Babak Shahbaba

Biostatistics with R

Solutions to Selected Exercises

Visit `http://www.ics.uci.edu/~babaks/BWR` for the most up-to-date version.

Springer

Chapter 1
Introduction

Question 2

Objective: To investigate associations of chocolate consumption (explanatory variable) with measured blood pressure (BP) and incidence of myocardial infarction and stroke (response variables) in middle-aged individuals.
Population: middle-aged German individuals; individuals with prevalent CVD and patients using antihypertensive medication were excluded from this study so the findings of this study cannot be extended to this part of the population.
Sample size: 19,357 participants after exclusion.
Type of study: prospective observational study. They collect cross-sectional and survival data.
Findings: Chocolate consumption is negatively associated with CVD risk; that is, higher amount of chocolate consumption tends to coincide with lower CVD risk. The inverse association seem to be stronger for stroke than for MI. In the abstract, the authors claims that "chocolate consumption appears to *lower* CVD risk". This may imply *causal relationship*, which is not possible to establish based on observational studies in general. The authors however use their language more carefully throughout the paper.

Question 3

The study by Taubert et al. is a randomized block (by gender) experiment that includes 44 Germans aged 56 through 73 years (24 women, 20 men). The target population is healthy middle-aged individuals.
Advantages: we could make inference about possible cause-effect relationships.
Disadvantages: in this particular study, the sample size is small. Findings: Inclusion of small amounts of polyphenol-rich dark chocolate as part of a usual diet efficiently

reduced BP and improved formation of vasodilative nitric oxide.
The study by Buijsse et al., on the other hand, is an observational study.
Advantages include large sample size which could produce better estimates. Many variables were included in the study to allow for more thorough analyses.
Disadvantages: we can only identify associations not causal relationships; no distinction between milk and dark chocolate as their cocoa content is different; this was done more properly through the randomized experiment by Taubert et al. Dietary intake, risk factors, and BP were assessed at baseline only; therefore, the authors assume that these variables remained stable over time; it would have been better to measure these variables over time and perform longitudinal analysis.

Question 4

The objective of this retrospective observational study is to examine the hypothesis that fatal medication errors spike in July, when thousands begin medical residencies. They looked at medication errors in the US from 1979 to 2006 ($n = 244,388$). They found a significant July spike in fatal medication errors inside medical institutions. Further, they concluded that the July mortality spike results at least partly from changes associated with the arrival of new medical residents. Although their findings seem to confirm the "New Resident Hypothesis," the authors provide alternative explanations by listing some possible confounding variables, for example, more alcohol consumption during summer, increase in injuries from accidents during summer, and increase in summer tourism. Although they state that the "New Resident Hypothesis" is still the best explanation for the July effect, this example nevertheless shows that why we should not make causal inference based on observational studies: because there might be some confounding variables influencing the observed relationships.

Based on their findings, the authors suggest several policy changes: 1) re-evaluating responsibilities assigned to new residents; 2) increasing supervision of new residents; 3) increasing education concerned with medication safety.

Question 5

Seeman discusses three studies examining the hypothesis that suggests low estrogen is associated with more sever symptoms in women with schizophrenia (target population). All three studies are prospective observational studies. The study by Hallonquist et al. (1993) includes a sample of 5 women, whose estrogen phase and symptoms are observed longitudinally. The second study also collected the data longitudinally from a sample of 32 women. The last study (Gattaz et al.) seems to be a case-control, cross-sectional study that includes 65 women aged 18 to 45

with schizophrenia (the case group) and 35 women with affective disorder (the control group). Seeman concludes that while the methodologies are different, the three studies suggest that there is a relationship between estrogen and symptoms of schizophrenia.

Question 8

The objective of this cross-sectional, observational study is to examine the relationship between BMI (response variabel) and neck circumference (predictor), and to find the best NC cutoff to identify children with him BMI. They specify their target population as children who were aged 6 to 18 years. However, they take their sample of 1102 children from those who had elective, noncardiac surgical procedures; this may affect how they can generalize their finding to the whole population of children aged 6 to 18. They found NC and BMI are strongly associated. Therefore, they suggest that NC could be an inexpensive and useful screening instrument for identifying overweight or obese children. Table 4 shows their recommended NC cutoff for different age groups.

Question 10

The objective of this study is to show that habituation to a food item can occur after imagining its consumption. To this end, the authors conduct 5 randomize experiments.

- In experiment 1, participants (N = 51 participants) randomly assigned to 3 groups:
  - Imagined inserting 33 quarters into a laundry machine (control group)
  - Imagined inserting 30 quarters into a laundry machine and then imagined eating 3 M&M's (3-repetition condition)
  - Imagined inserting 3 quarters into a laundry machine and then imagined eating 30 M&M's (30-repetition condition)
- In experiment 2, participants (N = 51 participants) randomly assigned to 4 groups:
  - Imagined eating 3 M&M's
  - Imagined eating 30 M&M's
  - Imagined inserting 3 quarters
  - Imagined inserting 30 quarters
- In experiment 3, participants (N = 68 participants) imagined:
  - Eating 3 M&M's

  - Eating 30 M&M's
  - Placing 3 M&M's into a bowl
  - Placing 30 M&M's into a bowl

- In experiment 4, participants (N = 80) imagined (the following) before consuming cheddar cheese:

  - Eating 3 M&M's
  - Eating 30 M&M's
  - Eating 3 cheddar cheese cubes into a bowl
  - Eating 30 cheddar cheese cubes into a bowl

- In experiment 5, (N = 81 participants)

  1. All participants rate how much they like cheddar cheese on 7-point scale: dislike extremely (1) and like extremely (7)
  2. Participants are divided into 2 groups:
     - Imagined performing 30 repetitions of the control task (as in experiment 1) and then imagined eating 3 cheddar cheese cubes and participants
     - Imagined performing three repetitions of the control task and then imagined eating 30 cheddar cheese cubes.
  3. All participants played the reinforcement game. At the end of the game, participants re-rated how much they liked cheddar cheese on a scale identical to the scale used in the beginning of the experiment.

Conclusions: Five experiments showed that people (target population) who repeatedly imagined eating a food (such as cheese) many times subsequently consumed less of the imagined food than did people who repeatedly imagined eating that food fewer times, imagined eating a different food (such as candy), or did not imagine eating a food.

Chapter 2
Data Exploration

Question 1

After you download "Calcium.txt", click `Data` → `Import data` → `from text file, clip board, or URL...` to upload the data into R-Commander. To create histograms, click `Graphs` → `Histogram` and select a numerical variable. Your plots should be similar to those in Figure 2.1.

The histogram of blood pressure at the beginning (before treatment) is unimodal and slightly skewed to the right. The frequency of observed values is high in the neighborhood of 115. The histogram of blood pressure at the end of the experiment is bimodal. Having multiple modes in a histogram usually indicates that our sample is not homogeneous and includes subgroups. In this case, it is trivial to identify the two subgroups: they are the calcium and placebo groups; this is of course how the experiment is designed (i.e., dividing the subjects into two treatments). While the experiment started with a homogeneous (in terms of BP) group of subjects, by the end of the experiment, those assigned to the calcium group had lower BP on average so they their distribution became different from that of the placebo group, hence the existence of the two modes.

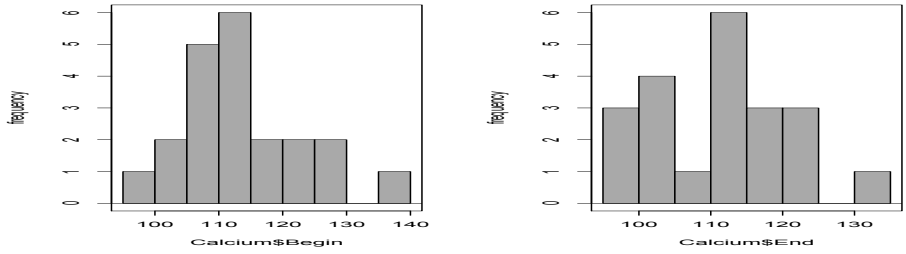

Fig. 2.1 Chapter 2 - Question 1

Question 2

After downloading “Survival.txt", click `Data → Import data → from text file, clip board, or URL...` to upload the data into R-Commander. To create a boxplot for Age, click `Graphs → Boxplot`, select “Age", then click “OK". To create bargragh for Status, click `Graphs → Bar graph`, select “Status", then click “OK". To create histogram for “Nodes”, click `Graphs → Histogram`, select “Nodes", then click “OK".

To plot $\sqrt{Nodes}$, you first need to create it as a new variable. Click `Data → Manage variables in active data set → Compute new variable`, then select “Nodes"; under “Variable name". You can name this new variable anything you want; to make it clear, however, let’s type in “SqrtNodes"; and under “Expression to compute", type in “sqrt(Nodes)". To create histogram for the new variable, SqrtNodes, click `Graphs → Histogram`, select “SqrtNodes", then click “OK". You graphs should be similar to those in Figure 2.2. Histogram for SqrtNodes is less skewed than that of Nodes.

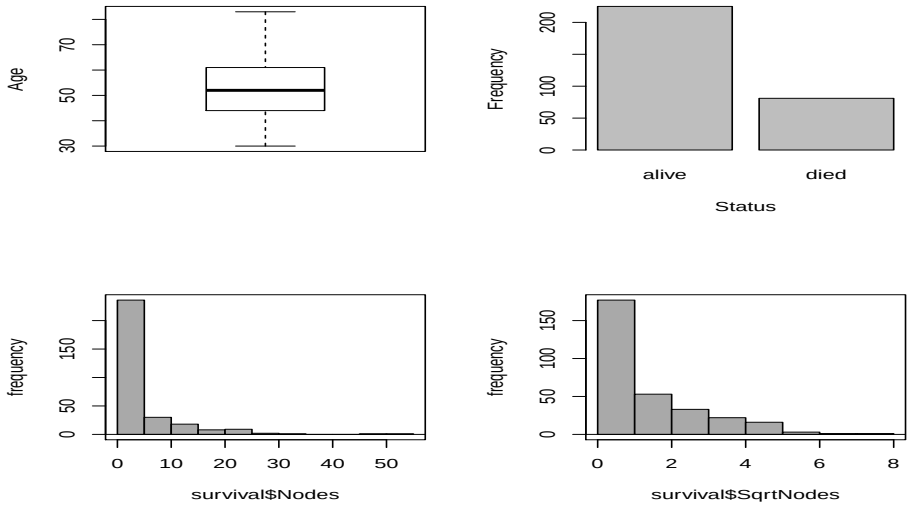

Fig. 2.2 Chapter 2 - Question 2

Question 4

Mean of Height = $\frac{\sum x_i}{n} = \frac{18+21+17+16+19}{5} = 18.2$
Mean of Weigth = 8.06
Standard deviation of Height = $\frac{\sum (x_i-\bar{x})^2}{n-1} = \frac{(18-18.2)^2+...+(19-18.2)^2}{5-1} = 1.92$
Standard deviation of Weight = 1.06

Question 5

Five number summary = (-10, -6, -4, -2, 2)
Range = 2 - (-10) = 12
IQR = -2 - (-6) = 4

Question 6

First, download "BodyTemperature.txt" and import it into R-Commander. To find five number summaries for numerical variables, go to `Statistics` → `Summaries` → `Numerical Summaries`, select all numerical variables, "Age", "HeartRate", and "Temperature", then click "OK". Here are the results:

```
               0%   25%  50%  75%  100%   n
Age          21.0 33.75 37.0 42.0  50.0 100
HeartRate    61.0 69.00 73.0 78.0  87.0 100
Temperature 96.2 97.70 98.3 98.9 101.3 100
```

Figure 2.3 shows histograms and boxplots for "Age", "HeartRate", and "Temperature". For "Age", the histogram is slightly skewed to the left; there is no outlier; the central tendency is around 35-40; we could use the sample mean (37.62) or the sample median (37.00) as a measure of central tendency. For "HeartRate", the histogram is almost symmetric; again, there is no outlier; the central tendency is around 70-75; as before we can use the sample mean (73.66) or sample median (73) as a measure of central tendency. For "Temperature", there seems to be bimodal: there is one mode around 98.5 and another one after 100. The sample might have included a group of individuals who had mild fever even though the target population was healthy individuals. On the other hand, because there are only few (4) individuals with body temperature above 100, they might be simply outliers. The boxplot shows that two of them can be in fact considered as outlier (denoted with dots). The central tendency is around 98-99. (You can use the sample mean and median to provide a more precise values for the central tendency.)

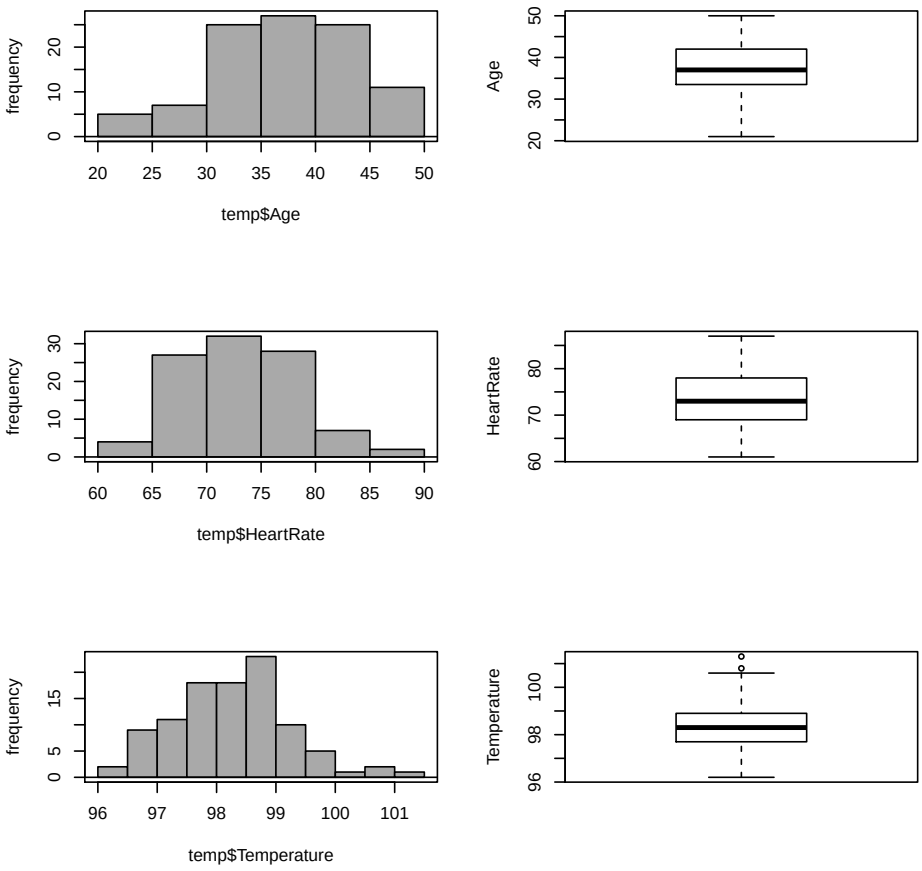

Fig. 2.3 Chapter 2 - Question 6

Question 7

CV for Age = $6.43/37.62 = 0.17$
CV for Temperature = $0.95/98.33 = 0.01$
If we multiple age by 12 to change its unit to month, its standard deviation and mean become $6.42 \times 12 = 77.16$ and $37.62 \times 12 = 451.44$ respectively. The resulting CV is $77.16/451.44 = 0.17$, which is the same as before.
To change the temperature unit to Celsius, we need to subtract 32 from the temperature values and multiply the result by 0.556. By doing so, the sample standard deviation and sample mean change to $0.556 \times 0.95 = 0.53$ and $0.556 \times (98.33 - 32) =$

0.36.88. The resulting CV is $0.53/36.88 = 0.014$, which is not the same as CV in terms of Fahrenheit.

Question 9

Download “AsthmaLOS.txt" and import it into R-Commander. For variables “race" and “owner.type", we need to identify and remove (assuming we cannot correct them) data entry errors. After reading description of variables provided in Section 2.5, we know that the range for race is integers from 1 to 5. If we find any values in “race" that are outside of this range, they are considered data entry errors. Similarly, only 1 and 2 are possible values for “owner.type". To check for data entry errors, let’s look at the frequencies for different values of “race" and “owner.type". Because these variables have numerical coding, we first need to convert them to factors first. To do this, click `Data` → `manage variables in active data set` → `Convert numeric variable to factors`, choose `owner.type` and `race`, and select the option `Use numbers`. Then, you can obtain the frequency table for the two variables. You will find that there are two cases with `owner.type=0` and race bigger than 5. You can simply identify these cases and remove them from the dataset by clicking `Data` → `Active data set` → `Remove row(s) from active data set` and providing their row number. This would be of course difficult for larger datasets. For large dataset, we could use other method such as taking a subset of data with the acceptable values.
Figure 2.4 shows histogram of “Age". The shape of this variable is skewed to the right. The sample mean and median are 6.7 and 5 respectively; variance is $4.34^2 = 18.85$; range=max-min=17-2=15 and IQR=Q3-Q1=10-3 = 7. Note that if you do not remove the outliers, minimum value of age will be 0.

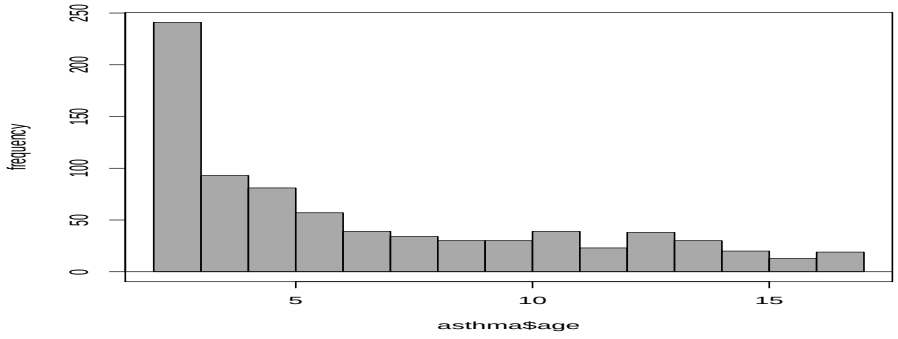

Fig. 2.4 Chapter 2 - Question 9

Question 10

Make sure `Animals` from the `MASS` package is the active dataset. Follow the steps discussed in Section 2.5.3 to create two new variables by log-transforming `body` and `brain`. Figure 2.5 shows the histogram of textttbody and `brain` before and after log-transformation. While the original variable are highly skewed (to the extent that using histograms to visualize the data becomes problematic), they become more symmetric after log-transformation.

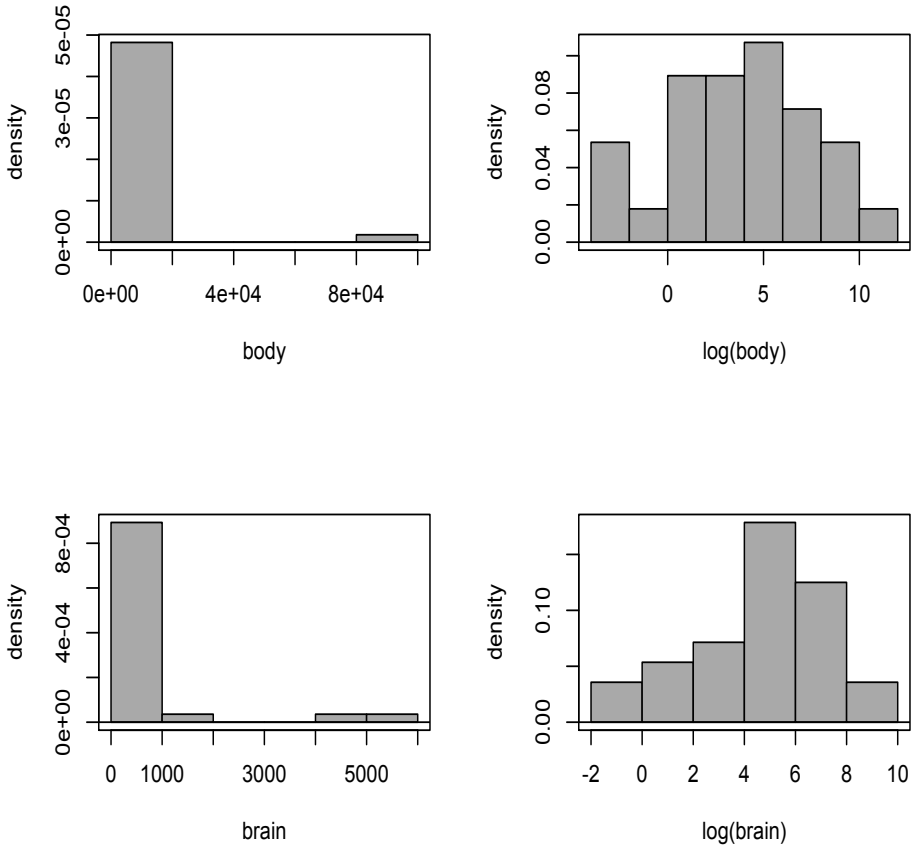

Fig. 2.5 Chapter 2 - Question 10

Chapter 3
Exploring Relationships

Question 1

We need to create a table similar to Table 3.2, and follow the steps discussed in pages 66 and 67. The sample covariance and correlation coefficient are 1.76 and 0.86 respectively.

Question 2

After you upload BodyTemperature into R-Commander, to create the scatterplot, click `Graphs` → `scatterplot`, select "HeartRate" for x-variable and "Temperature" for y-variable. To make a scatterplot with just the least-squares line (i.e., trend line), you should unmark other options, such as "Smooth line", "Show spread", and "Marginal boxplots", then click OK. The scatterplot between body temperature and heart rate is shown in the left panel of Figure 3.1. The plot suggests that the increase in heart rate tends to coincide with the increase in body temperature. The two variables seem to have a positive linear relationship. To find correlation coefficient between body temperature and heart rate, go to `Statistics` → `Summaries` → `Correlation matrix...`, select "Temperature" and "HeartRate", then click OK. You should get $correlation = 0.448$. This correlation coefficient is in accordance to what we found from examining the scatterplot. Again, to create boxplots, point to `Graphs` → `boxplot`, highlight "Temperature", click on "Plot by groups" to select Gender, then click OK. This will create boxplots of temperature separately for men and women. Boxplots for temperature by gender is shown in the right panel of Figure 3.1. Men's body temperature tends to be slightly lower. Further, body temperature for men seems to be more dispersed compared to that of women.

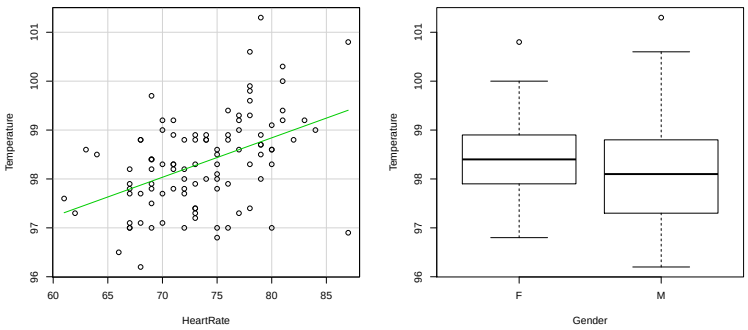

Fig. 3.1 Chapter 3 - Question 2

Question 3

As shown in page 62, you can upload into R-Commander by first entering these two commands in R Console:

```
install.packages("mfp", dependencies=TRUE)
library(map)
```

Then, you can access `body fat` by following the steps in page 62. The steps to create scatterplots are also presented in page 62. Before creating scatterplots, however, you need to create a new variable for BMI by following the steps in page 44. Your plots should look similar to Figure 3.2 before and after removing the outlier (see page 65). The correlation coefficient between `siri` and `neck` is 0.49. The

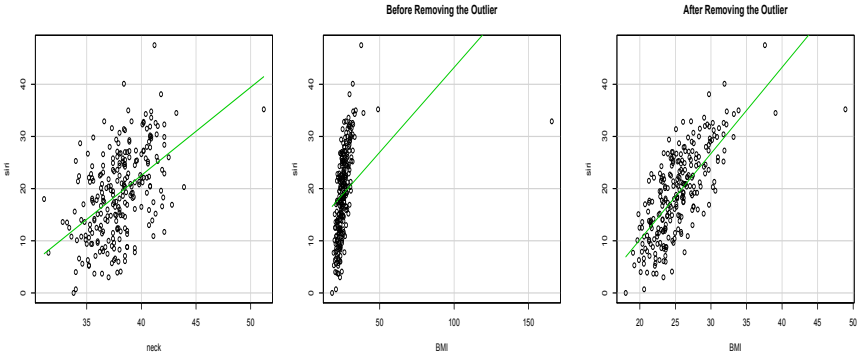

Fig. 3.2 Chapter 3 - Question 3

correlation coefficients between `siri` and `BMI` are 0.37 and 0.72 before and after removing the outlier respectively. Assuming the outlier was in fact data entry error, percent body fat has a stronger positive linear relationship with BMI than neck circumference. (Additionally, we can investigate the relationship between `neck` and `BMI`.)

Question 4

We start by creating a contingency table as follows:

	Heart Disease	No Heart Disease	Total
Never snore	24	1355	1379
Snore	86	1019	1105
Total	110	2374	2484

Table 3.1 Chapter 3- Question 4

Let p_1 be the proportion of people with heart disease for the “Never Snore" group and p_2 be the proportion of people with heart disease for the “Snore" group.

$$p_1 = \frac{24}{1379} = 0.017$$

$$p_2 = \frac{86}{1105} = 0.078$$

Difference of proportions, $p_2 - p_1 = 0.078 - 0.017 = 0.061$. The proportion of people suffering from heart disease increases by 0.061 in the snoring group compared to the non-snoring group.
Relative risk of heart disease is $p_2/p_1 = 0.078/0.017 = 4.59$. This means that the risk of a heart disease in the snoring group is 4.59 times of the risk in the non-snoring group.
Odds ratio is $OR = \frac{p_2/(1-p_2)}{p_1/(1-p_1)} = \frac{0.078/(1-0.078)}{0.017/(1-0.017)} = 4.89$. This means that the odds of a heart disease in the snoring group is 4.89 times higher than that of the non-snoring group.

Question 6

After importing “birthwt" data to Rcmdr, notice that variables “ht" and “low" are categorical variables with 0 denoting no history of hypertension and not having low birthweight babies respectively. Since Rcmdr wouldn’t know these variables are

categorical, we need to convert them to categorical by using `Data` → `Manage variables in active data set` → `Convert numeric variable to factors`, then select `ht` and `low`, choose `Use numbers`, and click "OK".

To create a 2-way contingency table, go to `Statistics` → `Contingency tables` → `Two-way table`, highlight `ht` for `Row variable` and `low` for `Column variable`; unselect `Chi-squares test of independence`, then click "OK". You should obtain the following table:

```
   low
ht    0   1
  0 125  52
  1   5   7
```

Now let p_1 be the proportion of low birthweight babies of mothers with history of hypertension and p_0 be the proportion of low birthweight babies of mothers with no history of hypertension,

$$p_0 = 52/177 = 0.294$$

$$p_1 = 7/12 = 0.583$$

Relative risk of having low birhtweight babies is $p_1/p_0 = 0.583/0.294 = 1.98$. This means that the risk of having low birthweigth babies among mothers with history of hypertension is almost double the risk among mothers with no history of hypertension.
Odds Ratio is $OR = \frac{0.583/(1-0.583)}{0.294/(1-0.294)} = 3.36$. This means that the odds of having low birthweigth babies among mothers with history of hypertension is 3.36 times higher than that of mothers with no history of hypertension.

Question 8

Figure 3.3 shows the three boxplots. All three variables tend to be higher in the diabetic group compared to the non-diabetic group. There is especially a substantial difference between the two groups in terms of the distribution of `glu`. Therefore, `glu` (and the other two variables to some extent) seems to be associated with diabetes.

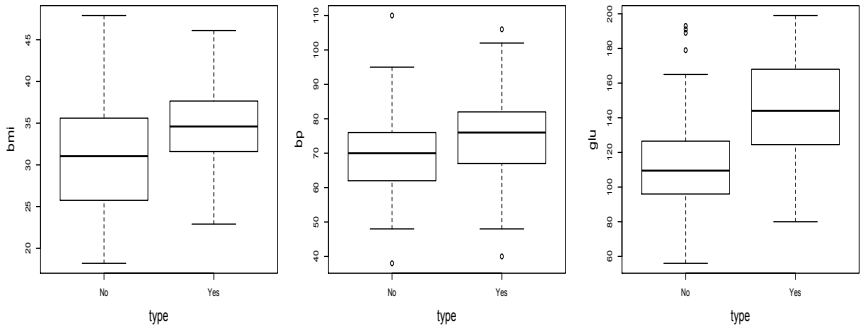

Fig. 3.3 Chapter 3 - Question 8

Chapter 4
Probability

Question 1

Given $P(E_1) = 0.3$ and $P(E_2) = 0.5$,

1. If E_1 and E_2 are disjoint, $P(E_1 \cup E_2) = P(E_1) + P(E_2) = 0.3 + 0.5 = 0.8$. In this case, these events are not partitioning the sample space; If they did, their union would have been equal to the sample space, whose probability is 1.
2. $P(E_3) = P((E_1 \cup E_2)^C) = 1 - P(E_1 \cup E_2) = 1 - 0.8 = 0.2$
3. If E_1 and E_2 are independent, $P(E_1 \cap E_2) = P(E_1)P(E_2) = (0.3)(0.5) = 0.15$
4. If E_1 and E_2 are independent, $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1)P(E_2) = (0.3) + (0.5) - 0.15 = 0.65$
5. $P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)} = \frac{(0.35)(0.5)}{0.3} = 0.58$. In this case, these two events are not independent since $P(E_1|E_2) \neq P(E_1)$

Question 2

$$
\begin{aligned}
P(aa) &= 0.1^2 = 0.01 \\
P(Aa) &= 2 \times 0.1 \times 0.9 = 0.18 \\
P(AA) &= 0.9^2 = 0.81
\end{aligned}
$$

Question 3

Figure 4.1 shows the sample space and probabilities of each possible outcome.

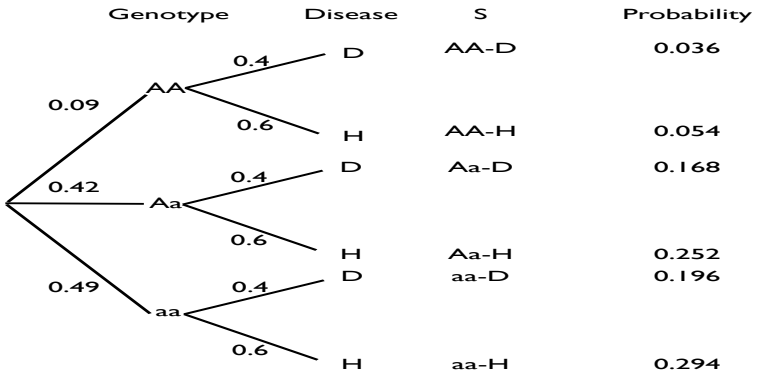

Fig. 4.1 Chapter 4 - Question 3

Question 4

Figure 4.2 shows the sample space and probabilities of each possible outcome.

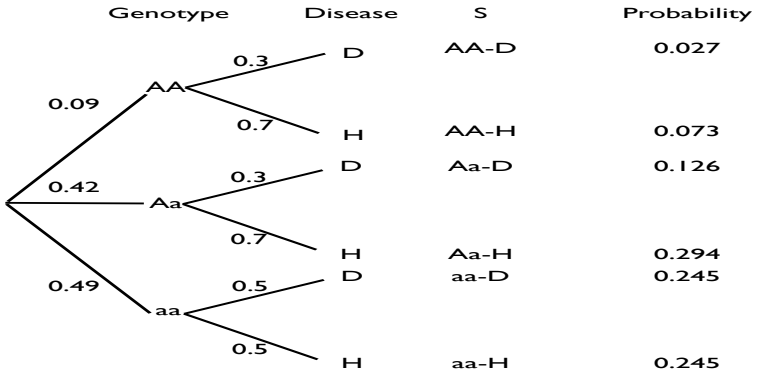

Fig. 4.2 Chapter 4 - Question 4

Question 5

Figure 4.3 shows the sample space and probabilities of each possible outcome.

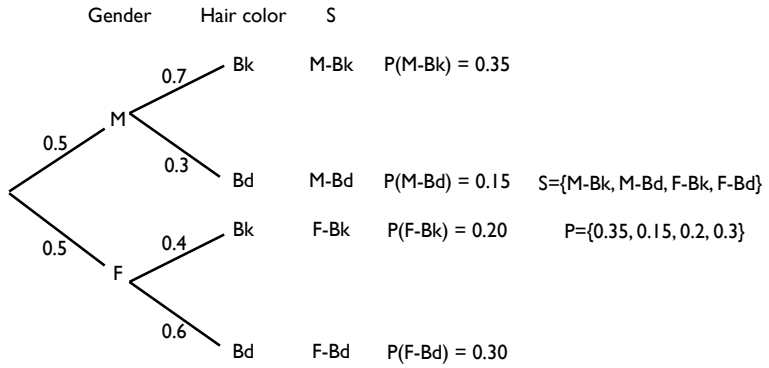

Fig. 4.3 Chapter 4 - Question 5

Question 6

Let F be the event that someone is affected with H1N1 flu and W be the event that someone washes her hands regularly. Given $P(F) = 0.02$, $P(W) = 0.6$, and $P(W|F) = 0.3$, then

$$P(F|W) = \frac{P(W|F)P(F)}{P(W)} = \frac{(0.3)(0.02)}{0.6} = 0.01.$$

Thus, the probability of getting H1N1 flu if a person washes her hands regularly is 0.01.

Chapter 5
Random Variables and Probability Distributions

Question 2

For this question, the correct plots are shown in Figure 5.1. (*In the book, the left plot is wrong.*)

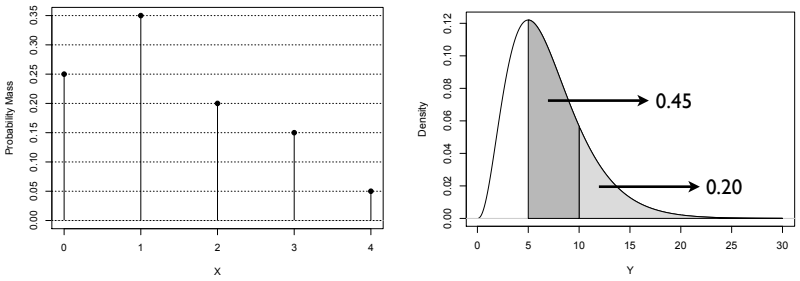

Fig. 5.1 Chapter 5 - Question 2

a) $P(X < 3) = 0.8$
b) $P(1 < X \leq 4) = 0.4$
c) $P(Y > 5) = 0.65$

Question 3

a) Figure 5.2 shows the probability mass function and cumulative distribution function.

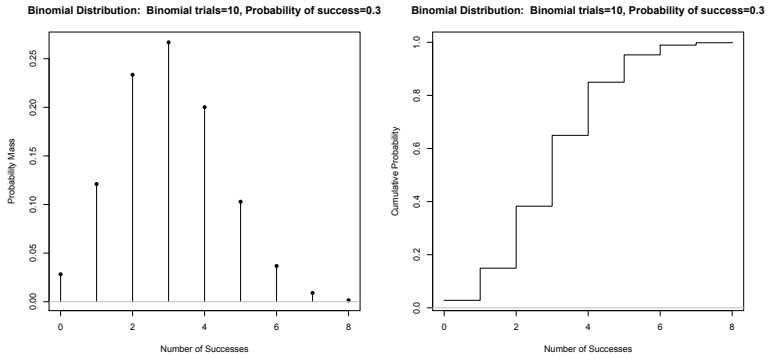

Fig. 5.2 Chapter 5 - Question 3-a

b) $\mu = 10 \times 0.3 = 3$ and $\sigma = \sqrt{10 \times 0.3 \times (1-0.3)} = 1.45$
c) $P(X \leq 4) = 0.85$
d) $P(X = 2) = 0.23$
e) $P(2 < X \leq 4) = 0.85 - 0.38 = 0.47$

Question 4

a) Figure 5.3 shows the probability density function and cumulative distribution function.

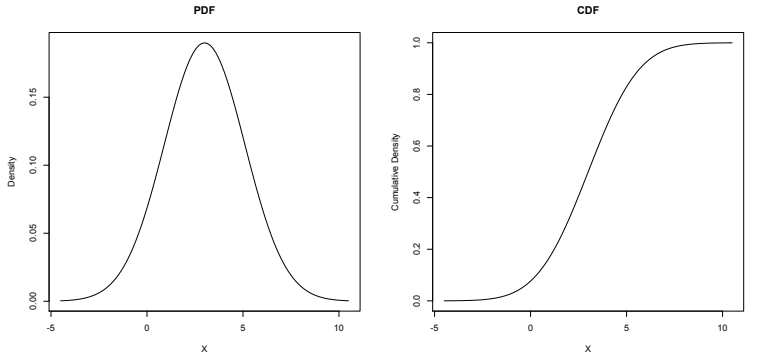

Fig. 5.3 Chapter 5 - Question 4-a

b) $\mu = 3, \sigma = \sqrt{2.1} = 1.45$

c) $P(X \leq 4) = 0.75$
d) $P(X = 2) = 0$
e) $P(2 < X \leq 4) = 0.75 - 0.24 = 0.51$

Question 7

a) Here are the probabilities of different categories:

$P(\text{Normal}) = 0.37$

$P(\text{Prehypertension}) = 0.84 - 0.37 = 0.47$

$P(\text{High blood pressure}) = 0.16$

b) Here are the three intervals:

68%: (125-15, 125+15] = (110, 140]
95%: (125 - 2×15, 125 + 2×15] = (95, 155]
99.7%: (125 - 3×15, 125 + 3×15] = (80, 170]

c) $P(SBP \leq 115) = 0.25$ and $P(SBP > 115) = 1 - 0.25 = 0.75$.

Question 8

a) Here are the probabilities of different categories:

$P(\text{Underweight}) = 0.078$
$P(\text{Normal Weight}) = 0.291$
$P(\text{Overweight}) = 0.322$
$P(\text{Obesity}) = 0.309$

b) 68%: (21, 33], 95%: (15, 39], 99.7%: (9, 45]
c) $P(\text{Obese or Underweight}) = 0.078 + 0.309 = 0.387$; the underlying events are mutually exclusive.
d) $P(BMI \leq 29.2) = 0.643$, $P(BMI > 29.2) = 0.357$

Question 9

To find x such that $P(X \leq x)$= 0.2, we follow the steps discussed in this chapter and find 0.2-quantile of BMI, which is 21.95. To find x such that $P(X > x)$= 0.2, we can follow similar steps but instead of lower tail, we select upper tail. In this case, x = 32.05.

Question 11

We denote the value of SBP after everyone takes the drug as $W = X - Y$. The mean and variance of W are 153 - 4 = 149 and 16 + 1 = 17 respectively. Note that we are adding the variances even though we are subtracting the original random variables. If $Y \sim N(4,1)$, then $W \sim N(149,17)$.

Chapter 6
Estimation

Question 1

a) $\bar{X} \sim N(\mu, \sigma^2/n)$; The standard deviation of is $\sigma/\sqrt{n} = 6/\sqrt{9} = 2$.
b) Using the standard normal distribution, z_{crit} for 0.8 confidence level is 1.28. Therefore, the 80% confidence interval for μ is

$$\bar{x} \pm z_{crit} \times \sigma/\sqrt{n} = 110 \pm 1.28 \times 2 = [107.44, 112.56]$$

Question 2

In this case, SE = $6/\sqrt{9} = 2$. Because σ is unknown, we use t-distributions to find confidence intervals. In this case, we use the t-distribution with $n-1 = 8$ degrees of freedom and fine $t_{crit} = 1.40$. Therefore, the 80% confidence interval is:

$$\bar{x} \pm t_{crit} \times s/\sqrt{n} = 110 \pm 1.40 \times 2 = [107.20, 112.80]$$

Note that this interval is slightly wider (reflecting a higher level of uncertainty) compared to what we found in the previous question, even though they both have the same sample size and variance. This is because one of them uses the exact value of σ, the other one uses its estimate, which has its own uncertainty.

Question 3

The sample size is $n = 189$, and the sample proportions are $p_{low} = 0.31$ and $p_{ht} = 0.064$. Using the standard normal distribution, $z_{crit} = 1.44$ for the 0.85-confidence level. Therefore, the 85% confidence intervals for the population proportions of `low` and `ht` are as follows:

$$\text{low: } 0.312 \pm 1.44 \times \sqrt{0.312 \times (1-0.312)/189} = [0.263, 0.361]$$
$$\text{ht: } 0.064 \pm 1.44 \times \sqrt{0.064 \times (1-0.064)/189} = [0.038, 0.089]$$

Question 5

The sample size is $n = 100$. Using the t-distribution with $n-1 = 99$ degrees of freedom, we find $t_{crit} = 1.29$. For heart rate, the sample mean and standard deviation are 73.66 and 5.31 respectively. Therefore, the 80% confidence interval for the population mean of heart rate is

$$\bar{x} \pm t_{crit} \times s/\sqrt{n} = 73.66 \pm 1.29 \times 5.31/10 = [72.97, 74.34].$$

For normal body temperature, the sample mean and standard deviation are 98.33 and 0.96. Therefore, the 80% confidence interval for the population mean of normal body temperature is

$$\bar{x} \pm t_{crit} \times s/\sqrt{n} = 98.33 \pm 1.29 \times 0.96/10 = [98.21, 98.45].$$

Question 6

Using the standard normal distribution, $z_{crit} = 1.64$ for the 0.9 confidence level. Our rough estimate of σ is $(11-2)/4 = 2.25$. Therefore, the required sample size is:

$$n = \left(\frac{1.64 \times 2.25}{0.5}\right)^2 \approx 55.$$

Question 7

Using the standard normal distribution, $z_{crit} = 1.96 \approx 2$ for 0.95 confidence level. Therefore, the required sample size is:

$$n = \left(\frac{2 \times 0.5}{0.02}\right)^2 = 2500.$$

Chapter 7
Hypothesis Testing

Question 1

Here, $\mu_0 = 115$, $H_A : \mu < 115$, $H_0 : \mu = 115$, $n = 100$, $\bar{x} = 111$, and $s = 32$. We calculate the t-score as follows:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{111 - 115}{32/\sqrt{100}} = -1.25$$

Because this is a one-sided test of the form $H_A : \mu < \mu_0$, $p_{obs} = P(T \leq t)$, where the distribution of T is $t(99)$. Using R-Commander, $p_{obs} = 0.11$. Because p-value is not less than the pre-specified cutoff, 0.1, we can not reject the null hypothesis at 0.1 level.

Question 2

Here, $H_A : \mu \neq 70$ vs. $H_0 : \mu = 70$. Based on the `Pima.tr` dataset, $\bar{x} = 71.26$ and $s = 11.48$. We calculate t-score = 1.55. Using the t-distribution with $df = 200 - 1 = 199$, p-value is 0.12. We fail to reject H_0 at commonly used significance levels (0.01, 0.05, and 0.1). In other words, there is not enough evidence to conclude that the mean blood pressure for Pima Indian women is different from 70. We can of course use R-Commander directly to test our hypothesis: `Statistics` $\rightarrow$ `Means` $\rightarrow$ `Single-sample t-test`.

Question 3

$$H_0 : \mu_0 = 0.2 \text{ vs. } H_A : \mu_0 < 0.2$$
$$p = 27/150 = 0.18$$
$$z = \frac{0.18 - 0.2}{\sqrt{0.2 \times 0.8/150}} = -0.612$$

Using the standard normal distribution, $p_{obs} = P(Z \leq -0.612|H_0) = 0.27$. Therefore, we fail to reject H_0 at commonly used significance levels (0.01, 0.05, 0.1).

Question 5

We need to upload the data into R-Commander, click `Statistics` → `Means` → `Single-sample t-test` and set $\mu_0 = 75$ (i.e., `null hypothesis: mu`). For the first part, where $H_A : \mu < 75$ vs. $H_0 : \mu = 75$, we use a one-sided t-test by selecting `Population mean < m0`. In this case, p-value = 0.007, which is less than the cutoff 0.01 so we can reject the null hypothesis at 0.01 significance level. Therefore, the observed difference from 75 is statistically significant at this level.

```
        One Sample t-test

data:  NormTemp$HeartRate
t = -2.5222, df = 99, p-value = 0.006629
alternative hypothesis: true mean is less than 75
95 percent confidence interval:
     -Inf 74.54215
sample estimates:
mean of x
    73.66
```

For the second part, where $H_A : \mu \neq 75$ vs. $H_0 : \mu = 75$, we use a one-sided t-test by selecting `Population mean != m0`. In this case, p-value = 0.013, which is bigger than the cutoff 0.01 so we cannot reject the null hypothesis at 0.01 significance level. Therefore, the observed difference from 75 is not statistically significant at this level.

```
        One Sample t-test

data:  NormTemp$HeartRate
t = -2.5222, df = 99, p-value = 0.01326
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 72.60581 74.71419
sample estimates:
```

```
mean of x
    73.66
```

Chapter 8
Statistical Inference for the Relationship Between Two Variables

Question 1

Using R-Commander we find $\bar{x}_{diabetic} = 74.59, \bar{x}_{nodiabetc} = 69.55$. Therefore, the observed difference between the sample means is $69.55 - 74.59 = -5.04$. We want to examine $H_A : \mu_{diabetic} \neq \mu_{nondiabetic}$ vs. $H_0 : \mu_{diabetic} = \mu_{nondiabetic}$. For this, we use an independent-samples t-test.

```
        Welch Two Sample t-test

data:  bp by type
t = -2.9592, df = 130.278, p-value = 0.003665
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.414080 -1.671482
sample estimates:
 mean in group No mean in group Yes
         69.54545          74.58824
```

Because p-value is 0.0037, we can reject the null hypothesis at 0.01 level and conclude that the observed difference between the sample means of the two groups is statistically significant. This indicates a statistically significant relationship between `bp` and `type`.

Question 3

We examine $H_A : \mu_{c39} \neq \mu_{c52}$ vs. $H_0 : \mu_{c39} = \mu_{c52}$,

```
        Welch Two Sample t-test

data:  VitC by Cult
```

```
t = -6.3909, df = 56.376, p-value = 3.405e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.94296  -8.85704
sample estimates:
mean in group c39 mean in group c52
             51.5              64.4
```

As we can see, p-value is quite small so we can reject H_0 and conclude that the mean vitamin C for cultivar c52 is greater than the mean vitamin C for cultivar c39. Therefore, the relationship between `VitC` and `Cult` is statistically significant.

Question 4

We use a paired t-test for this problem. The average of pairwise differences is 20.93, and the standard deviation 37.74. The t-score is $20.93/(37.74/\sqrt{15}) = 2.15$. Using the $t(14)$ distribution, p-value is 0.0497, which is less than 0.05. Therefore, we can reject the null hypothesis at 0.05 level and conclude that the difference between average heights for the two groups is statistically significant.

Question 5

Sample proportions of heart attack in the placebo and aspirin groups are $p_1 = 189/(189+10845) = 0.017$ and $p_2 = 104/(104+10933) = 0.009$. Therefore, $p_{12} = 0.017 - 0.009 = 0.008$ and

$$SE_{12} = \sqrt{0.017(1-0.017)/11034 + 0.009(1-0.009)/11037} = 0.0015.$$

The corresponding z-score and p-value are 5.25 and 1.5×10^{-7}. Therefore, we can comfortably reject the null hypothesis and conclude that the relationship between the two categorical variables (group and disease status) is statistically significant.

Question 9

Using R-Commander, we can exam the linear relationship between body temperature and heart rate, i.e., $H_A : \rho \neq 0$ vs. $H_0 : \rho = 0$, by clicking `Statistics` $\rightarrow$ `Summaries` $\rightarrow$ `Correlation test` and selecting the two variable.

```
        Pearson's product-moment correlation

data:  NormTemp$HeartRate and NormTemp$Temperature
```

```
t = 4.9562, df = 98, p-value = 3.011e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2754869 0.5920393
sample estimates:
      cor
0.4476808
```

The sample correlation coefficient in this case is 0.45, which is statistically significant (p-value = 2.01 $^{-6}$). That is, we can reject the null hypothesis $H_0 : \rho = 0$ and conclude that there is a strong linear relationship between the two variable.

Chapter 9
Analysis of Variance (ANOVA)

Question 1

Figure 9.1 shows the boxplots and plot of means (with the standard-deviation bars). The equal-variance assumption for ANOVA seems appropriate. Using R-Commander, we can perform ANOVA to examine the effectiveness of feed supplements: the observed value of F-statistic is $f = 15.37$, and the corresponding p-value is quite small. Therefore, we can reject the null hypothesis and conclude that based on this experiment, various feed supplements have quite different effects on the growth rate and the relationship between the two variables (feed type and weight) is statistically significant.

```
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 2

Figure 9.2 shows the plot of means, along with the corresponding confidence intervals, for the new variable `WeightGain = Postwt - Prewt`. While the amount of weight gain is close to zero on average for the control group (Cont), the averages are 3.0 and 7.2 for Cognitive Behavioral treatment (CBT) and Family Treatment (FT) respectively. Using R-Commander, we can perform ANOVA to examine the significance of overall changes in weight gain across different treatments. For this dataset, the observed value of F-statistic is $f = 5.422$, and the corresponding p-value is 0.0065. Therefore, we can reject the null hypothesis at 0.01 significance level and conclude that the treatments are in fact effective.

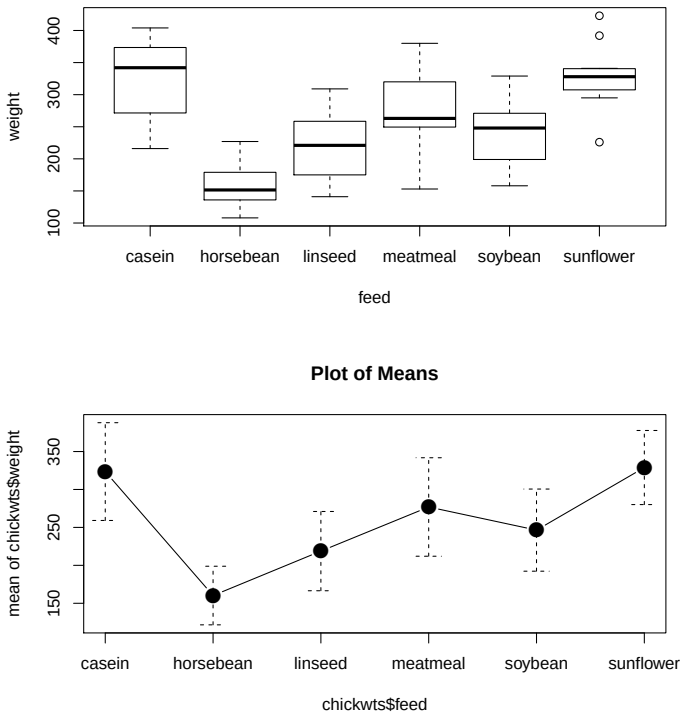

Fig. 9.1 Chapter 9 - Question 1

```
            Df Sum Sq Mean Sq F value Pr(>F)
Treat        2    615  307.32   5.422 0.0065 **
Residuals   69   3911   56.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5

Using a two-way ANOVA, where `Cult` and `Date` are the two factors, we find that `Cult` is significantly associated with `VitC` (p-value = 1.089×10^{-09}).

```
Anova Table (Type II tests)

Response: VitC
          Sum Sq Df F value    Pr(>F)
```

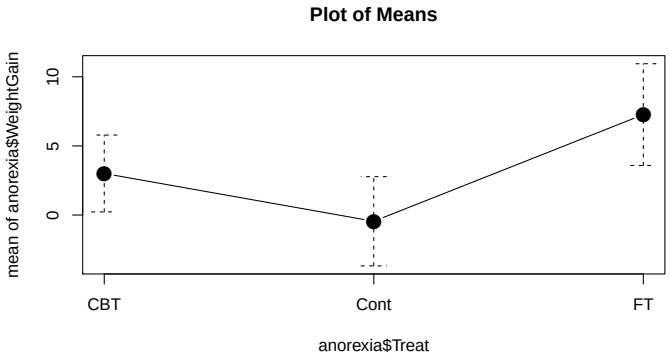

Fig. 9.2 Chapter 9 - Question 4

```
Cult      2496.2  1 54.1095 1.089e-09 ***
Date       909.3  2  9.8555 0.0002245 ***
Cult:Date  144.3  2  1.5640 0.2186275
Residuals 2491.1 54
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 10
Analysis of Categorical Vriables

Question 1

Under the null hypothesis, $\mu_0 = 0.2$. Therefore, the expected number of smokers out of $n = 150$ is $E_1 = n\mu_0 = 150 \times 0.2 = 30$, and the expected number of non-smokers is $E_2 = n(1-\mu_0) = 150 \times 0.8 = 120$. The observed numbers of smokers and non-smokers are $O_1 = 27$ and $O_2 = 123$ respectively. Therefore,

$$\begin{aligned} Q &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \\ &= \frac{(27-30)^2}{30} + \frac{(123-120)^2}{120} \\ &= 0.375 \end{aligned}$$

Using the χ^2 distribution with 1 degree of freedom, p-value = 0.54. Therefore, we cannot reject the null hypothesis.

Question 2

Import the `birthwt` dataset from the `MASS` package, convert `ht` and `low` to factors, and use `Statistics` → `Contingency tables` → `Two-way table` to examine the relationship between these two categorical variables. The results show that the relationship is statistically significant (p-value = 0.036) at 0.05 level so we can reject the null hypothesis.

```
   low
ht    0   1
  0 125  52
  1   5   7
```

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 4.388, df = 1, p-value = 0.03619
```

Question 4

In R-Commander, click `Statistics` → `Contingency tables` → `Enter and analyze two-way tables`, then create a 4×2 table, enter the frequencies as shown in Figure 10.1, and press OK. The results of Pearson's χ^2 test of independence show that the relationship between snoring and heart disease is statistically significant (p-value = 1.082×10^{-15}).

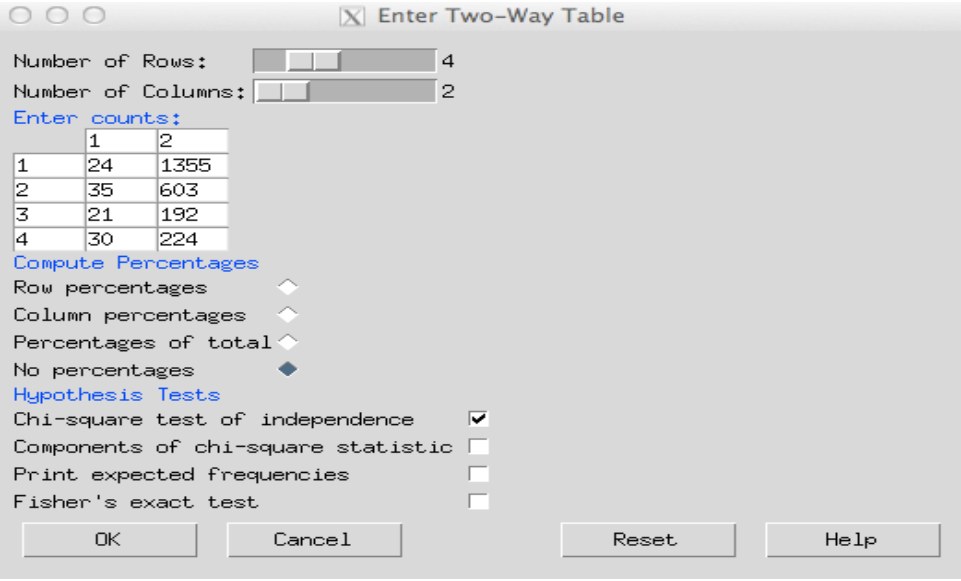

Fig. 10.1 Chapter 10 - Question 4

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 72.7821, df = 3, p-value = 1.082e-15
```

Chapter 11
Regression Analysis

Question 1

a) After uploading BodyTemperature.txt to R-Commander, click `Statistics` → `Fit models` → `Linear model` and select `Temperature` and `HeartRate` as the response variable and the predictor respectively (Figure 11.1).

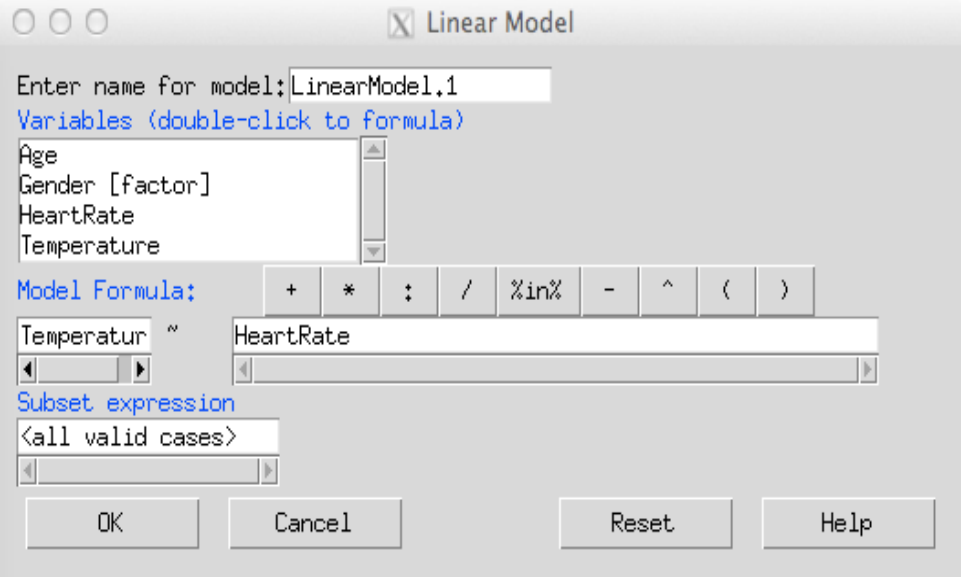

Fig. 11.1 Chapter 11 - Question 1

```
Call:
lm(formula = Temperature ~ HeartRate, data = NormTemp)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.50562 -0.46473  0.00543  0.48943  2.53943

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.39068    1.20144  76.900  < 2e-16 ***
HeartRate    0.08063    0.01627   4.956 3.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.86 on 98 degrees of freedom
Multiple R-squared: 0.2004,        Adjusted R-squared: 0.1923
F-statistic: 24.56 on 1 and 98 DF,  p-value: 3.011e-06
```

b) The estimate of the regression coefficient of heart rate is $\hat{\beta}_1 = 0.08$. This means that on average, one unit increase in heart rate coincides with $0.08\,^{\circ}$F increase in body temperature. The relationship is statistically significant with p-value= 3.01×10^{-6}.

c) 95% CI for β_1 is $[0.08 - 2 \times 0.016, 0.08 + 2 \times 0.016] = [0.048, 0.112]$. You can also obtain the confidence interval for β_1 by clicking `Models` $\rightarrow$ `Confidence intervals`.

d) $R^2 = 0.2004$ which is equal to the square of sample correlation coefficient $r = 0.4477$.

e) Click `Models` $\rightarrow$`Graphs`$\rightarrow$`Basic diagnostic plots` to obtain simple diagnostic plots similar to Figure 11.2. Observations 6, 75, and 86 are identified as potential outliers. We should not remove these observations from the data; rather, we investigate them to make sure their values are not entered by mistake.

f) According to our model,

$$\begin{aligned} \hat{y} &= 92.39 + 0.08x \\ &= 92.39 + 0.08 \times 75 \\ &= 98.39 \end{aligned}$$

Question 2

a) We repeat the steps in Question 1, but this time, under `Model Formula`, we enter `temperature` $\sim$ `HeartRate + Gender`.

```
Call:
lm(formula = Temperature ~ HeartRate + Gender, data = NormTemp)
```

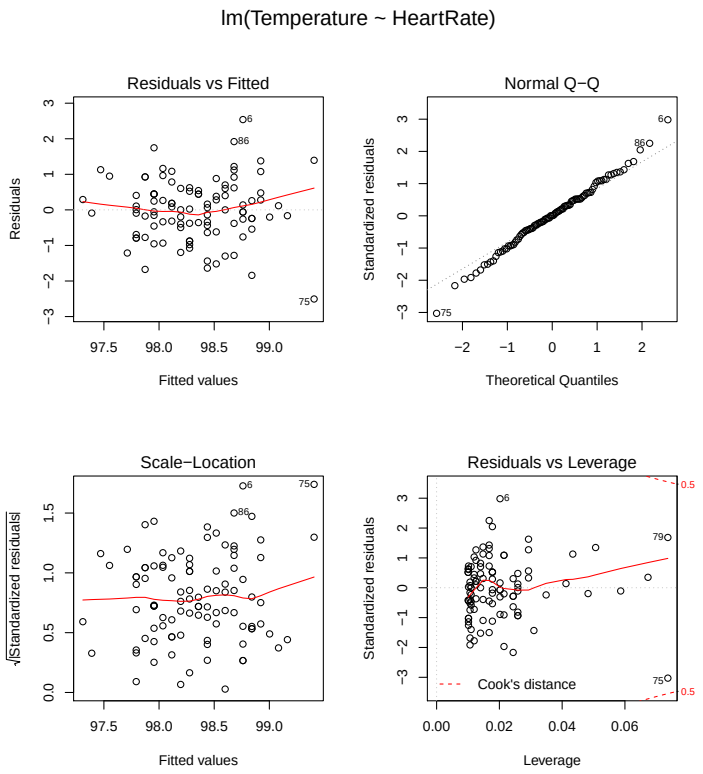

Fig. 11.2 Chapter 11 - Question 1-e

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.37056 -0.48862 -0.00963  0.53575  2.68538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.43764    1.18902  77.743  < 2e-16 ***
HeartRate    0.08199    0.01612   5.088 1.77e-06 ***
GenderM     -0.30044    0.17041  -1.763    0.081 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8509 on 97 degrees of freedom
Multiple R-squared: 0.2252,        Adjusted R-squared: 0.2093
F-statistic:  14.1 on 2 and 97 DF,  p-value: 4.212e-06
```

b) R^2 increases from 0.2004 to 0.2252 because of the additional predictor (gender) added to the model.

c) Based on this model, the estimates of regression coefficients for heart rate and gender are $\hat{\beta}_1 = 0.08$ and $\hat{\beta}_2 = -0.30$ respectively. Among people with the same gender (i.e., keeping the gender variable fixed), one unit increase in heart rate coincides with $0.08\,^{\circ}$F increase in body temperature on average. On the other hand, for a given heart rate (i.e., keeping the heart rate variable fixed), the expected (average) body temperature drops by $0.3\,^{\circ}$F for the male group. The latter is not statistically significant at 0.05 level (p-value = 0.081), but it is significant ant 0.1 level. Note that "M" at the end of `GenderM` (under `Coefficients`) indicates that the estimated regression coefficient is based on regarding the male group as 1 and the female group as 0 (i.e., the baseline group) for the binary random variable `gender`.

d) The 95% confidence intervals for β_1 and β_2 are as follows:

$$\begin{aligned}
\beta_1 : &\quad [0.08 - 2 \times 0.016, 0.08 + 2 \times 0.016] = [0.048, 0.112] \\
\beta_2 : &\quad [-0.30 - 2 \times 0.17, 0.08 + -0.30 - 2 \times 0.17] = [-0.64, 0.04]
\end{aligned}$$

You can also obtain the confidence intervals by clicking `Models` $\rightarrow$ `Confidence intervals`.

e) For a woman whose heart rate is 75,

$$\begin{aligned}
\hat{\text{Temperature}} &= 92.44 + 0.08\text{HeartRate} - 0.3\text{Gender} \\
&= 92.44 + 0.08 \times 75 - 0.3 \times 0 \\
&= 98.44
\end{aligned}$$

For a man whose heart rate is 75,

$$\begin{aligned}
\hat{\text{Temperature}} &= 92.44 + 0.08\text{HeartRate} - 0.3\text{Gender} \\
&= 92.44 + 0.08 \times 75 - 0.3 \times 1 \\
&= 98.14
\end{aligned}$$

Chapter 12
Clustering

Question 3

Click `Statistics` → `Dimensional analysis` → `Cluster analysis` → `k-means cluster analysis`, select all four variables, and set the number of clusters to 3.

```
> .cluster$size # Cluster Sizes
```

```
[1] 38 62 50
```

```
> .cluster$centers # Cluster Centroids
```

```
  new.x.Petal.Length new.x.Petal.Width new.x.Sepal.Length
1           5.742105          2.071053           6.850000
2           4.393548          1.433871           5.901613
3           1.462000          0.246000           5.006000
  new.x.Sepal.Width
1          3.073684
2          2.748387
3          3.428000
```

Informally (simply by focusing on the differences among the three centroids and assuming the four variables have comparable variances), it seems that the three clusters are very different with respect to petal length and width. In general, judging the importance of variables simply based on centroids could be misleading. This would be a more reasonable approach if we first standardize the variables (`Data` → `Manage variables in active data set` → `Standardize variables`).

The following contingency table shows the relationship between clusters (identified by the newly created variable `KMeans`) and the three species of flowers,`Species`. As we can see, there is a strong relationship between the identified clusters and `Species`.

```
      Species
KMeans setosa versicolor virginica
     1      0          2        36
     2      0         48        14
     3     50          0         0
```

Question 4

We follow similar steps as discussed above, but this time, we use `Hierarchical cluster analysis` instead of `k-means cluster analysis` and select `Complete Linkage` as the clustering method. R-Commander clusters the data hierarchically and provides the corresponding dendrogram. Next, we create a cluster identifier by cutting the three and dividing the data into three clusters: click `Statistics` → `Dimensional analysis` → `Cluster analysis` → `Add hierarchical clustering to data set` and set the number of clusters to 3. We can then create contingency table to examine the relationship between clusters and `Species`:

```
           Species
hclus.label setosa versicolor virginica
          1     50          0         0
          2      0         23        49
          3      0         27         1
```

Chapter 13
Bayesian Analysis

Question 1

For this problem, $\alpha = 1, \beta = 10, n = 50$, and $y = 8$. Therefore, given the observed data, the posterior probability distribution of μ is Beta(1+8, 10+50-8) = Beta(9, 52). We can now use R-Commander to plot the pdf of the posterior distribution: click `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Beta distribution` $\rightarrow$ `Plot beta distribution`, and set `Shape 1` and `Shape 2` to 9 and 52 respectively (Figure 13.1).

For the 95% credible interval, we can use the 0.025 and 0.975 quantiles: click `Distributions` $\rightarrow$ `Continuous distributions` $\rightarrow$ `Beta distribution` $\rightarrow$ `Plot quantiles`. The interval is [0.071, 0.246]. We can use the posterior mean as our point estimate: $\hat{\mu} = 9/(9+52) = 0.148$. This is slightly lower than the sample proportion: $p = 8/50 = 0.160$. Finally, to examine our hypothesis that $\mu < 0.2$, we use R-Commander to find the lower probability of 0.2: $P(\mu \leq 0.2|Y = 8) = 0.87$.

Question 2

We now use Beta(9, 52) as our prior so $\alpha = 9$ and $\beta = 52$. In the new study, $n = 30$ and $y = 6$. Therefore, the posterior distribution of μ is Beta (9+6, 52 + 30 - 6) = Beta(15, 76). The 95% credible interval for μ based on this distribution is [0.096, 0.247].

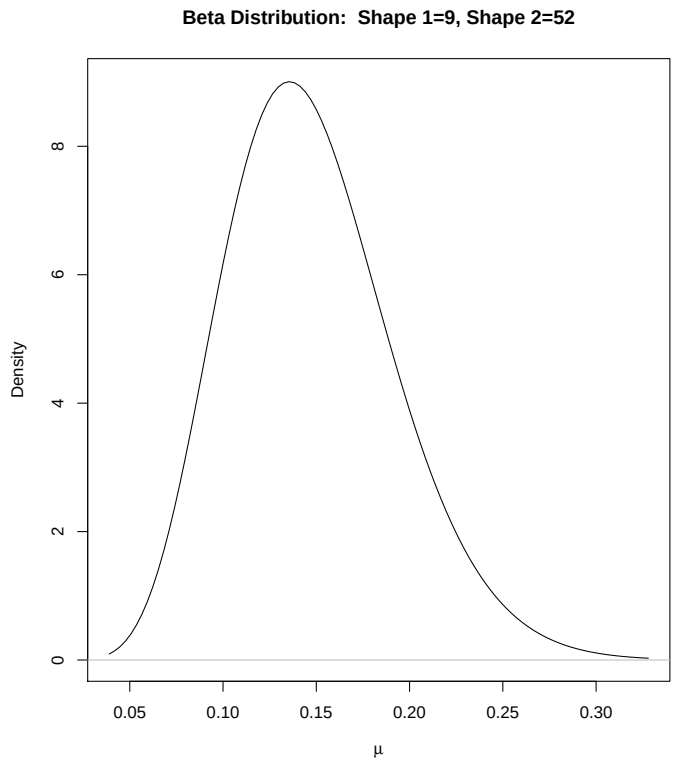

Fig. 13.1 Chapter 13 - Question 1

Question 3

We now suppose that we start with our initial prior, Beta(1, 10), and update the prior at the end of the second study; that is, $\alpha = 1, \beta = 10, n = 80$, and $y = 14$. The posterior distribution in this case is Beta(1+14, 10+80-14) = Beta(15, 76). This is the same distribution we obtained by updating our prior gradually. Therefore, our inference does not change whether we use the data gradually as they arrive or wait until we have all the data.

Question 6

We can use Beta(3, 7), shown in Figure 13.2, as our prior. According to this distribution, the range of values in the neighborhood of 0.3 has a large prior probability, whereas for the range values close to 1, the prior probability is close to zero (it is

not exactly zero; in general, we should never give zero prior probability to a range of possible values even if we do not think those values are very probable; we should always give data a chance to change our mind.) We now use the `birthwt` data to

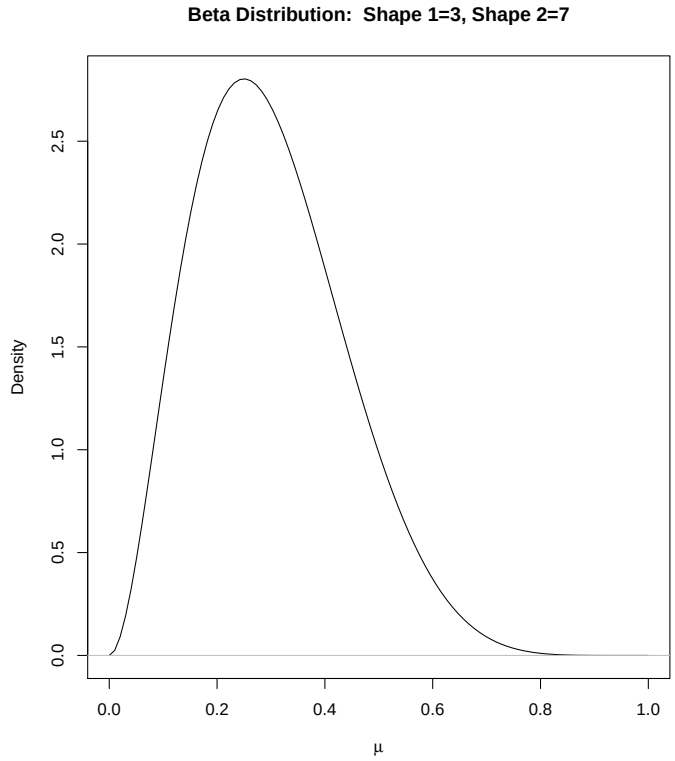

Fig. 13.2 Chapter 13 - Question 6

update this prior. Out of $n = 189$ babies in this dataset, $y = 59$ of them have low birthweight. Therefore, the posterior distribution is Beta(3+59, 7+189-59) = Beta(62, 137). We now use this distribution to find the probability that μ is in [0.25, 0.35] interval (i.e., within 0.05 from 0.3): $P(0.25 < \mu \leq 0.35) = 0.878 - 0.026 = 0.852$. The probability that μ would be outside of this range is 1-0.852 = 0.148.