Chapter 11 Solutions

11.1 As the number of trials increases, the graph of the likelihood becomes less dispersed. The maximum value occurs at the observed proportion. The graph of the log-likelihood is flatter, but it has its maximum at exactly the same place as the likelihood curve.

11.2 To find the G-statistic, we need to compute $G = -2\log(L_0) - (-2\log(L))$, where L_0 is the likelihood under the null model (same proportion for either drug) and L is the likelihood when we allow different proportions.

For the combined sample, we see 61 successes (pain free) among the 200 patients, so the combined proportion is $\hat{\pi}_0 = 61/200 = 0.305$. To find the likelihood under this null model we use

$$L_0 = 0.305^{61}0.695^{139} \text{ and so } \log(L_0) = 61\log(0.305) + 139\log(0.695) = -123.01$$

If we allow separate proportions for TMS and the placebo, the estimates are $\hat{p}_1 = 39/100 = 0.39$ and $\hat{p}_1 = 22/100 = 0.22$, respectively. Using the counts in the two-way table, this give a likelihood

$$L = 0.39^{39}0.61^{61}0.22^{22}0.78^{78} \text{ and } \log(L) = 39\log(0.39)+61\log(0.61)+22\log(0.22)+78\log(0.78) = -119.57$$

Note that -119.57 is the value labeled "Log-Likelihood" in the Minitab output. To compute the G-statistic we use

$$G = -2\log(L_0) - (-2\log(L)) = -2(-123.01) - (-2(-119.57)) = 6.88$$

This matches the value of the G-statistic in the Minitab output.

11.3 a. $L(\pi_1) = 1/4$, $L(\pi_2) = 1/3$, $L(\pi_3) = 1/2$

b. $L(\pi_1) = (1/4)(3/4) = 3/16$, $L(\pi_2) = (1/3)(2/3) = (2/9)$, $L(\pi_3) = (1/2)(1/2) = 1/4$

c. $L(\pi_1) = (1/4)(3/4)^2 = 9/64$, $L(\pi_2) = (1/3)(2/3)^2 = (4/27)$, $L(\pi_3) = (1/2)(1/2)^2 = 1/8$

11.4 a. $L(\pi_1) = 1/4$, $L(\pi_2) = 1/2$, $L(\pi_3) = 2/3$; the largest of these is 2/3, so the spinner with maximum likelihood is the third spinner, which has π_3.

b. $L(\pi_1) = (1/4)(3/4) = 3/16$, $L(\pi_2) = (1/2)(1/2) = 1/4$, $L(\pi_3) = (2/3)(1/3) = 2/9$; the largest of these is 1/4, so the spinner with maximum likelihood is the second spinner, which has π_2.

c. $L(\pi_1) = (1/4)(3/4)^2 = 9/64$, $L(\pi_2) = (1/2)(1/2)^2 = 1/8$, $L(\pi_3) = (2/3)(1/3)^2 = 2/27$; the largest of these is 9/64, so the spinner with maximum likelihood is the first spinner, which has π_1.

11.5 a. The likelihood is

$$\frac{e^{0.5+0.7}}{1+e^{0.5+0.7}} = \frac{3.3201}{1+3.3201} = 0.769$$

b. The likelihood is

$$\frac{e^{0.5+0.7(2.2)}}{1+e^{0.5+0.7(2.2)}} = \frac{7.6906}{1+7.6906} = 0.885$$

c. The likelihood is the product of the answers from parts (a) and (b): $0.769(0.885) = 0.680$.

11.6 a. The likelihood is

$$\frac{e^{0.5-1}}{1+e^{0.5-1}} = \frac{0.60653}{1+0.60653} = 0.378$$

b. The likelihood is

$$\frac{e^{0.5-2}}{1+e^{0.5-2}} = \frac{0.22313}{1+0.22313} = 0.182$$

c. The likelihood is the product of the answers from parts (a) and (b): $0.378(0.182) = 0.0689$.

11.7 a. We are interested in building a model for π, the proportion of students whose archery scores improved. For the null deviance, we compare the full model $logit(\pi) = \beta_0 + \beta_1 Attendance$ to a reduced model with just a constant term $logit(\pi) = \beta_0$.

b. For the residual deviance, we compare a full model that allows a different proportion for each value of *Attendance*, $logit(\pi) = p_i$, to one based on a linear function of *Attendance*, $logit(\pi) = \beta_0 + \beta_1 Attendance$.

11.8 a. For a logistic model using *Attendance* as a single predictor, we can test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ by comparing the residual deviance to the null deviance, $\chi^2 = 16.220 - 16.071 = 0.149$. Under the null hypothesis, this quantity should follow a chi-square distribution with one degree of freedom (the number of parameters being tested). From the upper tail of this distribution beyond $\chi^2 = 0.149$, we find a P-value of 0.6995.

b. The χ^2 P-value is not the same as the z-test for the *Attendance* coefficient in the output (P-value $= 0.718$), but the values are close and both show little evidence that attendance is related to improvement.

c. The archery data contain only 18 data cases, and only three students failed to improve their scores. With a sample this small, the tests may be unreliable.

d. We could use a randomization test in cases where conditions for a traditional test may not be met.

11.9 a. The residual deviance compares a reduced model based on a linear function of length with a saturated model that estimates a separate proportion for each different length. This is a comparison of the proportions in the last two rows of the table: the logistic model $\hat{\pi}$ values and the sample proportion $\hat{p}$ values at each length.

b. The null deviance compares the predicted logit proportions to the proportion of successful putts estimated from a constant model. This estimate is just the overall proportion of putts made for the combined sample (ignoring length), $338/587 = 0.576$.

c. Test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$, where β_1 is the coefficient of *Length* in the logistic regression model. The test statistic is the difference between the null and residual deviances $\chi^2 = 87.1429 - 6.8257 = 80.3172$.

We find a P-value for this statistic using the upper tail of a chi-square distribution with one degree of freedom. This gives a P-value that is essentially zero, giving strong evidence of a relationship between the length of putts and the chance of successfully making them.

We could also test this relationship using the z-test for the slope coefficient in the original computer output. The test statistic is $z = -8.391$ and P-value is listed as "< 2e-16," which means less than 2×10^{-16}, or computer notation for "essentially zero."

d. The P-values from the chi-square and z-tests yield the same conclusion in this case. Both are very small and provide strong evidence of a relationship between putting success and the length of the putts. However, the conclusions will not always match exactly.

e. The advice in the text is to consider overdispersion to be an issue when the residual deviance is much larger than its degrees of freedom. In this case, the residual deviance is 6.8257 with 3 degrees of freedom.

11.10 a. To use both *Location* and *Address*, we need to include indicators for two of the locations and one of the addresses. One such model is

$$logit(\pi) = \beta_0 + \beta_1 GrinnellTown + \beta_2 GrinnellCampus + \beta_3 Peaceworks$$

Here is some output from fitting this model.

```
Logistic Regression Table

                                                     Odds     95% CI
Predictor           Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant       -0.424347  0.299532   -1.42   0.157
GrinnellTown     1.25197  0.478586    2.62   0.009   3.50   1.37   8.94
GrinellCampus   -1.50895  0.596497   -2.53   0.011   0.22   0.07   0.71
Peaceworks       1.41655  0.398286    3.56   0.000   4.12   1.89   9.00

Log-Likelihood = -79.202
Test that all slopes are zero: G = 31.542, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests
Method           Chi-Square  DF       P
Pearson             1.09921   2   0.577
Deviance            1.18619   2   0.553
Hosmer-Lemeshow     1.00640   3   0.800
```

We see that all coefficients are significant at a 5% level, so we can conclude that the return rate probably differs between Des Moines and Grinnell town, Des Moines and Grinnell campus, and the two addresses—even after accounting for the other factors in the model. By looking at the signs of the coefficients, we see that letters are more likely to be returned if lost in the town of Grinnell (compared to Des Moines), less likely in the Grinnell campus (compared to Des Moines), and more likely if addressed to Iowa Peaceworks (compared to Friends of the Confederacy).

From the G-statistic (31.542, P-value = 0.000), we can conclude that this model (as a whole) has some effectiveness at explaining whether or not lost letters are returned.

b. The residual deviance (Deviance = 1.18619 in the output) reflects how much is gained if we include a parameter for each of the six cells in a 3×2 table of *Location* by *Address*, compared to our logistic regression model with four parameters. We compare this value to a chi-square distribution with 2 degrees of freedom (difference in number of parameters in the saturated and reduced models) to get a P-value of 0.553. This is not a small P-value, so we don't find evidence that we need to expand beyond the logistic regression model based on *Location* and *Address*.

c. We find the saturated model probabilities by finding the proportion of letters returned within each cell of the 3×2 table for *Location* by *Address*. We compare these to the probabilities obtained from the fitted logistic regression model (shown in parentheses).

	Confederacy	Peaceworks
Des Moines	0.425 (0.395)	0.700 (0.730)
Grinnell campus	0.100 (0.126)	0.400 (0.374)
Grinnell town	0.650 (0.700)	0.950 (0.904)

d. The residual deviance (1.186) is smaller than its degrees of freedom (2), so we do not see a need to worry about overdispersion.

11.11 For either method, we first use software to compute predicted probabilities of a nest being closed, using the fitted logistic model based on *Length* and *TotCare*. For example, the first bird in the sample (Eastern Kingbird) with *Length* = 20 and *TotCare* = 34 gets a predicted probability of 0.496 to have closed nest.

a. The table below compares the actual nest type versus the predicted nest type, where nests are predicted to be closed if the fitted probability is bigger than 0.50.

		Predicted	
		$\widehat{Open}$	$\widehat{Closed}$
Actual	Open	48	9
	Closed	14	12

We see that $48 + 12 = 60$ of the 83 species (72%) are predicted correctly with the 0.5 cutoff.

b. The table below compares the actual nest type versus the predicted nest type, where nests are predicted to be closed if the fitted probability is bigger than 0.687 (the proportion of open nests in the sample).

		Predicted	
		$\widehat{Open}$	$\widehat{Closed}$
Actual	Open	56	1
	Closed	17	9

We see that $56 + 9 = 65$ of the 83 species (78%) are predicted correctly with the 0.687 cutoff.

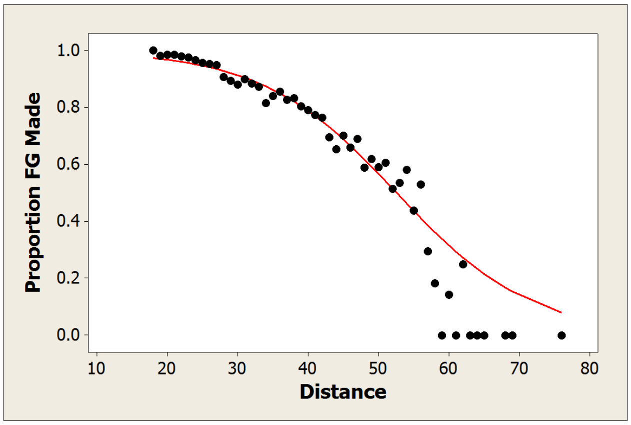
c. The model does fairly well at classifying the nest type correctly (72% to 78% depending on cutoff). The 0.687 cutoff does very well with the nests that are actually open (missing only 1 out of 57), but misses more of the closed nests since it's harder with that cutoff for a nest to be called closed.

11.12 a. We fit the logistic regression model $logit(\pi) = \beta_0 + \beta_1 Dist$, where π measures the proportion of field goals made and *Dist* is the distance of the attempt (in yards). Some output for fitting this model is shown below.

```
Variable  Value      Count
Makes     Event       6849
          Non-event   1671
N         Total       8520

Logistic Regression Table
                                                     Odds     95% CI
Predictor       Coef     SE Coef       Z      P  Ratio  Lower  Upper
Constant     5.48343    0.147606   37.15  0.000
Dist       -0.104269  0.0034906  -29.87  0.000   0.90   0.89   0.91

Log-Likelihood = -3656.652
Test that all slopes are zero: G = 1121.255, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             60.9645  49  0.117
Deviance            67.7289  49  0.039
Hosmer-Lemeshow     23.1194   7  0.002
```

The z-test for the slope ($z = -29.87$, P-value ≈ 0) and chi-square test based on the G-statistic ($\chi^2 = 1121.255$, P-value ≈ 0) give very strong evidence that the coefficient of *Dist* differs from zero. This indicates that distance is useful to help predict the proportion of field goals made.

Using the deviance statistic (67.7289, P-value $= 0.039$) we test the goodness-of-fit by comparing this model to one that uses a separate proportion for each field goal distance (rather than the logit function depending on a linear function of distance). This P-value is less than 5%, so we might suspect that the model can be improved. A plot of the proportion of successful kicks at each distance (black dots) and the predicted proportions from the logit model (red line) shows relatively good agreement, except for the longer distances.



If we wanted to improve this model, we might eliminate the very long attempts (perhaps over 55 yards) since these are much less successful and have small sample sizes. We might also consider other predictors to add to the model, such as an indicator for grass versus artificial turf fields, or a variable that tells something about weather conditions.

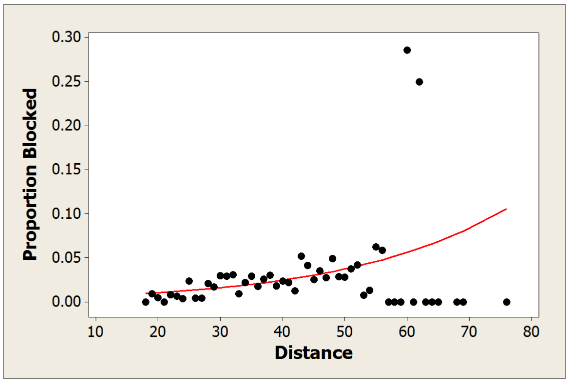
b. To model blocked field goals, we fit the logistic regression model $logit(\pi) = \beta_0 + \beta_1 Dist$, where now π measures the proportion of field goals that are blocked and $Dist$ is the distance of the attempt (in yards). Some output for fitting this model is shown below.

```
Variable  Value      Count
Blocked   Event        196
          Non-event   8324
N         Total       8520
                                                Odds      95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant     -5.37807   0.313426  -17.16  0.000
Dist        0.0426525  0.0076062    5.61  0.000   1.04   1.03   1.06

Log-Likelihood = -916.818
Test that all slopes are zero: G = 32.467, DF = 1, P-Value = 0.000
```

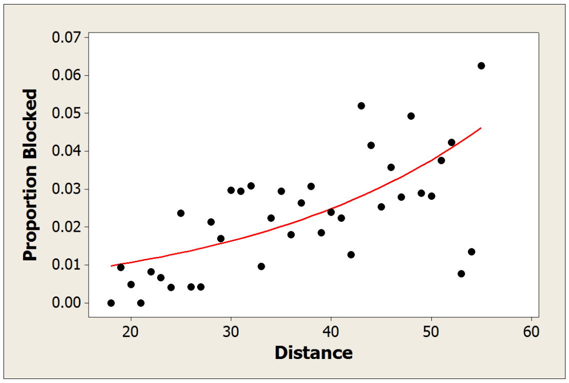
```
Goodness-of-Fit Tests
Method            Chi-Square  DF      P
Pearson              53.8297  49  0.295
Deviance             56.6338  49  0.212
Hosmer-Lemeshow      16.1287   7  0.024
```

The z-test for the slope ($z = 5.61$, P-value ≈ 0) and chi-square test based on the G-statistic ($\chi^2 = 32.47$, P-value ≈ 0) give very strong evidence that the coefficient of *Dist* differs from zero. This indicates that distance is useful to help predict the proportion of field goals that are blocked. The positive coefficient for *Dist* indicates that field goals are more likely to be blocked as the distance of the attempt increases.

Using the deviance statistic (56.63, P-value $= 0.212$), we test the goodness-of-fit by comparing this model to one that uses a separate proportion for each field goal distance (rather than the logit function depending on a linear function of distance). This P-value is not very small, so we don't see much evidence that the fit of the model could be improved with a different function of distance. A plot of the proportion of blocked kicks at each distance (black dots) and the predicted proportions from the logit model (red line) shows relatively good agreement, except again at the longer distances, where there are few attempts.



If we eliminate the attempts over 55 yards (which occur rarely and are rarely successful), we can refit the logistic regression model. The plot that follows shows the proportion blocked at each distance and the (red) line for the predicted probabilities based on the logit fit. Although this looks more scattered than the plot above, not that the vertical scale is much different—and more appropriate for what is usually a rare event.



11.13 a. The table below shows that $44 + 53 = 97$ (78.9%) of the 123 blue jays are correctly classified:

	KnownSex	
MalePred	F	M
FALSE	44	10
TRUE	16	53

b. The table below shows that $50+53 = 103$ (83.7%) of the 123 blue jays are correctly classified:

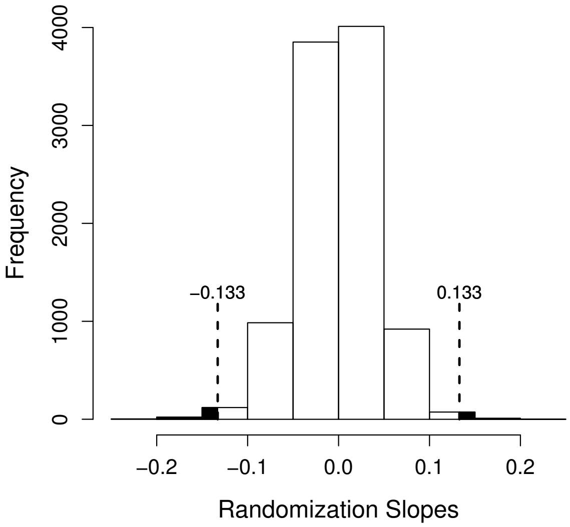
	KnownSex	
MalePred	F	M
FALSE	50	10
TRUE	10	53

11.14 a. We fit a logistic regression model to predict $Result2$ (whether or not a field goal is made) using $Yards$, the distance of each attempt (in yards). Some output for fitting this model is shown below.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.21805    2.59795   2.393   0.0167 *
Yards        -0.13291    0.05968  -2.227   0.0260 *

    Null deviance: 36.652  on 29  degrees of freedom
Residual deviance: 29.069  on 28  degrees of freedom
```

If we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using the z-test in the output ($z = -2.27$, P-value $= 0.0260$), we have fairly strong evidence that the chance of making a field goal is related to the $Yards$ of the attempt.

b. To repeat the test in (a) using a randomization procedure, we scramble the values in $Results2$ so they are not related to the $Yards$, refit the logistic regression, and record the slope. After repeating this process for 10,000 randomizations (under H_0), we get a histogram of slopes shown as follows.



Only 48 of the 10,000 randomization slopes are less than the slope observed in the original sample ($\hat{\beta}_1 = -0.133$), and only 18 more are as extreme in the upper tail. This gives an estimated P-value of $(48 + 18)/10,000 = 0.0066$. This is very small, giving strong evidence that a relationship between results and distance as strong as we see in this sample of 30 field goal attempts would happen very rarely by random chance alone.

c. Both of the tests in (a) and (b) give significant results at a 5% level, but only the randomization test is also significant at a 1% level. This is a small sample. We might be concerned that the z-test is less reliable since it depends on having a large sample for the normal approximation to be appropriate. The randomization procedure has no such conditions, so we would be more confident in its conclusion.

11.15 a. Below we have the two-way table, followed by the table of row proportions, rounded to ease the viewing. We see that the incidence rate for gunnels is definitely lowest when there are no crustaceans in the quadrat, but for higher densities the first density (1) shows a clearly higher incidence rate of about 8.5% than the higher densities, which show incidence rates between 5 and 5.8%.

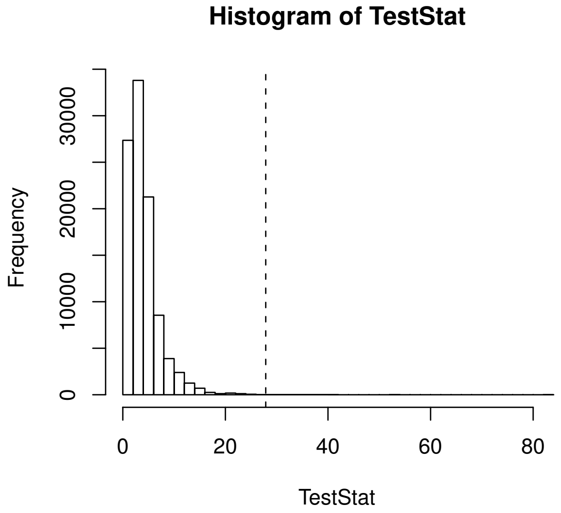
	GUNNEL	
AMPH_ISO	0	1
0	1165	21
1	151	14
2	160	9
3	49	3
4	19	1

Row proportions (rounded):

	GUNNEL	
AMPH_ISO	0	1
0	0.982	0.018
1	0.915	0.085
2	0.947	0.053
3	0.942	0.058
4	0.950	0.050

b. To perform the randomization test, we first computed the chi-square test statistic on the original two-way table from the data and obtain a value of TestStat0 = 27.86596. Any statistical package (e.g., Minitab or R) will give you this value with a simple menu or function call. We then repeatedly scramble at random the vector of 1592 *Amphiso* labels to create a succession of new versions of the *Amphiso* vector. With each of these iterations, we form a new two-way table with *Amphiso* as the row variable and *Gunnel* as the column variable and each time we calculate a value of the chi-square test statistic.

We performed this iteration 100,000 times, and the following histogram shows the 100,000 values we obtained. The dashed line is drawn at the value of TestStat0 = 27.86596. Notice that very few values of TestStat fall beyond TestStat0. In fact, only 41 of the 100,000 values are in this upper tail, giving us an empirical P-value of 0.00041, which is highly statistically significant.



11.16 a. (a) Following is the two-way table, followed by the table of row proportions, both unrounded and then rounded to ease the viewing. We see that the incidence rate for gunnels is definitely higher for substrata involving cobbles. For example, the categories 2, 7, 8, and 9 have the highest incidence rate of gunnels and these all involve cobbles. Two-way table of counts:

	Gunnel	
Subst	0	1
1	781	11
2	97	3
3	186	2
4	147	0
5	50	0
6	11	0
7	81	5
8	141	25
9	9	2
10	20	0
11	4	0
12	10	0
13	7	0

Row proportions for two-way table:

	Gunnel	
Subst	0	1
1	0.98611111	0.01388889
2	0.97000000	0.03000000
3	0.98936170	0.01063830
4	1.00000000	0.00000000
5	1.00000000	0.00000000
6	1.00000000	0.00000000
7	0.94186047	0.05813953
8	0.84939759	0.15060241
9	0.81818182	0.18181818
10	1.00000000	0.00000000
11	1.00000000	0.00000000
12	1.00000000	0.00000000
13	1.00000000	0.00000000

b. To perform the randomization test, we first computed the chi-square test statistic on the original two-way table from the data and obtain a value of TestStat0 = 110.6712. Any statistical package (e.g., Minitab or R) will give you this value with a simple menu or function call. We then repeatedly scramble at random the vector of 1592 *Subst* labels to create a succession of new versions of the *Subst* vector. With each of these iterations, we form a new two-way table with *Subst* as the row variable and Gunnel as the column variable and each time we calculate a value of the chi-square test statistic.

We performed this iteration 100,000 times, and the histogram below shows the 100,000 values we obtained. As we can see from the histogram, none of the 100,000 values reach or exceed the 100.6712 value from the original data, thus giving us an empirical P-value of 0.00000, which is highly statistically significant.

c. The two subtables are:

Subtable 1:

		Gunnel	
	Subst	0	1
	2	97	3
	7	81	5
	8	141	25
	9	9	2
	10	20	0
	11	4	0
	12	10	0
Totals:		362	35

So the overall incidence of quadrats with gunnels is $35/397 = 0.088$.

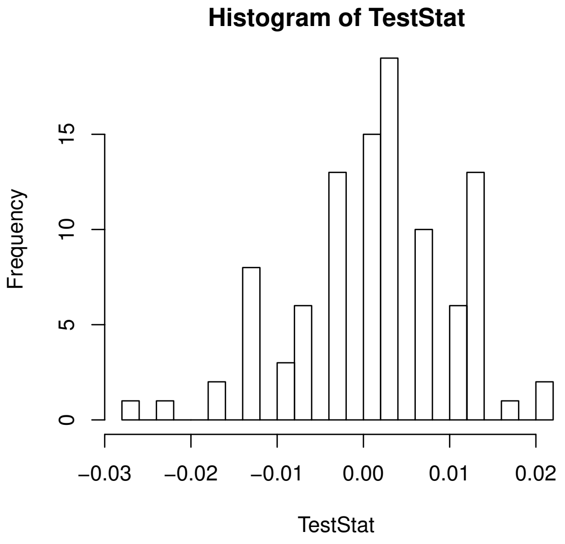
Substable 2:

		Gunnel	
	Subst	0	1
	1	781	11
	3	186	2
	4	147	0
	5	50	0
	6	11	0
	13	7	0
Totals:		1182	13

So the overall incidence of quadrats with gunnels is $13/1195 = 0.011$.

Thus the value of the test statistic is $35/397 - 13/1195 = 0.0773$.

d. We iterated the randomization process 100,000 times and obtained the test statistic values given in the histogram that follows. None of these values were as large as the observed 0.0773. Thus our empirical P-value $= 0.00000$, and we conclude a highly significant result. The data support the expectation that cobbled substrata are preferred by the gunnels.



11.17 a. The two-way table shows that, as expected, the Peaceworks address elicits a higher estimated return rate: 95% versus 65%.

	Peaceworks	
Returned	0	1
0	7	1
1	13	19

	Peaceworks	
Returned	0	1
0	0.35	0.05
1	0.65	0.95

b. The coefficient of 2.3254 shows that letters with Iowa Peaceworks on the envelope are more likely to be returned. The odd ratio is $e^{(}2.3254) = 10.23077$. Thus the odds of a return about 10 times greater for an Iowa Peaceworks letter than for a Friends of the Confederacy letter. The P-value of 0.0392 meets the conventional 0.05 criterion for statistical significance, so we would say the relationship is statistically significant.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6190     0.4688   1.320   0.1867
Peaceworks    2.3254     1.1280   2.062   0.0392 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. We iterated the randomization process 100,000 times. Each iteration consisted of a random permutation of the vector of 40 0-1 values of the *Peaceworks* variable. Then using this random version of *Peaceworks* alongside the original vector of 40 returned values, we fit a logistic model from which we computed the estimated odds ratio (exponentiating the slope coefficient). We accumulated this odds ratio into a vector of length 100,000. Of these 100,000 simulated odds ratios, only 1406 were as large as or larger than the 10.23077 value of the original data. This represents a P-value of 0.01406, which is more statistically significant that estimated by the logistic model in (b). Note that we have conducted a direction test; using a nondirectional alternative roughly doubles the P-value.

11.18 a. Here are the two-by-two tables, first to many digits and then rounded to simplify the look:

	Gunnel	
Crust	0	1
False	0.98229342	0.01770658
True	0.93349754	0.06650246

	Gunnel	
Crust	0	1
False	0.982	0.018
True	0.933	0.067

There is a much greater incidence of gunnel presence when there is a non-negligible density of the crustaceans in the quadrat: 6.7% versus 1.8%.

b. The value of the test statistic is 22.99, and the P-value is tiny.

c. The code below confirms the statistical significance observed above. The P-value is 3.68e-06, very much consistent with the chi-square results. (Note: These two significance tests are not mathematically equivalent, but should usually confirm one another.)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0160     0.2201 -18.242  < 2e-16 ***
CRUSTTRUE     1.3743     0.2969   4.629 3.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 409.58  on 1590  degrees of freedom
```

d. The estimated odds ratio is $e^{1.3743} = 3.95$. So the odds of finding gunnels is nearly 4 times greater when there is a non-negligible supply of the food source. A gunnel colony marches on its belly. The 95% confidence interval can be calculated by hand:

$$1.3743 \pm 1.96(0.2969) = 1.3743 \pm 0.5819 = (0.7924, 1.9562)$$

(0.7924, 1.9562) is the 95% confidence interval for the slope parameter. Exponentiate this to get the confidence interval for odds-ratio, getting: 2.21 to 7.07. So we are 95% confident the odds ratio is between 2.2 and 7.1, with 3.95 being our point estimate.

11.19 a. The following is the two-way table, followed by the table of row proportions, both unrounded and then rounded to ease the viewing. We see that the incidence rate for gunnels is definitely lowest when there are no crustaceans in the quadrat, but for higher densities the first density (1) shows a clearly higher incidence rate of about 8.5% than the higher densities, which show incidence rates between 5 and 5.8%.

	Gunnel	
Amphiso	0	1
0	1165	21
1	151	14
2	160	9
3	49	3
4	19	1

Row proportions:

Amphiso	0	1
0	0.98229342	0.01770658
1	0.91515152	0.08484848
2	0.94674556	0.05325444
3	0.94230769	0.05769231
4	0.95000000	0.05000000

Row proportions, rounded:

Amphiso	0	1
0	0.982	0.018
1	0.915	0.085
2	0.947	0.0533
3	0.942	0.058
4	0.950	0.050

b. The results of the chi-square test on the contingency table follow. The *P*-value of 1.328e-05 tells us that there is a clear statistically significant relationship between these variables. This test does not tell us the nature of the relationship.

```
        Pearson’s Chi-squared test

data:  tab
X-squared = 27.866, df = 4, p-value = 1.328e-05

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

However, we note the warning message, alerting us to an expected table with several small expected counts. Combing categories 2, 3, and 4 into a new category 2, we obtain the following results, with the expected values given as well.

	Gunnel	
Amph2	0	1
0	1165	21
1	151	14
2	228	13

Table of expected counts:

	Gunnel	
Amph2	0	1
0	1150.2412	35.758794
1	160.0251	4.974874
2	233.7337	7.266332

```
Pearson’s Chi-squared test

data:  tab
X-squared = 27.8276, df = 2, p-value = 9.064e-07
```

Note: One expected count is very slightly below 5, not a situation to create doubt in the validity of the highly significant P-value.

c. We define indicator variables for categories 1, 2, 3, and 4 and name them AI1, AI2, AI3, and, AI4. We do use category 0 as the omitted case and leave it out of the model. We get the following output when fitting the logistic model:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0160     0.2201 -18.242  < 2e-16 ***
AI1           1.6377     0.3557   4.604 4.14e-06 ***
AI2           1.1380     0.4072   2.795   0.0052 **
```

```
AI3           1.2227     0.6342   1.928   0.0539 .
AI4           1.0715     1.0493   1.021   0.3072
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 408.08  on 1587  degrees of freedom
```

The significance test that answers whether there is a relationship between *Amphiso* and *Gunnel* is a drop-in-deviance between the null and residual deviances. That is: $430.69 - 408.08 = 22.61$ on 4 degrees of freedom $(1591 - 1587)$, which is a P-value of 0.00015, a highly significant result that confirms what we found with the chi-square tests in (b).

11.20 a.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0160     0.2201 -18.242  < 2e-16 ***
AI1           1.6377     0.3557   4.604 4.14e-06 ***
AISUM         1.1516     0.3602   3.197  0.00139 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 408.10  on 1589  degrees of freedom
```

b. First model (full)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.01595    0.22014 -18.242  < 2e-16 ***
AI1          1.63773    0.35569   4.604 4.14e-06 ***
AISUM        1.07151    1.04933   1.021    0.307
AI2          0.06649    1.08166   0.061    0.951
AI3          0.15123    1.18591   0.128    0.899
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 408.08  on 1587  degrees of freedom
```

Second model (reduced):

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0160     0.2201 -18.242  < 2e-16 ***
AI1           1.6377     0.3557   4.604 4.14e-06 ***
AISUM         1.1516     0.3602   3.197  0.00139 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 408.10  on 1589  degrees of freedom
```

c. The model in (b) is:

$$b_1 AI1 + b_2(AI2 + AI3 + AI4) + b_3 AI2 + b_4 AI3$$

If we let $b_3 = b_4 = 0$ in this model, we get: $b_1 AI1 + b_2(AI2 + AI3 + AI4)$, which is the model from part (a).

This shows the model from part (a) is a reduced version of the model from part (b).

d. Here is the nested LRT test calculation:

```
Residual deviance: 408.08  on 1587  degrees of freedom (full)
Residual deviance: 408.10  on 1589  degrees of freedom (reduced)
```

$408.10 - 408.08 = 0.02$ on $1589 - 1587 = 2$ d.f. The P-value is 0.9900, which is clearly nonsignificant. This suggests the adequacy of the reduced model, that is, the coefficients for categories 2, 3, and 4 are effectively the same.

11.21 a. The table below gives the rounded figures.

	Water	
Gunnel	1	0
1	0.044	0.015
0	0.956	0.985

We see that with standing water there is an increased probability of the gunnel being present: 4.4% chance versus 1.5% chance.

b. Following is computer output for a two-sample z-test for proportions, which is equivalent to the chi-square test.

```
2-sample test for equality of proportions without continuity correction
X-squared = 11.5215, df = 1, p-value = 0.000688
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01249688 0.04570147
sample estimates:
    prop 1     prop 2
0.04444444 0.01534527
```

The output gives a P-value of 0.000688 and a 95% confidence interval of 0.01249688 to 0.04570147 for the difference in the two proportions (0.044 and 0.015) from (a). This difference in proportions is statistically significant.

c. The output in (b) gives the chi-square results.

d. Here is the computer output from the logistic model:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1615     0.2909 -14.305  < 2e-16 ***
Water         1.0934     0.3372   3.243  0.00118 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 418.61  on 1590  degrees of freedom
AIC: 422.61
```

We get corroboration of the significant relationship between the *Water* presence and the odds of finding a gunnel. Here the P-value is 0.00118.

e. The odds ratio computed from the model is $e^{1.0934} = 2.9844$, which is equal (up to some round-off error) to the OR directly computed from the table: $\frac{770(36)}{774(12)} = 2.9845$.

11.22 a. The computer output below gives the two-way table of counts.

```
        Spacing
Deer95  10  15
     0 411 251
     1 105 104
```

The two-way table of column percentages is:

```
          Spacing
Deer95     10    15
     0 0.797 0.707
     1 0.203 0.293
```

When the spacing is 10 feet, there is about a 20% chance of deer browsing; with 15-foot spacing, there is about a 29% chance. Thus deer are more likely to browse with the wider spacing.

b. The odds ratio is $\frac{411(104)}{251(105)} = 1.62$. So the odds for browsing are about 1.6 greater for the larger spacing.

c. The computer output below gives the results of the two-sample z-test. The "X-squared" statistic is the square of the usual z, so $z = 3.04$. The two-sided P-value is 0.0024.

```
data:  PinesTable
X-squared = 9.23, df = 1, p-value = 0.00238
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0307 0.1482
sample estimates:
prop 1 prop 2
 0.797  0.707
```

d. The computer output below gives the results of fitting the logistic model; the variable *Space* is an indicator variable for 15-foot spacing. These results confirm the result from part (c): The effect of spacing is highly significant (P-value $= 0.0025$). Note that $e^{0.484} = 1.62$, which is the odds ratio reported in (b).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3318     0.4025   -5.79  6.9e-09 ***
Space          0.484      0.160    3.02   0.0025 **
```

e. For 1997, the browsing behavior was different. Here, the proportion of trees browsed was close to the same for the two spacings: 11.1% for 10-foot spacing and 9.9% for 15-foot spacing.

```
        Spacing
Deer97   10  15
     0 455 318
     1  57  35
```

```
          Spacing
Deer97    10    15
     0 0.889 0.901
     1 0.111 0.099
```

The odds ratio (15 versus 10) is $\frac{455(35)}{57(318)} = 0.88$.

This difference in browsing percentages is not statistically significant; the P-value is 0.568:

```
data:  PinesTable97
X-squared = 0.33, df = 1, p-value = 0.568
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.136  0.074
sample estimates:
prop 1 prop 2
  0.59   0.62
```

The logistic fit confirms these findings:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.077      0.141  -14.78   <2e-16 ***
Space         -0.129      0.227   -0.57     0.57
```

We conclude that the preference for wider spacing in 1995 has disappeared by 1997. This may be due to the trees growing closer together as they get larger, making everything more crowded and mitigating the effect of greater spacing when the trees were planted.

11.23 a. The computer output below gives the two-way table of counts:

```
         Cover95
Deer95   0   1   2   3
     0 151 158 177 176
     1  60  76  44  29
```

The two-way table of column percentages is:

```
           Cover95
Deer95     0     1     2     3
     0 0.716 0.675 0.801 0.859
     1 0.284 0.325 0.199 0.141
```

As thorny cover goes up, the browsing percentage tends to go down.

b. The computer output below gives the results of the chi-square test. The P-value is 0.00002.

```
data:  CoverTable
X-squared = 24.4, df = 3, p-value = 2.02e-05
```

c. The computer output below gives the results of fitting the logistic model. These results confirm the result from part (b): The effect of thorny cover is highly significant for the two highest categories, when these are compared to the baseline category of Cover95 = 0.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.923      0.153   -6.05  1.5e-09 ***
as.factor(Cover95)1    0.191      0.207    0.92  0.35557
as.factor(Cover95)2   -0.469      0.227   -2.06  0.03907 *
as.factor(Cover95)3   -0.880      0.252   -3.49  0.00048 ***
```

d. To estimate the odds ratio for each level of thorny cover, we take that indicator's coefficient and exponentiate it; thus we are comparing each level to category 0, the absence of thorny cover. The calculations are

$$\begin{aligned} e^{0.191} &= 1.21; \\ e^{-0.469} &= 0.626; \\ e^{-0.880} &= 0.415. \end{aligned}$$

Thus the odds of browsing for category 1 are 1.21 times the odds for category 0; the odds of browsing for category 2 are 0.626 times the odds for category 0; and the odds of browsing for category 3 are 0.415 times the odds for category 0.

e. The table below gives a cross tabulation of browsing (no = 0, yes = 1) versus predicted browsing (no = FALSE, yes = TRUE). We see that of the 871 total cases, 353 were correctly predicted as not browsed trees and 136 were correctly predicted as browsed trees, giving a percentage correct of $(353 + 136)/871 = 489/871 = 56.1\%$. This is a rather weak probability of success.

```
   FALSE TRUE
 0   353  309
 1    73  136
```

f. For 1997, the overall level of browsing was lower than in 1995. The browsing percentages ranged from 8.6% to 12.4%, whereas the lowest percentage in 1995 was 14.1% (for category 3).

```
            Cover95
Deer97      0      1      2      3
     0 0.9139 0.8974 0.8761 0.8873
     1 0.0861 0.1026 0.1239 0.1127
```

Also, the relationship between cover category and browsing is not close to exceeding the kinds of fluctuations that would be seen by chance alone; the P-value is 0.631.

```
data:  CoverTable97
X-squared = 1.73, df = 3, p-value = 0.6313
```

The logistic fit confirms these findings:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.362      0.247   -9.58   <2e-16 ***
as.factor(Cover95)1    0.193      0.327    0.59     0.56
as.factor(Cover95)2    0.405      0.321    1.26     0.21
as.factor(Cover95)3    0.299      0.331    0.90     0.37
```

In summary, there was much less deer browsing in 1997 than in 1995 and none of the thorny cover categories stands apart from the others in 1997 deer browsing.

11.24 a. The computer output below gives the fit of the logistic model. The P-value for *Spacing* is less than 0.01, with *Spacing* having a positive coefficient so that greater spacing is associated with more browsing. Two of the levels of $Cover95$ have small P-values and negative coefficients, consistent with reduced browsing when thorny cover is greater.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.0009     0.4340   -4.61    4e-06 ***
as.factor(Cover95)1   0.2033     0.2079    0.98  0.32821
as.factor(Cover95)2  -0.4499     0.2283   -1.97  0.04878 *
as.factor(Cover95)3  -0.8341     0.2532   -3.29  0.00099 ***
Spacing               0.0870     0.0325    2.68  0.00736 **
---

Null deviance: 959.89  on 870  degrees of freedom
Residual deviance: 927.56  on 866  degrees of freedom
```

We can test the overall utility of the model by comparing it to a null model (i.e., a model with only a constant). The resulting chi-square test has a P-value that is tiny:

```
Model 1: Deer95 ~ 1
Model 2: Deer95 ~ as.factor(Cover95) + Spacing
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       870        960
2       866        928  4     32.3  1.6e-06 ***
```

b. The odds ratio is $e^{0.087} = 1.09$. The 95% confidence interval is (1.02,1.16) so the shift above 1.0 is statistically significant but not very large.

```
data:  CoverTable
X-squared = 24.4, df = 3, p-value = 2.02e-05
```

c. The full model was fit in part (a). The reduced model has only *Spacing* as a predictor; the computer output below shows the fit:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3318     0.4025   -5.79  6.9e-09 ***
Spacing       0.0967     0.0320    3.02   0.0025 **
---
    Null deviance: 959.89  on 870  degrees of freedom
Residual deviance: 950.76  on 869  degrees of freedom
```

The test statistic is 23.2 on 3 degrees of freedom, giving a tiny P-value:

```
Model 1: Deer95 ~ Spacing
Model 2: Deer95 ~ as.factor(Cover95) + Spacing
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       869        951
2       866        928  3     23.2  3.7e-05 ***
```

d. The fit of the full model is shown below. All of the P-values for the predictors are large, which is consistent with the test of overall usefulness of the model: The P-value is 0.74, which means that there is no evidence that the predictors are helpful.

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.0856     0.6077   -3.43   0.0006 ***
as.factor(Cover95)1   0.1898     0.3275    0.58   0.5623
as.factor(Cover95)2   0.3993     0.3213    1.24   0.2139
as.factor(Cover95)3   0.2849     0.3326    0.86   0.3916
Spacing              -0.0226     0.0456   -0.50   0.6205
---
```

```
    Null deviance: 586.18  on 864  degrees of freedom
Residual deviance: 584.19  on 860  degrees of freedom


Model 1: Deer97 ~ 1
Model 2: Deer97 ~ as.factor(Cover95) + Spacing
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       864        586
2       860        584  4        2     0.74
```

11.25 a. Here is some computer output for doing a chi-square test on the 3 × 2 table with the three *Location*s as rows and the *Returned* responses (1 = returned, 0 = not) as columns.

```
 Rows: Location   Columns: Returned

                        0      1     All

DesMoines              35     45      80
                    33.14  46.86   80.00

GrinnellCampus         15      5      20
                     8.29  11.71   20.00

GrinnellTown            8     32      40
                    16.57  23.43   40.00

All                    58     82     140
                    58.00  82.00  140.00

Cell Contents:        Count
                      Expected count

Pearson Chi-Square = 17.036, DF = 2, P-Value = 0.000
Likelihood Ratio Chi-Square = 17.771, DF = 2, P-Value = 0.000
```

The small P-values (0.000) give strong evidence that the return rate is related to the location where the letters were "lost." While expected and observed counts are fairly close in Des Moines, there were quite a few more letters returned (than expected) in Grinnell town, and the actual return rate was less than expected at Grinnell campus.

b. Since *Location* is a categorical variable with three categories, we use two indicator variables in the model and leave the third out as the reference group. The output below comes from fitting a logistic model for *Returned* based on indicators for *GrinnellTown* and *GrinnellCampus*.

```
Variable  Value  Count
Returned  1         82  (Event)
          0         58
          Total    140

Logistic Regression Table
                                                             Odds     95% CI
Predictor           Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant        0.251314  0.225374    1.12  0.265
GrinnellTown     1.13498  0.455019    2.49  0.013   3.11   1.28   7.59
GrinellCampus   -1.34993  0.563436   -2.40  0.017   0.26   0.09   0.78

Log-Likelihood = -86.088
Test that all slopes are zero: G = 17.771, DF = 2, P-Value = 0.000
```

Based on the small P-values of the individual t-tests for the coefficients, $GrinnellTown$ (0.013) and $GrinnellCampus$ (0.017), we have evidence that both of these coefficients are different from zero. Thus the return rate for lost letters in both of these locations would appear to differ from the reference location, Des Moines.

c. We'll do a test for the overall effectiveness of the model.

To assess the overall effectiveness of the model, we test $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 = 0$. This compares the constant model, $logit(\pi) = \beta_0$, to the model with both predictors, $logit(\pi) = \beta_0 + \beta_1 GrinnellTown + \beta_2 GrinnellCampus$.

From the output in part (b), we use the G-statistic (17.771) and compare it to a chi-square distribution with 2 degrees of freedom. The P-value in the output (0.000) indicates that the two indicators together really help predict whether or not letters are returned. Note that this chi-square statistic and P-value exactly match those of the likelihood ratio chi-square of the output in part (a).

d. Here is output for doing the chi-square analysis using just the letters addressed to Friends of the Confederacy.

```
Rows: Location   Columns: Returned
                          0      1    All

DesMoines                23     17     40
                      22.29  17.71  40.00

GrinnellCampus            9      1     10
                       5.57   4.43  10.00
```

```
GrinnellTown           7      13      20
                   11.14    8.86   20.00

All                   39      31      70
                   39.00   31.00   70.00

Pearson Chi-Square = 8.294, DF = 2, P-Value = 0.016
Likelihood Ratio Chi-Square = 9.176, DF = 2, P-Value = 0.010
```

Here is output for doing the chi-square analysis using just the letters addressed to Iowa Peaceworks.

```
Rows: Location   Columns: Returned
                       0       1     All

DesMoines             12      28      40
                   10.86   29.14   40.00

GrinnellCampus         6       4      10
                    2.71    7.29   10.00

GrinnellTown           1      19      20
                    5.43   14.57   20.00

All                   19      51      70
                   19.00   51.00   70.00

Pearson Chi-Square = 10.583, DF = 2, P-Value = 0.005
Likelihood Ratio Chi-Square = 11.584, DF = 2, P-Value = 0.003
```

The relationship between *Location* and *Return* is similar for both types of addresses. Both chi-square analyses show evidence of a relationship (likelihood ratio chi-square equal to 0.01 for Confederacy and 0.003 for Peaceworks). The nature of the relationship is also the same for both addresses: return rate close expected in Des Moines, less than expected in Grinnell campus, and higher than expected in Grinnell town.

e. To use both *Location* and *Address* we need to include indicators for two of the locations and one of the addresses. One such model is

$$logit(\pi) = \beta_0 + \beta_1 GrinnellTown + \beta_2 GrinnellCampus + \beta_3 Peaceworks$$

Following is some output from fitting this model.

```
Logistic Regression Table

                                                       Odds     95% CI
Predictor           Coef   SE Coef      Z       P  Ratio  Lower  Upper
Constant       -0.424347  0.299532  -1.42  0.157
GrinnellTown     1.25197  0.478586   2.62  0.009   3.50   1.37   8.94
GrinellCampus   -1.50895  0.596497  -2.53  0.011   0.22   0.07   0.71
Peaceworks       1.41655  0.398286   3.56  0.000   4.12   1.89   9.00

Log-Likelihood = -79.202
Test that all slopes are zero: G = 31.542, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             1.09921   2  0.577
Deviance            1.18619   2  0.553
Hosmer-Lemeshow     1.00640   3  0.800
```

We see that all coefficients are significant at a 5% level, so we can conclude that the return rate probably differs between Des Moines and Grinnell town, Des Moines and Grinnell campus, and the two addresses—even after accounting for the other factors in the model. By looking at the signs of the coefficients, we see that letters are more likely to be returned if lost in the town of Grinnell (compared to Des Moines), less likely in the Grinnell campus (compared to Des Moines), and more likely if addressed to Iowa Peaceworks (compared to Friends of the Confederacy).

From the G-statistic (31.542, P-value = 0.000), we can conclude that this model (as a whole) has some effectiveness at explaining whether or not lost letters are returned.

f. The logistic regression analysis allows us to test more specifically which groups have different return rates. It also allows us to include both explanatory factors (*Location* and *Address*) in the same model.

11.26 a. From the computer output, we find predicted logit values for each of the three groups:

$$\begin{aligned}
\text{None: } logit(\hat{\pi}) &= -1.119 + 0.783(0) - 0.336(0) = -1.119 \\
\text{Joke: } logit(\hat{\pi}) &= -1.119 + 0.783(1) - 0.336(0) = -0.336 \\
\text{Ad: } logit(\hat{\pi}) &= -1.119 + 0.783(0) - 0.336(1) = -1.455
\end{aligned}$$

If *Ad* is the omitted reference indicator, its logit value becomes the constant term, $\hat{\beta}_0 = -1.455$. The coefficients of *Joke* and *None* are the deviations needed to get to the same predicted logit values for those groups.

$\hat{\beta}_{Joke} = -0.336 - (-1.455) = 1.119 \qquad \hat{\beta}_{None} = -1.119 - (-1.455) = 0.336$

The fitted logit model for Ad and $None$ is $logit(\hat{\pi}) = -1.455 + 1.119Joke + 0.336None$.

b. If $Joke$ is the omitted reference indicator, its logit value becomes the constant term, $\hat{\beta}_0 = -0.336$. The coefficients of Ad and $None$ are the deviations needed to get to the same predicted logit values for those groups.

$\hat{\beta}_{Ad} = -1.455 - (-0.336) = -1.119 \qquad \hat{\beta}_{None} = -1.119 - (-0.336) = -0.783$

The fitted logit model for Ad and $None$ is $logit(\hat{\pi}) = -0.336 - 1.119Ad - 0.783None$.

11.27 a. Here is some output from fitting a logistic regression model to predict $Survived$ based on sex, where $SexMale = 1$ for males and 0 for females.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.69315    0.09869   7.023 2.17e-12
Sexmale      -2.30118    0.13488 -17.061  < 2e-16
```

The small P-value ≈ 0 gives strong evidence that there is some relationship between sex and survival. The negative coefficients indicates that the probability of survival is lower for males than it is for females. This is consistent with the proportions found in part (c).

b. Here is a table of expected counts based on a null hypothesis that Sex and $Survived$ are unrelated.

```
        Survived
Sex          0      1
  female 303.7  158.4
  male   559.3  291.7
```

The value of the chi-square statistic for comparing these expected counts to the observed counts from part(c) is $\chi^2 = 329.8$, which gives a P-value ≈ 0 when compared to a chi-square distribution with one degree of freedom. This confirms the strong evidence that $Titanic$ survival is related to Sex.

11.28 a. To see if THC is more effective, we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$. We can use either the P-value for the individual z-test for the coefficient of THC or the G-statistic (both

P-values = 0.001). However, those are two-tailed tests, so the P-value for one tail is about 0.0005. This is a very small P-value, providing strong evidence that THC is more effective than prochlorperazine at preventing nausea with chemotherapy.

b. If we test $H_0 : p_{THC} = p_{Pro}$ versus $H_a : p_{THC} > p_{Pro}$ using a two-sample z-test, we use the sample proportions from part (a) and the pooled proportion $\hat{p} = (36+16)/(79+78) = 0.331$. The z-test statistic is

$$z = \frac{0.456 - 0.205}{\sqrt{0.331(1-0.331)\left(\frac{1}{79}+\frac{1}{78}\right)}} = 3.34$$

The P-value is the area in the upper tail of a standard normal, beyond 3.34, so P-value = 0.0004. This is a very small P-value, so we have evidence that the proportion of patients helped by THC is higher than the proportion helped by prochlorperazine.

c. Here is some output with observed and expected counts for a chi-quare test of the relationship between effectiveness and which drug is used.

```
                   Effective  NotEffective  Total
     THC             36           43          79
                     26.17        52.83

Prochlorperazine     16           62          78
                     25.83        52.17

      Total          52          105         157

Chi-Sq = 11.124, DF = 1, P-Value = 0.001
```

The P-value given in the output should be cut in half, so the P-value is 0.0005, which again demonstrates a relationship between the treatment drug and effectiveness. We see from the expected to observed counts to see that THC is the more effective drug.

d. These tests are consistent with each other, but not exactly the same. Up to round-off differences, the chi-square statistic in part (c) is the square of the z-statistic in part (b) and those P-values should match exactly. The z-statistic for the THC coefficient and chi-square G-statistic in the logistic regression are both close to these, but not exactly the same.

11.29 a. A two-way table of survival by age group is shown below, along with the proportion surviving in each group.

```
        AgeGroup
Survive  1  2  3
      0  5 17 18
      1 54 60 46
```

```
                           Age group
                        1      2      3
Proportion survive   0.915  0.779   0.719
```

The decreasing proportion surviving with increasing age seems to suggest that older people have lower odds of survival. However, the probability of survival for the second and third age groups do not appear very different from one another (probabilities of survival are 0.78 and 0.72, respectively). Those survival rates do appear quite a bit lower than the youngest age group, where the probability of survival is over 90%.

b. Here is some output for a logistic regression model to predict *Survive* using *AgeGroup* as a quantitative predictor.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.7566     0.5732   4.809 1.52e-06 ***
AgeGroup     -0.6399     0.2414  -2.651  0.00802 **

Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.66  on 198  degrees of freedom
AIC: 196.66
```

The estimated proportions surviving for each group based on this model are

$$\begin{aligned}
\hat{\pi}_1 &= \frac{e^{2.7566-0.6399(1)}}{1+e^{2.7566-0.6399(1)}} = \frac{8.30}{1+8.30} = 0.893\\
\hat{\pi}_2 &= \frac{e^{2.7566-0.6399(2)}}{1+e^{2.7566-0.6399(2)}} = \frac{4.38}{1+4.38} = 0.814\\
\hat{\pi}_3 &= \frac{e^{2.7566-0.6399(3)}}{1+e^{2.7566-0.6399(3)}} = \frac{2.31}{1+2.31} = 0.698
\end{aligned}$$

c. The estimated log(odds) for surviving in each age group are

$$\begin{aligned}
\log(\widehat{odds}_1) &= 2.7566 - 0.6399(1) = 2.117\\
\log(\widehat{odds}_2) &= 2.7566 - 0.6399(2) = 1.477\\
\log(\widehat{odds}_3) &= 2.7566 - 0.6399(3) = 0.837
\end{aligned}$$

The predicted log(odds) for survival in the middle age group is 0.6399 less the odds of survival of the younger age group and 0.6399 more than the older age group because for each age group unit increase, the log(odds) decreases by 0.6399.

The observed log(odds) for the middle group is log(0.779/(1 − 0.779)), or 1.26. The log(odds) for the younger group is log(0.915/(1 − 0.915)), or 2.38, which is 1.12 greater than than the log(odds) for the middle age group. The difference in these log(odds) is predicted to be 0.64 but is actually about twice that.

Note that each of the calculations above could have been more directly computed using odds instead of probabilities. The two approaches are mathematically equivalent, but there would be less round-off error using the odds as opposed to the probabilities. Here, with odds we have:

	notes	Youngest	Middle	Oldest
odds of survival	successes/failures	10.80	3.53	2.56
observed log(odds)	aka empirical logits	2.38	1.26	0.94
predicted log(odds)	`predict` cmd output	2.12	1.48	0.84
observed probabilities	from table	0.915	0.779	0.719
predicted probabilities	from model	0.893	0.814	0.698

The log(odds) for the oldest age group is log(0.719/(1 − 0.719)), or 0.940. This is about 0.32 less than the middle group and half of what the model predicts. It seems that the model does not fit all that well, which we could have anticipated by looking at sample odds.

In terms of logits, the model predictions are low for the youngest group, high for the middle group and low for the oldest group.

d. Model with age groups as a factor, that is, each category gets an indicator variable and two of the three indicator variables are included in a multiple logistic regression model to predict *Survive*.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.3795     0.4675   5.090 3.57e-07
factor(AgeGroup)2  -1.1184     0.5422  -2.063  0.03915
factor(AgeGroup)3  -1.4413     0.5439  -2.650  0.00805

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 191.59  on 197  degrees of freedom
AIC: 197.59
```

This model provides a separate estimate of log(odds) for each age group. Each estimated coefficient for the age group variables compares the log(odds) of survival of that age group to the youngest age group, the one that was left out of the model. So the odds of survival for someone in the middle age group is $e^{-1.12} = 0.327$, or about a third of the odds of someone from the younger group surviving. Similarly, the odds of survival for someone in the older age group is $e^{-1.44} = 0.237$, or about a quarter of the odds of someone from the younger group surviving.

e. The survival proportions estimated for each age group using the model in (d) are equal to the sample proportions found in part (a).

$$\begin{aligned}
\hat{\pi}_1 &= \frac{e^{2.3795}}{1+e^{2.3795}} = \frac{10.80}{1+10.80} = 0.915 \\
\hat{\pi}_2 &= \frac{e^{2.3795-1.1184(1)}}{1+e^{2.3795-1.1184(1)}} = \frac{3.53}{1+3.53} = 0.779 \\
\hat{\pi}_3 &= \frac{e^{2.3795-1.4413(1)}}{1+e^{2.3795-1.4413(1)}} = \frac{2.56}{1+2.56} = 0.719
\end{aligned}$$

f. A chi-square test on the two-way table from part (a) gives the following output.

```
Pearson's Chi-squared test
X-squared = 7.7467, df = 2, p-value = 0.02079
```

This small P-value gives evidence that the age groups is related to survival.

To test the overall effectiveness of the logistic regression model in part (d), we find the change in the residual deviance from from the null model.

$$G = 200.16 - 191.59 = 8.57$$

We find a P-value using the upper tail of a chi-square distribution with 2 degrees of freedom, $P(\chi^2_2 > 8.57) = 0.0138$. The results of these are consistent although not identical. Both suggest that the odds of survival is significantly associated with age.

g. A logistic model using the actual ages in the **ICU** data set, rather than the three age groups.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.05851    0.69608   4.394 1.11e-05 ***
Age         -0.02754    0.01056  -2.607  0.00913 **

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.31  on 198  degrees of freedom
AIC: 196.31
```

Note that the model with Age in years is not nested nor does it nest any of the other models. Thus we'll take a look at AICs.

Age categories numeric	196.66
Age category indicators	197.59
Age in years	196.3

Because none of these models is overwhelmingly preferred, other considerations may come to bear. Categorical results may be easier to interpret and convey to others. On the other hand, summarizing into three categories loses a lot of information. It is helpful to reflect on the purpose of this model. Will it be used solely for prediction and will the age in years always be available? If so, it seems that the model with age in years would be preferred.

	White def	Black def
Death penalty	19	17
No DP	141	149
Total	160	166

11.30 a.

b. When the victim is white the DP percentage for white defendants is 12.6% (19/151), which is lower than the DP percentage for black defendants, which is 17.5% (11/63). When the victim is black the DP percentage for white defendants is 0% (0/9), which is lower than the DP percentage for black defendants, which is 5.8% (6/103). In the aggregate table from part (a) the DP percentage for white defendants is 11.9% (19/160), which is greater than the DP percentage for black defendants, which is 10.2% (17/166).

c. White defendants tend to have white victims and having a white victim makes the death penalty much more likely. (For the white victim table the DP percentage is 14.0% (30/214), which is much higher than the DP percentage for the black victim table of 5.4% (6/112).)

	Smoker	Nonsmoker
Alive	443	502
Dead	139	230
Total	582	732

11.31 a.

b. When the woman is young, the percentage alive for smokers is 82.1% (437/532), which is lower than the percentage alive for nonsmokers, which is 87.9% (474/539). When the woman is old, the percentage alive for smokers is 12.0% (6/50), which is lower than the percentage alive for nonsmokers, which is 14.5% (28/193). In the aggregate table from part (a) the percentage alive for smokers is 76.1% (443/582), which is greater than the percentage alive for nonsmokers, which is 68.6% (502/732).

c. Smoking is associated with age, with younger women more likely to have been smokers at baseline and more likely to be alive after 20 years. (For the 18–64 table the percentage alive is 85.1% (911/1071), which is much higher than the percentage alive for the 65+ table of 14.0% (34/243).)

11.32 a. The output is below.

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.842      0.420   -6.76  1.4e-11 ***
White.Victim      1.324      0.519    2.55    0.011 *
DefendantWhite   -0.440      0.401   -1.10    0.272
```

White.Victim has a positive coefficient, which means that, controlling for race of the defendant, the probability that the death penalty will be imposed is higher when the victim is white than when the victim is black.

b. *DefendantWhite* has a negative coefficient, which means that, controlling for race of the victim, the probability that the death penalty will be imposed is lower when the defendant is white than when the defendant is black.

c. The magnitude of the z-value for *White.Victim* is much greater than for *DefendantWhite*, so *White.Victim* has the stronger relationship with the response variable.

11.33 a. The output is given below.

```
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)    1.971      0.127   15.46   <2e-16 ***
AgeGroup65+   -3.709      0.215  -17.23   <2e-16 ***
SmokerYes     -0.434      0.164   -2.64   0.0083 **
```

AgeGroup65+ has a negative coefficient, which means that, controlling for smoking status, the probability of being alive after 20 years is lower in the older age group (those who are 65+) than in the younger age group.

b. *Smoker* has a negative coefficient, which means that, controlling for age group, the probability of being alive after 20 years is lower for smokers than nonsmokers.

c. The magnitude of the z-value for *AgeGroup* is much greater than for *Smoker*, so *AgeGroup* has the stronger relationship with the response variable.

11.34 a. White: athlete graduation rate= $1088/1747 = 0.623$, nonathlete graduation rate = $112,906/188,176 = 0.600$. Black: athlete graduation rate = $970/2076 = 0.467$, nonathlete graduation rate = $7895/22,556 = 0.350$. For each race, the graduation rate is higher for athletes.

b. The athlete graduation rate is $2058/3823 = 0.538$; the nonathlete graduation rate is $120,801/210,732 = 0.573$. The graduation rate is higher for non-athletes.

c. Athletes are more likely to be black and blacks have a lower graduation rate, which makes the overall graduation rate for athletes low, despite the fact that for each race athletes have a higher graduation rate than nonathletes.

11.35 a. 2-point shots: Irving Hit rate $= 494/979 = 0.505$, Frye Hit rate $= 101/185 = 0.546$. 3-point shots: Irving Hit rate $= 177/441 = 0.401$, Frye Hit rate $= 137/335 = 0.409$. For each type of shot Frye had the higher Hit rate.

b. The Hit percentage for Irving is $671/1420 = 0.473$; the Hit percentage for Frye is $238/520 = 0.458$. Irving had the higher overall Hit percentage.

c. A large proportion of Frye's shots were 3-point shots (64.4% versus only 31.1% for Irving) and such shots have a lower Hit rate, which makes Frye's overall Hit percentage low, despite the fact that he did better than Irving for each type of shot.

11.36 a. The output is as follows:

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.1536     0.0324    4.73  2.2e-06 ***
StudentNonAthlete     0.1415     0.0327    4.32  1.5e-05 ***
```

We see that *StudentNonAthlete* has a positive coefficient, which means that the probability that a nonathlete graduates within 6 years is higher than the probability for an athlete.

b. The output is as follows:

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.2946     0.0340   -8.66   <2e-16 ***
RaceWhite             1.0063     0.0144   70.06   <2e-16 ***
StudentNonAthlete    -0.3080     0.0344   -8.95   <2e-16 ***
```

We see that *RaceWhite* has a positive coefficient, which means that, controlling for Student, the probability of graduating within 6 years is higher for whites than for blacks.

c. We see that *StudentNonAthlete* has a negative coefficient, which means that, controlling for race, the probability of graduating within 6 years for a nonathlete than for an athlete.

d. The magnitude of the z-value for *RaceWhite* is much greater than for *StudentNonAthlete*, so *Race* has a stronger relationship with the response variable.

11.37 a. The output is as follows:

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.1696     0.0880   -1.93    0.054 .
PlayerIrving   0.0597     0.1028    0.58    0.562
```

We see that *PlayerIrving* has a positive coefficient, which means that the probability that Irving makes a shot is higher than the probability that Frye makes a shot.

b. The output is as follows:

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.3335     0.0954   -3.50  0.00047 ***
ShotTypeTwo    0.4564     0.0985    4.63  3.6e-06 ***
PlayerIrving  -0.0930     0.1086   -0.86  0.39209
```

We see that *ShotTypeTwo* has a positive coefficient, which means that, controlling for *Player*, the probability of hitting a shot is higher for 2-point shots than for 3-point shots.

c. We see that *PlayerIrving* has a negative coefficient, which means that, controlling type of shot, the probability that Irving makes a shot is lower than the probability that Frye makes a shot.

d. The magnitude of the z-value for *ShotTypeTwo* is much greater than for *PlayerIrving*, so *ShotType* has a much stronger relationship with the response variable.