This assignment includes problems related to analysis of two-way categorical data, analysis of variance, and a couple review problems.

1. Exercise 10.38 (page 420).
   Data records sex and whether or not the individual has perfect pitch for 99 conservatory of music students. Use Fisher's exact test of the null hypothesis that sex is independent of having perfect pitch.

   Solution: The data can be tabulated as follows.

	Perfect pitch	Not perfect pitch	
Male	1	50	51
Female	9	39	48
	10	89	99

   For Fisher's exact test we may think of the row totals to be balls of different color to be partitioned into group totals of the columns size or vice versa. We wish to know the probability of this table or one more extreme. We can pick any of the four cells in the table as the random count and compute the appropriate probability.

   Here is one choice. Let X be the number of men with perfect pitch. We wish to find $\Pr\{X \leq 1\}$. (Note, had we chosen the upper right cell, we would have wished to compute $\Pr\{X \geq 50\} = 1 - \Pr\{X \leq 49\}$.) Let's think of there being 10 black balls and 89 white balls. We pick 51 without replacement at random. The probability of one or fewer black balls in the sample is the cumulative hypergeometric distribution probability computed in R.

```
> phyper(1, 10, 89, 51)

[1] 0.005909976
```

   Here is one of the eight equivalent alternatives.

```
> 1 - phyper(49, 89, 10, 51)

[1] 0.005909976
```

2. Exercise 10.44 (page 425). Data shows three different environments in which juvenile lobsters can develop and the resultant claw configurations.

   Solution:

   (a) **Do the chi-square test if the test statistic is 24.35.** There are $(3-1)(3-1) = 4$ degrees of freedom. The p-value is less than 0.0001.

```
> 1 - pchisq(24.35, 4)

[1] 6.795404e-05
```

   There is very strong evidence that the environment is related to claw development.

   (b) **Verify the calculation.**

```
> observed <- matrix(c(8, 2, 7, 9, 4, 9, 1, 20, 7), 3, 3)
> rsum <- apply(observed, 1, sum)
> csum <- apply(observed, 2, sum)
> gsum <- sum(rsum)
> expected <- (rsum %o% csum)/gsum
> observed
     [,1] [,2] [,3]
[1,]    8    9    1
[2,]    2    4   20
[3,]    7    9    7

> expected
```

```
         [,1]     [,2]      [,3]
[1,] 4.567164 5.910448  7.522388
[2,] 6.597015 8.537313 10.865672
[3,] 5.835821 7.552239  9.611940
> x2 <- sum((observed - expected)^2/expected)
> x2
[1] 24.36374
> 1 - pchisq(x2, 4)
[1] 6.752389e-05
```

(c) **Show the distributions of claw configuration for each treatment.**

```
> percents <- round(100 * observed/cbind(rsum, rsum, rsum), 1)
> percents
     rsum rsum rsum
[1,] 44.4 50.0  5.6
[2,]  7.7 15.4 76.9
[3,] 30.4 39.1 30.4
```

(d) **Interpret the results.** The treatment of developing lobsters in smooth plastic containers without chips with which to exercise is more likely to result in lobsters with two cutter claws.

3. Exercise 10.52 (page 432).
Explain why the data from the bedrest for twins example is inappropriate to use to find a confidence interval for a difference in proportions.

Solution: The 210 twins in the bedrest group and the 214 twins in the control group are not an independent samples, but rather are groups of 105 and 107 sets of twins, respectively. We would need to control for the dependence between twins.

4. Exercise 10.68 (page 447).
A cross produces 89 glandular plants and 36 glandless plants. Use a goodness-of-fit test to see if the data is consistent with an 11:5 theory or a 13:3 theory.

Solution:

(a) **11:5 theory**

```
> observed <- c(89, 36)
> expected <- sum(observed) * c(11/16, 5/16)
> x2 <- sum((observed - expected)^2/expected)
> x2
[1] 0.3492364
> 1 - pchisq(x2, 1)
[1] 0.5545457
```

The data is consistent with this theory.

(b) **13:3 theory**

```
> observed <- c(89, 36)
> expected <- sum(observed) * c(13/16, 3/16)
> x2 <- sum((observed - expected)^2/expected)
> x2
[1] 8.287385
> 1 - pchisq(x2, 1)
[1] 0.003992144
```

There is strong evidence that the 13:3 theory does not fit the observed cotton plant phenotype distribution.

5. Exercise 11.1 (page 466).
Make computations on ficticious data.

Solution:

```
> ex11.1 <- list(samp1 = c(48, 39, 42, 43), samp2 = c(40, 48, 44),
+     samp3 = c(39, 30, 32, 35))
> ex11.1

$samp1
[1] 48 39 42 43

$samp2
[1] 40 48 44

$samp3
[1] 39 30 32 35

> n <- unlist(lapply(ex11.1, length))
> n

samp1 samp2 samp3
    4     3     4

> m <- unlist(lapply(ex11.1, mean))
> m

samp1 samp2 samp3
   43    44    34

> v <- unlist(lapply(ex11.1, var))
> v

   samp1    samp2    samp3
14.00000 16.00000 15.33333

> grand <- sum(n * m)/sum(n)
> grand

[1] 40

> ssb <- sum(n * (m - grand)^2)
> ssb

[1] 228

> ssw <- sum((n - 1) * v)
> ssw

[1] 120

> sstot <- sum((c(ex11.1$samp1, ex11.1$samp2, ex11.1$samp3) - grand)^2)
> sstot

[1] 348

> ssb + ssw
```

```
[1] 348

> msb <- ssb/2
> msb

[1] 114

> msw <- ssw/sum(n - 1)
> msw

[1] 15

> sp <- sqrt(msw)
> sp

[1] 3.872983
```

6. Exercise 11.4 (page 467).
Complete the ANOVA table.

Solution:

Source	df	SS	MS
Between groups	3	135	45
Within groups	12	337	28.08333
Total	15	472	

There were 4 groups (3=4-1) in the study.

There were 16 total observations in the study.

7. Exercise 6.58 (page 222).
Construct a confidence interval for the proportion of pregnant female adult white-tailed deer in the central Adirondack area.

Solution:

```
> count <- 97
> total <- 127
> ptilde <- (97 + 2)/(127 + 4)
> ptilde

[1] 0.7557252

> se <- sqrt(ptilde * (1 - ptilde)/(total + 4))
> se

[1] 0.03753925

> ptilde - 1.96 * se

[1] 0.6821483

> ptilde + 1.96 * se

[1] 0.8293021
```

We are 95% confident that the proportion of adult female white-tailed deer in the central Adirondack area that are pregnant is between 0.682 and 0.829.

8. Exercises 7.86 and 7.87 (page 310).

Solution: Exercise 7.86.

(a) "We are 95% confident that $\mu_1 > \mu_2$ because most of the interval is greater than 0" is incorrect. The value 0 is one of the plausible explanations of the data.

(b) "We are 95% confident that $\mu_1 - \mu_2$ is between -2.3 and 16.1" is correct.

(c) "We are 95% confident that $\bar{y}_1 - \bar{y}_2$ is between -2.3 and 16.1" is incorrect. We are certain of this.

(d) "95% of the nitric oxide infants were hospitalized longer than the average control infant" is incorrect. Confidence intervals are not about locations of individual observations in the population.

Exercise 7.87.

We would not reject at the $\alpha = 0.05$ level because 0 is in the interval. If 0 were not in the interval we would reject at that level.