Solutions for 3.9 Exercises

1. **Drink driving**

   We have 189 'yes' answers and 111 'no' or 'unsure' answers.

```
> n <- 300
> (p.hat <- 189/n)

[1] 0.63

> (n.tilde <- n + 4)

[1] 304

> (p.tilde <- (189 + 2)/n.tilde)

[1] 0.6282895

> round(ci <- p.tilde + c(-1.96, 1.96) * sqrt(p.tilde * (1 - p.tilde)/n.tilde),
+     4)

[1] 0.5740 0.6826
```

   Using the `round()` function is not essential, but it is useful in presenting the proportions to a manageable number of decimal places.

2. **Mendel's peas**

   Enter the data:

```
> peas <- c(rep("A-RY", 315), rep("B-WY", 101), rep("C-RG", 108),
+     rep("D-WG", 32))
> table(peas)

peas
A-RY B-WY C-RG D-WG
 315  101  108   32
```

   Get point estimates of the proportions in each category:

```
> p.peas <- table(peas)/length(peas)
> round(p.peas, 3)

peas
 A-RY  B-WY  C-RG  D-WG
0.567 0.182 0.194 0.058
```

   The data appear to match the theory very well—too well, it has been suggested. The ratio of 9:3:3:1 corresponds to proportions of 0.56, 0.19, 0.19 and 0.06.

   Assumptions:

   - n is large enough – this looks ok, ≥ 5 in each category.
   - observations are independent – we would want to check the experimental design – for example, how were the peas/plants selected?
   - the sample is representative of the population — again, we would need to know something about how the data were collected to have confidence in this.

3. **Blood groups**

```
> group <- c(rep("O", 51), rep("A", 38), rep("B", 9), rep("AB",
+     2))
> table(group)

group
 A AB  B  O
38  2  9 51

> p.hat <- table(group)/length(group)
> p.hat

group
   A   AB    B    O
0.38 0.02 0.09 0.51
```

Wald:

```
> p.hat.AB <- p.hat[2]
> p.hat.AB + c(-1, 1) * 1.96 * sqrt(p.hat.AB * (1 - p.hat.AB)/length(group))

[1] -0.00744  0.04744
```

Jeffrey prior:

```
> qbeta(c(0.025, 0.975), 2 + 0.5, 100 - 2 + 0.5)

[1] 0.004178812 0.062605849
```

Jeffrey prior confidence interval is more appropriate because $\hat{p}$ is close to 0.

Assumptions:

- There are ≤ 5 students for blood group AB – not great for the normal approximation.
- The sample may not comprise independent students; we would hope that not all students are from one subject or friendship group, for instance.
- We would need to know more about the sampling design to make a judgment on this.

The data appear to be consistent with the claim. However, this data set may not relate well to the claim because it is probably not a representative sample of the Australian population or the world.

4. **Kitchen appliance preferences**

   (a) First we enter the data. Note that we will use "A-White", etc. to keep the ordering of the categories the same as in the table and not alphabetical.

```
> finish <- c(rep("A-White", 70), rep("B-Steel", 130), rep("A-White",
+     48), rep("B-Steel", 52))
> product <- c(rep("C-Fridge", 200), rep("D-Washer", 100))
> goods <- data.frame(product, finish)
> table(goods)

          finish
product    A-White B-Steel
  C-Fridge      70     130
  D-Washer      48      52

> table(goods$finish)
```

```
A-White B-Steel
    118     182
```

(b) Now find the proportions and confidence intervals – note that the question is really just about the finish.

```
> p.finish <- table(goods$finish)/length(goods$finish)
> round(p.finish,3)
A-White B-Steel
  0.393   0.607
> n.tilde <- length(goods$finish) + 4
> p.tilde <- (table(goods$finish) + 2)/n.tilde
> # lower confidence limit
> lcl <- p.tilde - 1.96 * sqrt(p.tilde * (1 - p.tilde)/n.tilde)
> round(lcl,3)
A-White B-Steel
   0.34    0.55
> # upper confidence limit
> ucl <- p.tilde + 1.96 * sqrt(p.tilde * (1 - p.tilde)/n.tilde)
> round(ucl,3)
A-White B-Steel
   0.45    0.66
```

So our point estimate of the proportion who prefer a stainless steel finish is 0.607, and our Agresti-Coull 95% confidence interval for this is (0.55, 0.66).

Assumptions:

- n is large enough – this looks ok.
- Observations are independent — if the purchases represented were all made by readers of one magazine this assumption would not hold. We would need to find out how the data were collected.
- The sample is representative of the population — if the data were all from the same shop this assumption would not hold. How were the data collected?

Note that drawing conclusions about preference from this data could be dubious, given that other factors like price and availability might influence buying decisions. But taking the data on face value, it would seem to contradict the claim by *Choice*, since 0.7 (70%) is outside our confidence interval for p.

(c)
```
> p.steel <- table(goods)[, 2]/(table(goods)[, 1] + table(goods)[,
+     2])
> p.steel
C-Fridge D-Washer
    0.65     0.52
> p.diff <- p.steel[1] - p.steel[2]
> s2.steel <- p.steel * (1 - p.steel)/(table(goods)[, 1] + table(goods)[,
+     2])
> var.diff <- sum(s2.steel)
> ci.diff <- p.diff + qnorm(c(0.025, 0.975)) * sqrt(var.diff)
> round(ci.diff, 3)
[1] 0.012 0.248
```

So our estimate of the difference is 0.13 and a confidence interval for this is given by (0.012, 0.248). Note that the difference is fairly large and that the confidence interval for the difference excludes 0, leading us to conclude that there is a real difference (though the interval is also quite wide, and only just excludes 0).

Note that we could have used `c(-1.96,1.96)` instead of `qnorm(c(0.025,0.975))` for the normal distribution quantiles.

5. **Wallet on the street**

(a)
```
> sex <- c("Female", "Male")
> n <- c(93, 75)
> yes <- c(84, 53)
> p.hat <- round(yes/n, 3)
> n.tilde <- n + 4
> p.tilde <- (yes + 2)/n.tilde
> lcl <- c(-1, -1)
> ucl <- c(1, 1)
> ci <- round(p.tilde + cbind(lcl, ucl) * 1.96 * sqrt(p.tilde *
+     (1 - p.tilde)/n.tilde), 3)
> data.frame(sex, p.hat, ci)
```

```
     sex p.hat   lcl   ucl
1 Female 0.903 0.823 0.950
2   Male 0.707 0.595 0.798
```

The R function `cbind()` is "column bind" and joins the columns together into an array. The final steps of the R code above are not necessary for calculation, but useful for presentation.

(b)
```
> p <- yes/n
> s2.p <- p * (1 - p)/n
> difference <- p[1] - p[2]
> s2.diff <- sum(s2.p)
> round(ci <- difference + qnorm(c(0.025, 0.975)) * sqrt(s2.diff),
+     3)
```

```
[1] 0.077 0.316
```

(c) We are 95% confident that the true difference in the proportions of college female and male students is between 0.077 and 0.316. A zero difference is therefore not very plausible, and so it reasonable to conclude that the proportion is larger for females.

6. **Aspirin and stroke**

(a) This study shows some good design features. Patients are randomised into treatment and control groups to prevent bias. The study was double-blind, helping to ensure impartiality.

(b)
```
> treat <- c("aspirin", "control")
> n <- c(78, 77)
> yes <- c(63, 43)
> p.hat <- round(yes/n, 3)
> data.frame(treat, p.hat)
```

```
    treat p.hat
1 aspirin 0.808
2 control 0.558
```

(c)
```
> p <- yes/n
> s2.p <- p * (1 - p)/n
> difference <- p[1] - p[2]
> s2.diff <- sum(s2.p)
> round(ci <- difference + qnorm(c(0.025, 0.975)) * sqrt(s2.diff),
+     3)
```

```
[1] 0.108 0.391
```