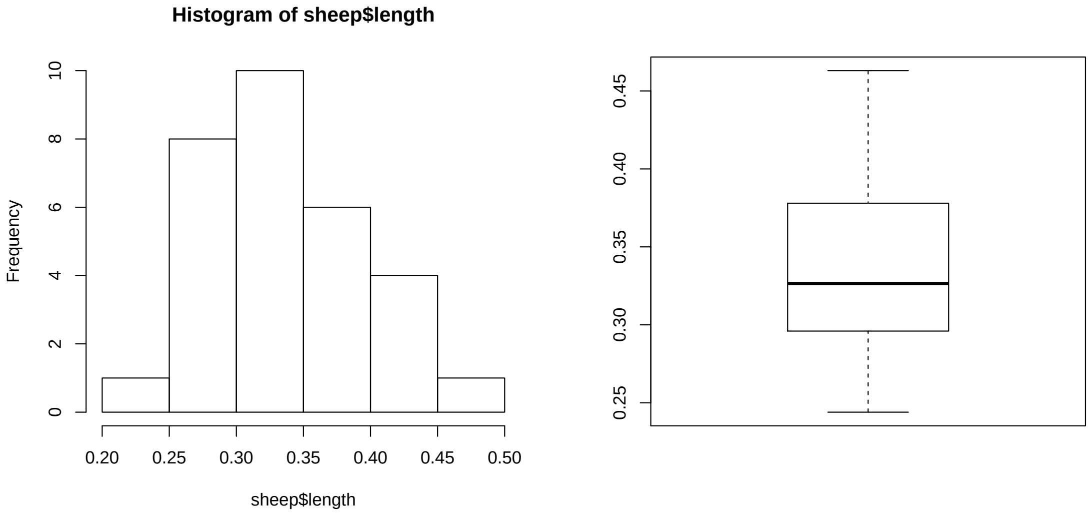
MAST90044 Thinking and Reasoning with Data
Semester 1 2019
Assignment 1 Solutions

Q.1.
```
> sheep <- read.csv("sheep.csv")
```

(a)
```
> summary(sheep$length)
```

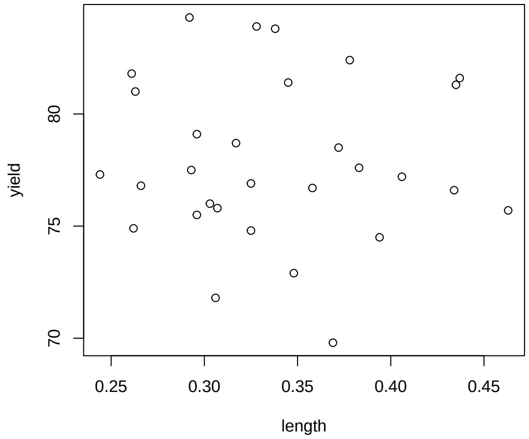
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2440  0.2960  0.3265  0.3381  0.3765  0.4630
```

```
> par(mfrow=c(1,2))
> hist(sheep$length)
> boxplot(sheep$length)
```



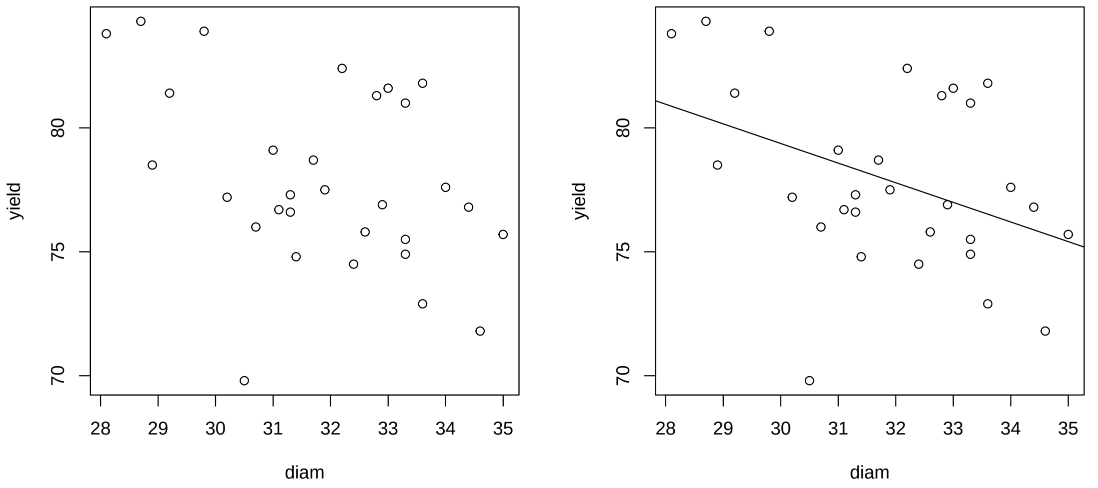
Distribution is consistent with normality – reasonably symmetrical, no strange observations.

(b)
```
> plot(yield˜length,data=sheep)
```



No apparent relationship.

(c)
```
> par(mfrow=c(1,2))
> plot(yield~diam,data=sheep)
> plot(yield~diam,data=sheep)
> abline(lm(yield~diam,data=sheep))
```



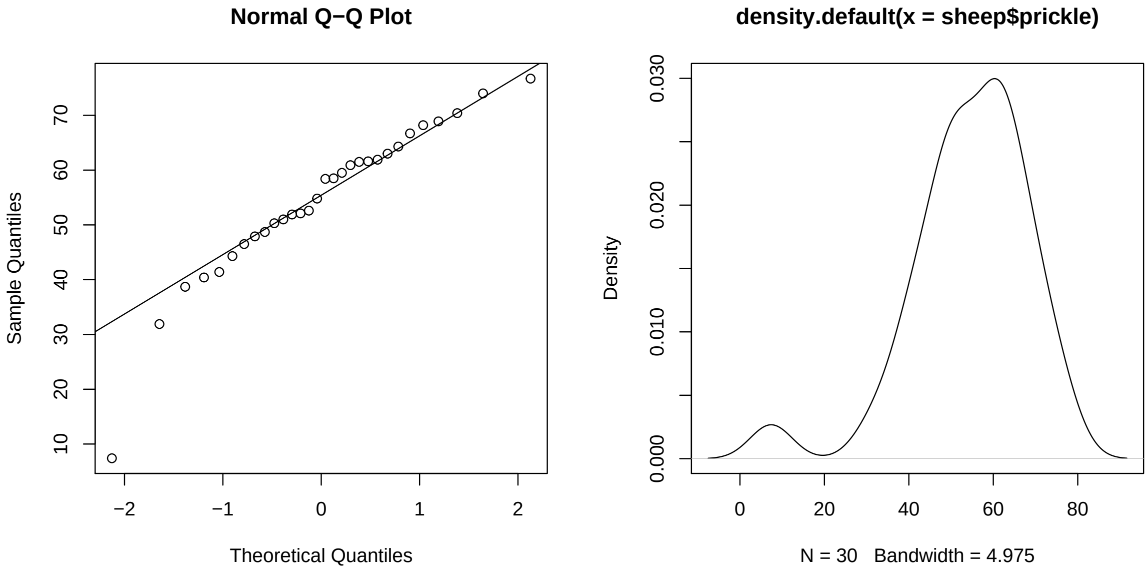
```
> cor(sheep$yield, sheep$diam)
[1] -0.4006844
```

$$
\begin{aligned}
\text{Model:} \qquad yield_i &= \alpha + \beta diam_i + e_i, \quad e_i \sim \mathrm{N}(0,\sigma) \\
y_i &= \alpha + \beta x_i + e_i, \quad e_i \sim \mathrm{N}(0,\sigma)
\end{aligned}
$$

$\hat{\alpha} \approx 100; \quad \hat{\beta} \approx -0.8.$

There appears to be a moderate negative relationship between yield and fibre diameter. There is an outlier with a very small yield.

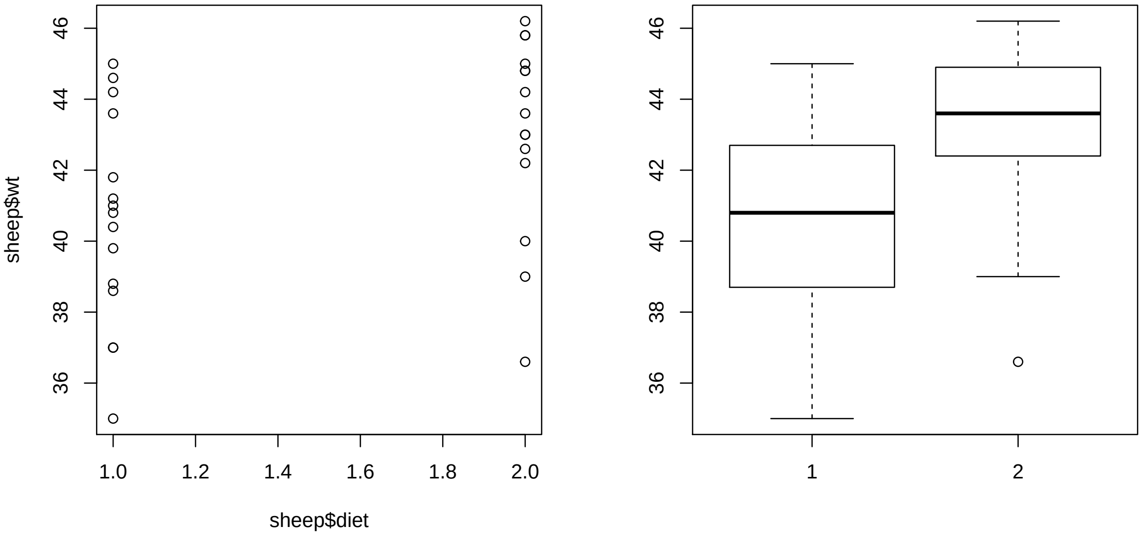
(d)
```
> par(mfrow=c(1,2))
> qqnorm(sheep$prickle)
> qqline(sheep$prickle)
> plot(density(sheep$prickle))
```



The distribution of prickle is reasonably consistent with a normal distribution.

(e)

```
> par(mfrow=c(1,2))
> plot(sheep$wt˜sheep$diet)
> boxplot(sheep$wt˜sheep$diet)
```



```
> dietlabel <- c(rep("no lupin", 15), rep("lupin", 15))
> tapply(sheep$wt, dietlabel, mean)
```

```
lupin no lupin
43.10667 40.58667
```

```
> tapply(sheep$wt, dietlabel, sd)
```

```
lupin no lupin
2.735864 2.982297
```

It appears that the lupin diet increases the average final ewe weight. It doesn't affect the spread much.

(f)

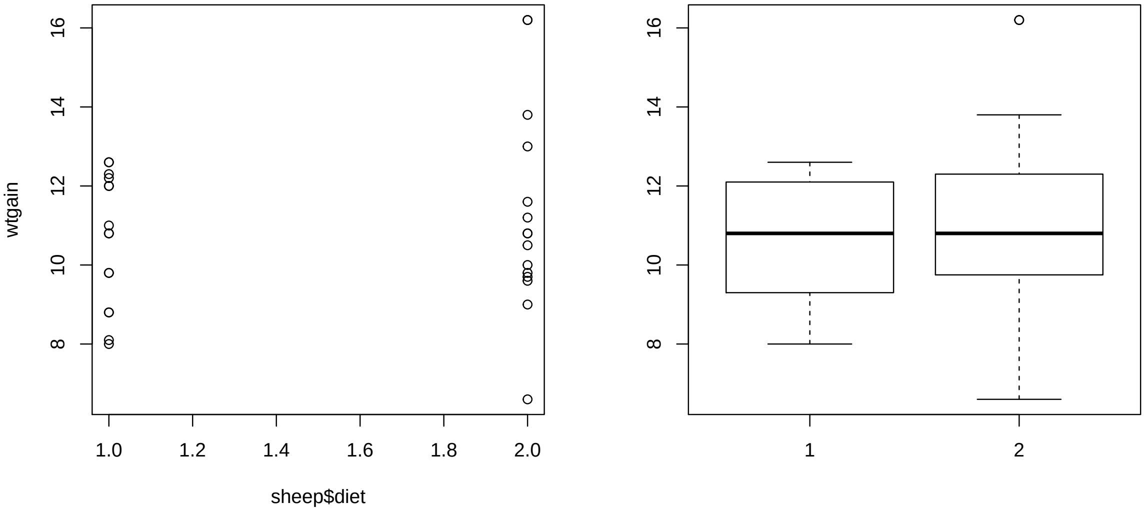
$$\text{Model:} \quad wt_{ij} = diet_i + e_{ij}$$
$$y_{ij} = \mu_i + e_{ij}$$

$i = 1, 2$ or (no lupin, lupin); $j = 1, \ldots, 15$.

$\hat{\mu}_1 = 43.1$; $\hat{\mu}_2 = 40.6$; $\hat{\sigma} \approx 2.8 - 2.9$.

Assumption: errors e_{ij} are separate draws from a normal distribution with mean 0 and standard deviation σ, i.e. $e_{ij} \sim \mathrm{N}(0, \sigma)$.

(g)
```
> wtgain <- sheep$wt - sheep$initwt

> par(mfrow=c(1,2))
> plot(wtgain~sheep$diet)
> boxplot(wtgain~sheep$diet)
```



```
> tapply(wtgain,dietlabel,mean)

lupin no lupin
11.25333 10.64000

> tapply(wtgain,dietlabel,sd)

lupin no lupin
2.606823 1.655208
```

It appears that the lupin diet does not have much effect on the average weight gain – the difference between diets is much less than it was for final ewe weight.

A report on the effect of these two diets on ewe weight would highlight the fact that although the lupin diet resulted in sheep with larger average weight, much of this effect was due to their initial weight being larger.

[4 + 2 + 9 + 3 + 7 + 6 + 7 + 7 = 45 marks]

Q.2. (a)

```
> n <- 30 + 224
> (p.hat <- 30/n)
[1] 0.1181102
```

1 mark for correct proportion.

Wald:

```
> p.hat + c(-1.96, 1.96) * sqrt(p.hat * (1 - p.hat)/n)
[1] 0.07841941 0.15780106
```

Wilson:

```
> prop.test(30,254)

1-sample proportions test with continuity correction

data:  30 out of 254, null probability 0.5
X-squared = 146.65, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.08233747 0.16581662
sample estimates:
p
0.1181102
```

Clopper-Pearson:

```
> binom.test(30,254)

Exact binomial test

data:  30 and 254
number of successes = 30, number of trials = 254, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.08112433 0.16430048
sample estimates:
probability of success
0.1181102
```

Agresti-Coull:

```
> n.tilde <- 30+224+4
> p.tilde <- (30 + 2)/n.tilde
> p.tilde + c(-1.96, 1.96) * sqrt(p.tilde * (1 - p.tilde)/n.tilde)
[1] 0.08380974 0.16425227
```

Jeffreys prior:

```
> qbeta(c(0.025, 0.975), 30 + 0.5, 254 - 30 + 0.5)
[1] 0.08278632 0.16207324
```

3 marks, 1 mart for each methods, for correct R code and output.

$n = 254$ is large and the estimated proportion is not close to 0 or 1, so the Wald method is acceptable here. Agresti-Coull is even better. It is easily-computed interval estimate that have substantially better coverage than the Wald intervals.

3 marks; 0.5 for mentioning large n, 0.5 mark for mentioning that the proportion far from 0/1, and 1 mark for mentioning that Wald test is ok. 1 mark for mentioning that Agresti-Coull has better coverage.

(b) Assumptions:

- n is large, so that is not a problem. [1 mark.]
- Observations are independent — this depends on the survey design, but should be OK. [1 mark.]
- The sample is representative of the population — this is unknown with the information given here. If the sample were randomly selected, it would not be a problem, but if, for example, the sample was taken from a sleep clinic, then it may not be representative. [1 mark.]

(c)

```
> n1 <- 21 + 30 + 192 + 224
> (p1.hat <- (21 + 30)/n1)
[1] 0.1092077
```

[1 mark.]

```
> n2 <- 24 + 35 + 1355 + 603
> (p2.hat <- (24 + 35)/n2)
[1] 0.02925136
```

[1 mark.]

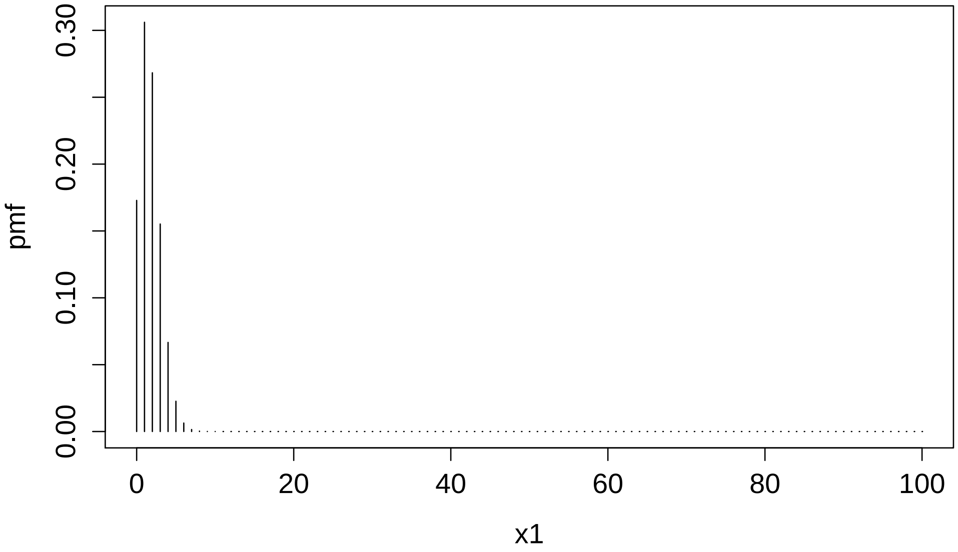
```
> p1.hat - p2.hat + c(-1.96, 1.96) * sqrt(p1.hat * (1 - p1.hat)/n1 +
+    p2.hat * (1 - p2.hat)/n2)

[1] 0.05072739 0.10918530
```

[1 mark.]

The claim is not supported by the 95% confidence interval – a zero difference is very implausible in light of the data. [2 marks.]

(d) The random variable has a binomial distribution with $n = 100$ and $p = 24/1379 = 0.0174$, i.e. $X \sim \text{Bi}(100, 0.0174)$.
$E(X) = np = 100 \times 0.0174 = 1.74$.
3 marks: 1 for recognizing binomial distribution and 1 for correct parameters. 1 Mark for expectation calculation.

```
> x1 <- 0:100
> pmf <- dbinom(x1, 100, 0.0174)
> plot(pmf ˜ x1, type = "h")
```



1 mark for correct R code and output.

(d)
```
> dbinom(2, 100, 0.0174)
[1] 0.2683034
```

[1 marks.]

```
> 1 - pbinom(1, 100, 0.0174)
[1] 0.5210578
```

[2 marks.]

[7 + 3 + 5 + 4 + 3 = 22 marks]

Q.3. (a) The mean of the difference between the proportions:

$$0.68 - 0.58 = 0.1$$

The standard deviation of the difference between the proportions:

$$s = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.68(1-0.68)}{100} + \frac{0.58(1-0.58)}{100}}$$
$$= \sqrt{0.0022 + 0.0024} = \sqrt{0.0046} = 0.0679$$

3 marks: 1 for calculation of the mean, 1 for calculation of the the sd; 1 for any reasonable comment.

(b) 95% confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$
$$0.1 \pm 1.96 \times 0.0679$$

The 95% confidence interval is $[-0.0331\,,\,0.2331]$.

Zero, i.e. no difference between Aki and Natasha, is inside of the confidence interval. Therefore we don't have enough evidence to reject the null hypothesis of equal abilities ($p_1 = p_2$).

5 marks: 3 marks for correct method and answer for the CI, 2 marks for any reasonable conclusion (note that in chapter 3 we had not yet covered using the CI for a hypothesis test).

(c) R code to calculate the 95% confidence interval:

```
> prop.test(x = c(68,58) , n = c(100,100))  # by default: correct = TRUE
2-sample test for equality of proportions with continuity correction
data:  c(68, 58) out of c(100, 100)
X-squared = 1.7375, df = 1, p-value = 0.1875
alternative hypothesis: two.sided
95 percent confidence interval:
-0.0431045  0.2431045
sample estimates:
prop 1 prop 2
0.68   0.58
```

Without the correction:

```
> prop.test(x = c(68,58) , n = c(100,100) , correct = FALSE)
2-sample test for equality of proportions without continuity correction
data:  c(68, 58) out of c(100, 100)
X-squared = 2.145, df = 1, p-value = 0.143
alternative hypothesis: two.sided
95 percent confidence interval:
-0.0331045  0.2331045
sample estimates:
prop 1 prop 2
0.68   0.58
```

2 marks for correct R code and output.

[3 + 5 + 2 = 10 marks]

Total marks = 77