Chapter 7: Comparing proportions and distributions

7.1 a) H_0: The 2019 distribution is equal to the historical distribution
H_1: The 2019 distribution is not equal to the historical distribution

7.1 b) Expected number = Historical % × Observed number

	Historical	Observed (O)	Expected (E)	$O - E$	$\frac{(O-E)^2}{E}$
Spring	4.87 (30%)	2.90	25.2	11.8	5.525
Summer	3.04 (40%)	4.50	33.6	-6.6	1.296
Fall	1.65 (15%)	4.94	12.6	-0.6	0.029
Winter	2.88 (15%)	3.28	12.6	-4.6	1.679
Total	2.31 (100%)	4.73	84		8.530

7.1 c) Since every season has an expected value above 5 you can do a chi-square test.

7.1 d) $X^2 = 8.530$

7.1 e) The calculated chi-square value of 8.530 is higher than the critical chi-square value of 7.8. H_0 is rejected. The 2019 distribution is significantly different from the historical distribution. There is a 5% risk of a type-I error.

7.1 f) You rejected the null hypothesis H_0 with 95% confidence, and so the p-value must be lower than 0.05.

7.1 g) $X^2 = 8.530$

```
Observed <- c(37, 27, 12, 8)
Historical <- c(0.3, 0.4, 0.15, 0.15)
Expected <- c(25.2, 33.6, 12.6, 12.6)

chi <- sum((Observed - Expected)^2 / Expected) # 8.530
```

7.1 h) Yes.

```
qchisq(p = 0.95, df = 3) # 7.185
```

7.1 i) R returns the same chi-squared as calculated, so the answer was correct. It also shows a p-value of below 0.05, as expected.

```
# chi-square test: x = observations p = model distribution
# rescale makes sure the model distribution adds up to 100%
chisq.test(x = Observed, p = Historical, rescale.p = TRUE)

# Chi-squared value: 8.5298
# p-value: 0.0362
```

7.1 j) The expected values are 25.2, 33.6, 12.6, and 12.6. R shows the same expected values.

```
chisq <- chisq.test(x = Observed, p = Historical)
chisq$expected # Extract expected values with $expected
# 25.2 33.6 12.6 12.6
```

7.2 a) The **sales** data frame contains 3 columns: **month**, **historical** and **newstore**. It contains the 'Historical' and 'New Store' distribution of sales over the months.

```
# These are the values for the sales data set
sales <- data.frame(month = seq(from = 1, to = 12, by = 1),
                    historical = c(5.1, 5.1, 6.7, 10, 11.4, 10,
                                   6.7, 5.1, 6.7, 10, 11.7, 11.7),
                    newstore = c(5.6, 6.2, 9.4, 8.6, 6.8, 4.8,
                                 5.6, 4.8, 8.8, 12.6, 13.1, 13.7))

summary(sales)
```

7.2 b) The chi-square test requires every cell in the expected distribution to have a value of at least 5 and it requires the total of both groups to be equal. Since the historical distribution contains more than 5 in every cell and both observed and expected values add up to the same number (100) we can use this for a chi-squared test.

7.2 c) H_0: The new distribution is equal to the historical distribution
H_1: The new distribution is not equal to the historical distribution

7.2 d) The p-value of 0.6963 is higher than the critical p-value of 0.10. H_0 is not rejected. The new store distribution is not significantly different from the historical distribution. There is a risk a type-II error.

```
chisq.test(x = sales$newstore, p = sales$historical, rescale.p = TRUE)
# p-value: 0.6963
```

7.3 a) The best estimate of the population proportion π is the sample proportion p.

$\pi_1 = \frac{k}{n} = \frac{8}{71} = 0.113$

$\pi_2 = \frac{k}{n} = \frac{16}{111} = 0.144$

7.3 b) Confidence interval sample 1:

$p \pm z_\alpha \times \sqrt{\frac{p(1-p)}{n}} = 0.113 \pm 1.960 \times \sqrt{\frac{0.113 \times (1-0.113)}{71}} = [0.039;\ 0.187]$

Confidence interval sample 2:

$p \pm z_\alpha \times \sqrt{\frac{p(1-p)}{n}} = 0.144 \pm 1.960 \times \sqrt{\frac{0.144 \times (1-0.144)}{111}} = [0.078;\ 0.210]$

7.3 c) H_0: $\pi_2 \leq \pi_1$ H_1: $\pi_2 > \pi_1$

Where π_2 and π_1 are the success proportions for the evening and afternoon calls respectively.

7.3 d) Combined success probability:

$p^* = \frac{k_1 + k_2}{n_1 + n_2} = \frac{8+16}{71+111} = 0.132$

7.3 e) Combined standard error:

$s_p = \sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})} = \sqrt{0.132(1-0.132)(\frac{1}{71} + \frac{1}{111})} = 0.0514$

7.3 f) z-score: $\frac{p_1 - p_2}{s_p} = \frac{0.144 - 0.113}{0.514} = 0.612$

Note that p_1 and p_2 were switched because in the hypotheses π_1 and π_2 were also switched.

7.3 g) The calculated z-score of 0.612 is lower than than the critical z-value of 1.645. H_0 is not rejected. The evening success rate is not shown to be significantly higher than the afternoon success rate. There is a risk of a type-II error.

7.3 h) The code creates a vector of successes `k` and a vector of sample sizes `n`. The proportion test then tests the equality. It shows the proportions you calculated earlier, and a p-value of 0.6984 which supports your conclusion if you do not reject H_0.

```
n <- c(71, 111)
k <- c(8, 16)
prop.test(x = k, n = n)
# p-value: 0.6984
```

Note that R actually returns a chi-squared value. Because it actually does a chi-square test it can in fact be used to test more than two proportions at the same time.