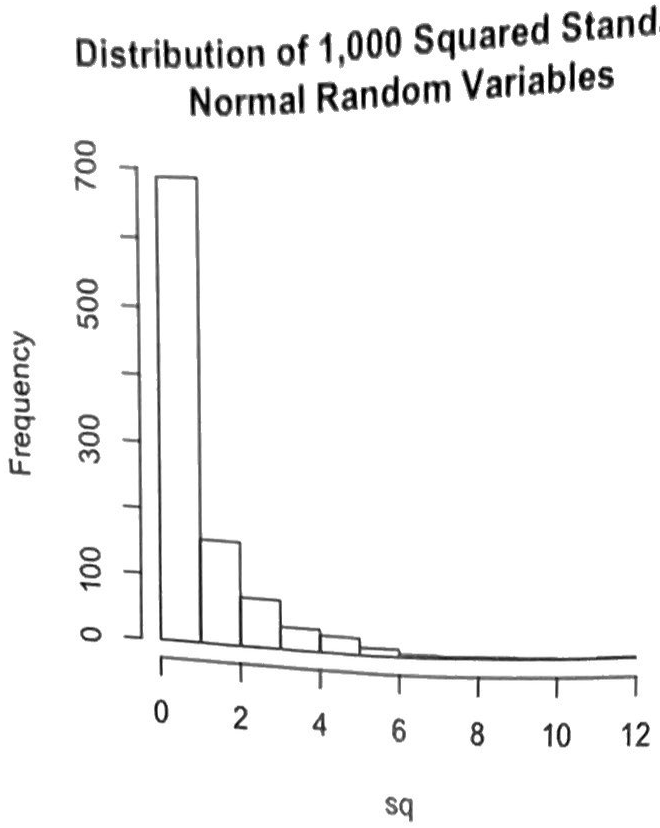


The theoretical mean is 0, and the theoretical variance is 1. The values obtained, as well as sample and theoretical quartiles, are shown in the R output below:

```
> mean(x)
[1] 0.01331869
> var(x)
[1] 0.9885846
> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.34845 -0.66343  0.08198  0.01332  0.72532  2.82358
> c(qnorm(0.25), qnorm(0.5), qnorm(0.75))
[1] -0.6744898  0.0000000  0.6744898
```

b.

```
> sq <- x^2
> hist(sq, main="Distribution of 1,000 Squared Standard \n Normal Random Variables")
```



The theoretical mean is 1 and the theoretical variance is 2 for a χ^2 distribution. The values obtained, as well as sample and theoretical quartiles, are shown in the R output below:

```
> mean(sq)
[1] 0.9877734
> var(sq)
[1] 1.833183
> summary(sq)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.000001  0.112705  0.489147  0.987773  1.286319 11.212131
> c(qchisq(0.25, 1), qchisq(0.5, 1), qchisq(0.75, 1))
[1] 0.1015310 0.4549364 1.3233037
```

c.

The requested simulation was performed in R, and the output is shown below:

```
> z1 <- rnorm(1000, 0, 1)
> z2 <- rnorm(1000, 0, 1)
> sum_sqrs <- (z1^2+z2^2)
> mean(sum_sqrs)
[1] 2.03372
> var(sum_sqrs)
[1] 3.965977
```

Thus, before any calculation or theory, it can be estimated from the simulation results that the expected value of the sum of squares of two independent standard normal deviates is 2, while the variance is 4. Indeed, when we observe that $E({Z_1}^2 + {Z_2}^2) = E({\chi_{df=2}}^2) = 2$ and $Var({Z_1}^2 + {Z_2}^2) = Var({\chi_{df=2}}^2) = 4$, we can confirm that our simulation is fairly accurate.

6.

a.

i. We can calculate this by finding the probability that the lower bound of one confidence interval is larger than the upper bound of the other, and then multiplying that probability by two. Thus:

$$P(no\ overlap) = 2P\left(\bar{X}_1 - 1.96\left(\frac{\sigma_1}{\sqrt{n_1}}\right) > \bar{X}_2 + 1.96\left(\frac{\sigma_2}{\sqrt{n_2}}\right)\right)$$

$$= 2P\left(\bar{X}_1 - \bar{X}_2 > 1.96\left(\frac{\sigma_2}{\sqrt{n_2}} + \frac{\sigma_1}{\sqrt{n_1}}\right)\right)$$

We proceed by standardizing:

$$= 2P\left(Z > \frac{1.96\left(\frac{\sigma_2}{\sqrt{n_2}} + \frac{\sigma_1}{\sqrt{n_1}}\right)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}\right)$$

Since $n_1 = n_2$ and $\sigma_1 = \sigma_2$,

$$= 2P\left(Z > \frac{1.96 \bullet 2\left(\frac{\sigma_1}{\sqrt{n_1}}\right)}{\sqrt{2\left(\frac{{\sigma_1}^2}{n_1}\right)}}\right)$$

$$= 2P\left(Z > 1.96 \bullet \sqrt{2}\right) = 2P(Z > 2.7719) = 0.0056$$ ✓

ii. We can calculate this by again finding the probability that the lower bound of one confidence interval is larger than the upper bound of the other, and then multiplying that probability by two. Thus:

$$P(no\ overlap) = 2P\left(\bar{X}_1 - 1.96\left(\frac{\sigma_1}{\sqrt{n_1}}\right) > \bar{X}_2 + 1.96\left(\frac{\sigma_2}{\sqrt{n_2}}\right)\right)$$

$$= 2P\left(\bar{X}_1 - \bar{X}_2 > 1.96\left(\frac{\sigma_2}{\sqrt{n_2}} + \frac{\sigma_1}{\sqrt{n_1}}\right)\right)$$

We proceed by standardizing:

$$= 2P\left(Z > \frac{1.96\left(\frac{\sigma_2}{\sqrt{n_2}} + \frac{\sigma_1}{\sqrt{n_1}}\right)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}\right)$$

Since $n_1 = 2n_2$ and $\sigma_1 = 3\sigma_2$, ✓

$$= 2P\left(Z > \frac{1.96\left(\frac{\sigma_2}{\sqrt{n_2}} + \frac{3\sigma_2}{\sqrt{2n_2}}\right)}{\sqrt{\frac{9{\sigma_2}^2}{2n_2} + \frac{{\sigma_2}^2}{n_2}}}\right) = 2P\left(Z > \frac{1.96 \bullet \frac{3+\sqrt{2}}{\sqrt{2}}\left(\frac{\sigma_2}{\sqrt{n_2}}\right)}{\frac{\sqrt{11}}{\sqrt{2}}\left(\frac{\sigma_2}{\sqrt{n_2}}\right)}\right)$$

$$= 2P\left(Z > 1.96 \bullet \frac{3+\sqrt{2}}{\sqrt{11}}\right) = 2P(Z > 2.6086) = 0.0091$$

b.

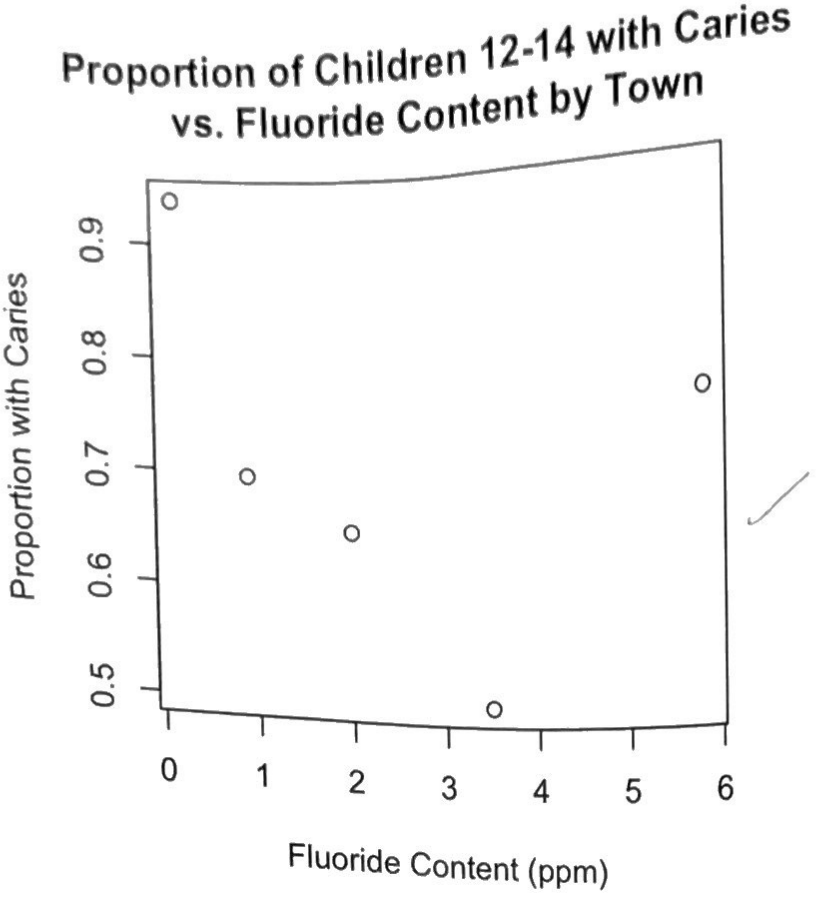
Based on the calculations performed in part (a), we can say that when two confidence intervals do not overlap at the $\alpha = 0.05$ confidence level, there is still a small chance that the population means being compared are equal, although it is very unlikely. Additionally, when they do overlap, there is a very high likelihood that the population means are not significantly different.

(could relate to a formal hypothesis test at $\alpha = .05$

7.

The following R output shows the given data table reproduced, as well as the output for Pearson's χ^2 test. Because the p-value is so incredibly small, we can state that at virtually any confidence level, there is a significant relationship between fluoride content and the presence or absence of caries. It is difficult to describe exactly what that relationship is without further investigation, but the data is graphed below to give some visual insight into the relationship. In general, it looks like higher fluoride content corresponds with lower presence of caries, but the Meresa data does not agree with that trend.

```
> dat_tab
        flour caries no_caries     car_p
Essex    0.15    243        16 0.9382239
Slough   0.90     83        36 0.6974790
Harwick  2.00     60        32 0.6521739
Burnham  3.50     31        31 0.5000000
Meresa   5.80     39        12 0.7647059
> chisq.test(dat_tab[,c(2,3)])

        Pearson's Chi-squared test

data:  dat_tab[, c(2, 3)]
X-squared = 80.216, df = 4, p-value < 2.2e-16
```



8.

a.

Each student's likelihood generating no more than 3 confidence intervals that miss is independent. Thus, we can find the probability that one student generates no more than 3 confidence intervals that miss, and then raise it to the 10^{th} power. We can essentially treat the random variable $X = \#$ of "misses" as though it is binomially distributed with parameters $n = 20$ and $p = 0.1$ and $q = 0.9$:

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= \binom{20}{0}(0.1)^0(0.9)^{20} + \binom{20}{1}(0.1)^1(0.9)^{19} + \binom{20}{2}(0.1)^2(0.9)^{18}$$
$$+ \binom{20}{3}(0.1)^3(0.9)^{17} = 0.8670$$

In order to determine the probability that all ten students generate no more than 3 confidence intervals that miss is found by raising this result to the 10^{th} power as indicated above:

b. $0.8670^{10} = 0.2401$ ✓

Based on the data given, it looks like 5% of the class generated more than 3 confidence intervals that miss. If all we are interested in is evaluating whether the appropriate proportion of students generated three or fewer confidence intervals that miss, we can define the r.v. $X \sim bin(100, 0.8670)$ to be the number of students that generate three or fewer misses, and then observe that the probability of attaining a result as or more extreme as the one we have is given by:

$P(X \geq 95)$

$$= \binom{100}{95}(0.8670)^{95}(0.1330)^{5} + \binom{100}{96}(0.8670)^{96}(0.1330)^{4}$$
$$+ \binom{100}{97}(0.8670)^{97}(0.1330)^{3} + \binom{100}{98}(0.8670)^{98}(0.1330)^{2}$$
$$+ \binom{100}{99}(0.8670)^{99}(0.1330)^{1} + \binom{100}{100}(0.8670)^{100}(0.1330)^{1}$$
$$= 0.0018$$

Such a result indicates to us, that this occurrence was very unlikely given our probability model. It could be that our model is an inappropriate one for this distribution. We can also check by performing a χ^2 goodness-of-fit test. This is perhaps a better way to evaluate the goodness-of-fit of our model, given the distribution of misses. Given our model, the expected results are as follows:

# of Misses	**# of Students (Observed)**	**# of Students (Expected)**
0	2	$100\binom{20}{20}(0.9)^{20}(0.1)^{0} = 12.16$
1	4	$100\binom{20}{19}(0.9)^{19}(0.1)^{1} = 27.02$
2	12	$100\binom{20}{18}(0.9)^{18}(0.1)^{2} = 28.52$
3	77	$100\binom{20}{17}(0.9)^{17}(0.1)^{3} = 19.01$
4	5	$100\binom{20}{16}(0.9)^{16}(0.1)^{4} = 8.98$
≥5	0	$100(1 - P(X \leq 4)) = 4.32$

$$\chi^2 = \frac{(O_0 - E_0)^2}{E_0} + \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} + \frac{(O_5 - E_5)^2}{E_5}$$

$$= \frac{(2 - 12.16)^2}{12.16} + \frac{(4 - 27.02)^2}{27.02} + \frac{(12 - 28.52)^2}{28.52} + \frac{(77 - 19.01)^2}{19.01} + \frac{(5 - 8.98)^2}{8.98} + \frac{(0 - 4.32)^2}{4.32} = 220.61$$

-1

df = 5

The χ^2 statistic we have obtained is unreasonably large, and without getting an exact p-value we can confidently state that our model does not provide an adequate fit for the data we have observed. Beyond the model not being a good fit, it would be difficult to say why we could get such an unusual result, other than the possibility that maybe the students all used the same output (i.e. there's a cheating problem that needs to be addressed).

Good!