7.6 (2×2 **contingency table):**

(a) The sample log odds ratio equals

$$\ln \widehat{\psi} = \ln \left(\frac{38/4}{14/7} \right) = \ln(4.75) = 1.558.$$

The $\widehat{\mathrm{Var}}(\widehat{\psi})$ formula simplifies to

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\psi}) &= \frac{1}{n_0\widehat{p}_0} + \frac{1}{n_0\widehat{q}_0} + \frac{1}{n_1\widehat{p}_1} + \frac{1}{n_1\widehat{q}_1} \\
&= \frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}} \\
&= \frac{1}{14} + \frac{1}{7} + \frac{1}{38} + \frac{1}{4} \\
&= 0.4906.
\end{aligned}$$

Hence the test statistic for testing $H_0 : \psi = 1$ or $H_0 : \ln \psi = 0$ equals

$$\frac{1.558}{\sqrt{0.4906}} = 2.224,$$

which is significant at $\alpha = 0.05$.

(b) Substitute in the test statistic
$\widehat{p}_0 = 14/21 = 0.6667, \widehat{p}_1 = 38/42 = 0.9047$ and
$\widehat{p} = (s_0 + s_1)/(n_0 + n_1) = 52/63 = 0.8254$. Thus

$$z = \frac{0.9047 - 0.6667}{\sqrt{0.8254 \times 0.1746(1/21 + 1/42)}} = 2.346.$$

7.7 (Simpson's Paradox):

(a) Simpson's paradox occurs for these data because men apply in large numbers do departments A and B which have high admit rates whereas women apply in small numbers to these departments. On the other hand, women apply in large numbers to departments C-F which have low admit rates.

(b) The sample log odds ratio for men vs. women is

$$\ln\left[\frac{(1198/1493)}{(557/1278)}\right] = 0.61035.$$

(c) The logistic regression model using only Gender as predictor is as shown below.

```
Call:
glm(formula = Admit ~ Gender, family = binomial, data = ucb)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83049    0.05077 -16.357   <2e-16 ***
Gender       0.61035    0.06389   9.553   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5950.9  on 4524  degrees of freedom
AIC: 5954.9

Number of Fisher Scoring iterations: 4
```

Note that $\widehat{\beta}_1 = 0.61035$ is exactly the sample log odds ratio. Also note that $\widehat{\beta}_1 > 0$.

(d) Adding Dept.A-Dept.E as predictors (Dept.F is taken as reference Dept.) we get the following logistic regression model.

```
Call:
glm(formula = Admit ~ Gender + Dept.A + Dept.B + Dept.C
            + Dept.D + Dept.E, family = binomial, data = ucb)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.62456    0.15773 -16.640   <2e-16 ***
Gender      -0.09987    0.08085  -1.235    0.217
Dept.A       3.30648    0.16998  19.452   <2e-16 ***
Dept.B       3.26308    0.17878  18.252   <2e-16 ***
Dept.C        2.04388    0.16787  12.176   <2e-16 ***
Dept.D        2.01187    0.16992  11.840   <2e-16 ***
Dept.E        1.56717    0.18044   8.686   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6044.3  on 4525  degrees of freedom
Residual deviance: 5187.5  on 4519  degrees of freedom
AIC: 5201.5

Number of Fisher Scoring iterations: 5
```

Note now the Gender coefficient $\widehat{\beta}_1 < 0$. After taking the Dept. effect into account the log odds ratio for men vs. women is now negative, that is, after adjusting for Dept. men are admitted at a lower rate than women. This illustrates Simpson's paradox.

7.8 (Radiation therapy):

(a) The fitted logistic response model is shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.81944    1.83518   2.081   0.0374 *
Days        -0.08648    0.04322  -2.001   0.0454 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32.601  on 23  degrees of freedom
Residual deviance: 27.788  on 22  degrees of freedom
AIC: 31.788

Number of Fisher Scoring iterations: 4
```

As the number of days of exposure increases the odds of success (absence of tumor) decrease. This may seem counterintuitive but the amount of radiation exposure is fixed and if it is spread over more days then its effect is diluted.

(b) The difference in the log-odds of absence of tumor vs. presence of tumor if the number of days of exposure is increased by 5 days is

$$\ln\psi(x+5) - \ln\psi(x) = \beta_0 + \beta_1(x+5) - \beta_0 - \beta_1 x = 5\beta_1.$$

The 95% CI for $5\beta_1$ is

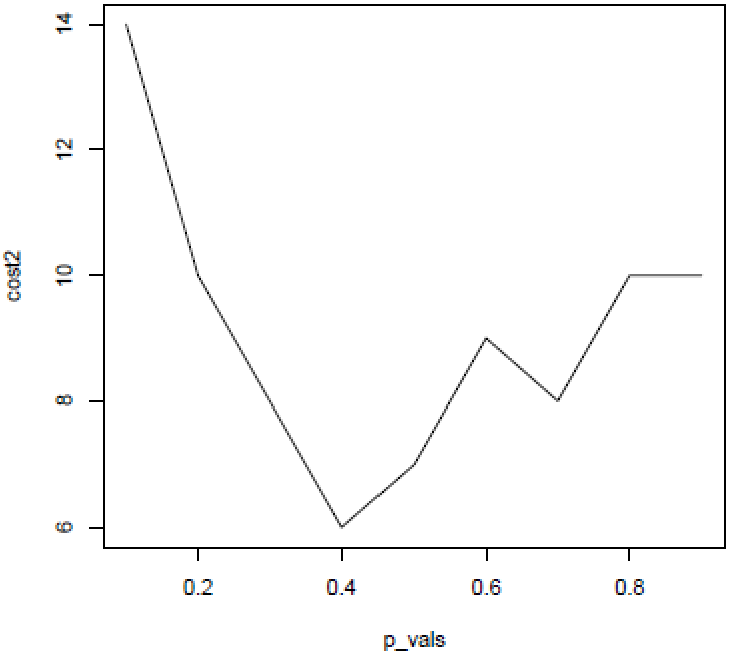
$$\begin{aligned} 5\widehat{\beta}_1 \pm 1.96\text{SE}(5\widehat{\beta}_1) &= -5(0.0865) \pm 1.96 \times 5 \times 0.0432 \\ &= -0.4325 \pm 0.4236 \\ &= [-0.8558, -.0086]. \end{aligned}$$

Hence the 95% CI on the odds of success is

$$[\exp(-0.08558), \exp(-0.0086)] = [0.4249, 0.9914].$$

Hence the odds of success are < 1 if the number of days of exposure is increased by 5 days.

(c) A plot of total misclassifications versus p^* is shown below. (Here costs for a false positive and a false negative are both assumed to be 1, so the total cost = total ♯ of misclassifications). Note that misclassifications are minimized and hence CCR is maximized when $p^* = 0.40$.



The confusion matrix using $p^* = 0.40$ is shown below. We see that the maximum CCR = 0.625.

```
    FALSE TRUE
  0     4    6
  1     3   11
> CCR=sum(diag(tab))/sum(tab)
> CCR
[1] 0.625
```

From this table we can calculate

$$\text{Sensitivity} = \frac{11}{14} = 0.7857, \text{Specificity} = \frac{4}{10} = 0.4000,$$
$$\text{Precision} = P = \frac{11}{17} = 0.6471, \text{Recall} = R = \frac{11}{14} = 0.7857,$$
$$F_1\text{-Score} = \left[\frac{1}{2}\left(\frac{1}{0.6471} + \frac{1}{0.7857}\right)\right]^{-1} = 0.710.$$

Thus this model is good for predicting successes but not for predicting failures.

7.10 (Measures of accuracy of classification):

Table 1:

$$\text{CCR} = \tfrac{9000}{10{,}000} = 0.90, \text{ Sensitivity} = \tfrac{900}{1000} = 0.90, \text{ Specificity} = \tfrac{8100}{9000} = 0.90,$$
$$\text{Precision} = \tfrac{900}{1800} = 0.50, \text{ Recall} = \tfrac{9000}{10{,}000} = 0.90,$$
$$F_1 = \tfrac{2\times 0.50\times 0.90}{0.50+0.90} = 0.543.$$

Table 2:

$$\text{CCR} = \tfrac{9000}{10{,}000} = 0.90, \text{ Sensitivity} = \tfrac{90}{100} = 0.90, \text{ Specificity} = \tfrac{8910}{9900} = 0.90,$$

$$\text{Precision} = \tfrac{90}{1080} = 0.083, \text{ Recall} = \tfrac{9000}{10{,}000} = 0.90,$$
$$F_1 = \tfrac{2\times 0.083\times 0.90}{0.083+0.90} = 0.152.$$

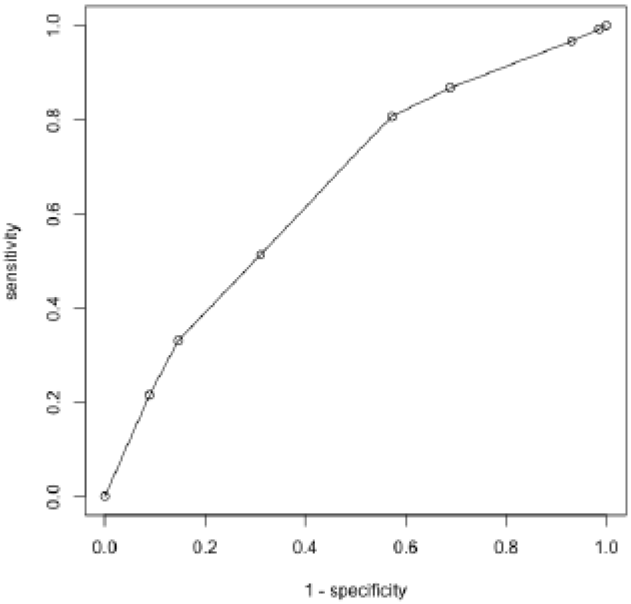
In Table 2, the percentage of relevant items is only $100/10,000 = 1\%$ while in Table 1 it is $1000/10,000 = 10\%$. That is the reason why the precision and hence F_1-score are much lower in Table 2 than in Table 1 even though other metrics are the same. Rare traits are difficult to detect accurately.

7.11 (ROC curve for simple logistic regression model for the art museum visits data):

(a) The table is as follows. The $\widehat{p}_i$ values are actually not needed for plotting the ROC curve but they are included in the table and are taken from the R output in the same example. Sensitivity is calculated as the proportion of Yes responses that are correctly classified as Yes. 1−Specificity is calculated as the proportion of No responses that are incorrectly classified as Yes. If we start with $p^* = 0$ then all Yes responses are correctly classified as Yes but all No responses are incorrectly classified as Yes. So Sensitivity = 1 and Specificity = 0 or 1−Specificity = 1. Once p^* increases just past 0.1213, 7 Yes responses are incorrectly classified as No, so Sensitivity drops to $(925 - 7)/925 = 0.9924$ but 23 No responses are correctly classified as No and hence Specificity increases by $24/1682 = 0.0143$ or 1−Specificity decreases to $1 - 0.0143 = 0.9857$.

	Visit				
Education	Yes	No	$\widehat{p}_i$	Sensitivity	1−Specificity
1	7	24	0.1213	1.0000	1.0000
2	24	92	0.1608	0.9924	0.9857
3	92	408	0.2102	0.9665	0.9310
4	53	196	0.2698	0.8670	0.6885
5	271	439	0.3390	0.8065	0.5719
6	172	277	0.4159	0.5135	0.3109
7	107	96	0.4972	0.3308	0.1463
8	199	150	0.5786	0.2151	0.0892
Total	925	1682	2607		

(b) To plot the ROC curve we first plot the 8 points Sensitivity vs. 1−Specificity and connect them by straight lines. This implies linear interpolation between the successive plotted points instead of step changes because the data are grouped, so within each group of the people with a given Education level, both quantities are assumed to change linearly. The resulting ROC curve is shown in the figure below, which exactly matches the ROC curve in Figure 7.6 (a). The AUC for this curve can be calculated to be 0.6559.


sensitivity
1 - specificity