Module 1 solutions

Module 2: Solutions to Learning Activities

Activity 2.4

Using the health survey data (`Activity_S2.4.xlsx`) described in the computing notes of this module, create a new variable, BMI, which is equal to a person's weight (in kg) divided by their height (in metres) squared (i.e. BMI $= \frac{\text{weight (kg)}}{[\text{height (m)}]^2}$. Categorise BMI using the WHO categories:

- Underweight: BMI < 18.5
- Normal weight: 18.5 $\leq$ BMI < 25
- Pre-obesity: 25 $\leq$ BMI < 30
- Obesity Class I: 30 $\leq$ BMI < 35
- Obesity Class II: 35 $\leq$ BMI < 40
- Obesity Class III: BMI $\geq$ 40

Create a two-way table to display the distribution of BMI categories by sex (sex: 1 = respondent identifies as male; 2 = respondent identifies as female). Does there appear to be a difference in categorised BMI between males and females?

Answers

Table 1: CAPTION

BMI category	Male	Female	Total
Underweight	6 (1.2%)	12 (1.9%)	18 (1.6%)
Normal weight	134 (26.1%)	228 (36.4%)	362 (31.8%)
Pre-obesity	216 (42.1%)	195 (31.1%)	411 (36.1%)
Obesity Class I	95 (18.5%)	106 (16.9%)	201 (17.6%)
Obesity Class II	46 (9.0%)	55 (8.8%)	101 (8.9%)
Obesity Class III	16 (3.1%)	31 (4.9%)	47 (4.1%)
Total	**513 (100.0%)**	**627 (100.0%)**	**1,140 (100.0%)**

From this health survey, it appears that men are more likely to have BMIs indicating Pre-Obesity (men 42% vs women 31%) and Obesity Class I (men 19% vs women 17%), compared to women who are more likely to have BMIs indicating Normal weight (women 36% vs men 26%).

Process

We first read the Excel data into R, using the readxl package. It is useful to examine the dataset - here using the `summary()` function:

```
library(readxl)
library(jmv)

survey <- read_excel("data/activities/Activity_S2.4-health-survey.xlsx")
summary(survey)
```

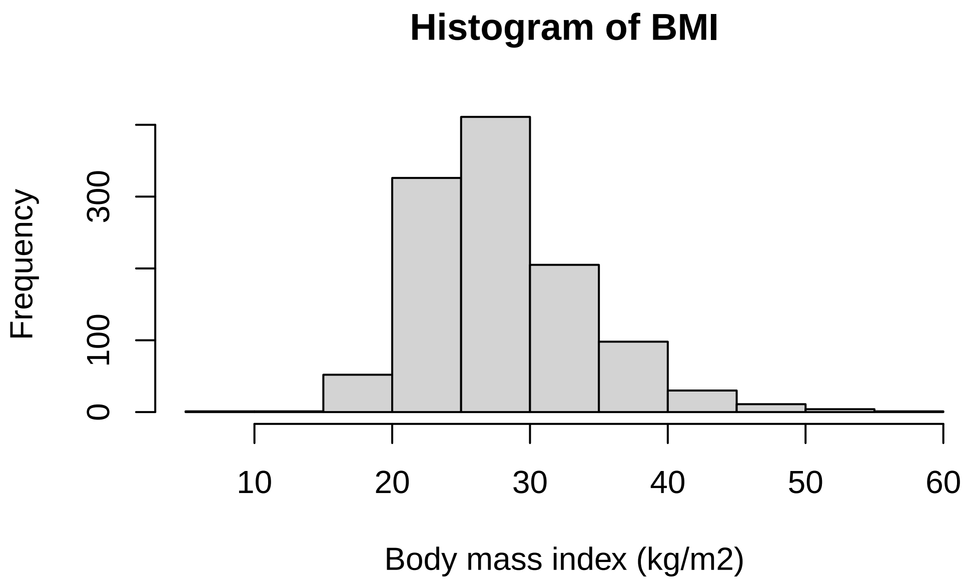
```
      sex            height          weight     
 Min.   :1.00   Min.   :1.220   Min.   : 22.70  
 1st Qu.:1.00   1st Qu.:1.630   1st Qu.: 68.00  
 Median :2.00   Median :1.700   Median : 79.40  
 Mean   :1.55   Mean   :1.698   Mean   : 81.19  
 3rd Qu.:2.00   3rd Qu.:1.780   3rd Qu.: 90.70  
 Max.   :2.00   Max.   :2.010   Max.   :213.20  
```

Note that has been entered as a numeric variable. We should define sex as a factor, and then create BMI. After creating BMI, we should examine its distribution using a histogram and/or a boxplot:

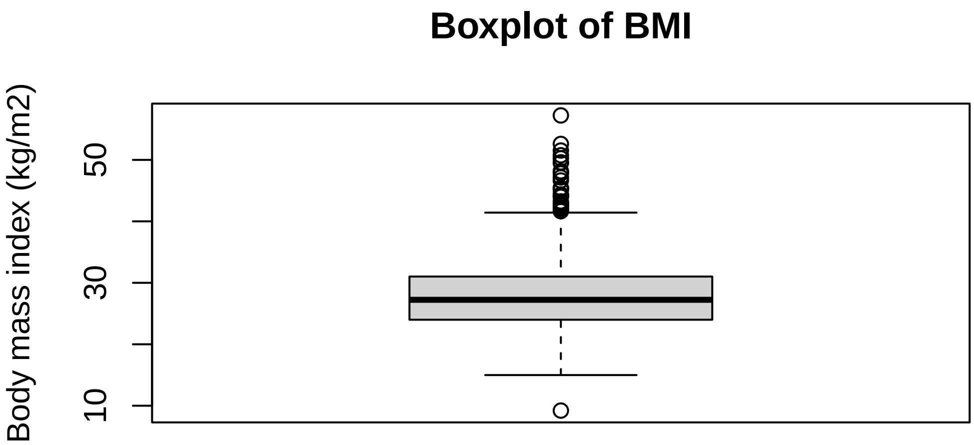
```r
# Define sex as a factor
survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))

# Create BMI
survey$bmi = survey$weight / (survey$height^2)

# Examine the distribution of BMI
hist(survey$bmi, main="Histogram of BMI", xlab="Body mass index (kg/m2)")
```




```r
boxplot(survey$bmi, main="Boxplot of BMI", ylab="Body mass index (kg/m2)")
```



The boxplot in particular shows that there are some extreme values of BMI. We can examine some records using the `subset()` function:

```
subset(survey, bmi<15)
```

sex	height	weight	bmi
Female	1.57	22.7	9.21
Female	1.65	40.8	15

```
subset(survey, bmi>45)
```

sex	height	weight	bmi
Female	1.52	105	45.4
Male	1.85	174	50.8
Female	1.22	74.8	50.3
Male	1.93	213	57.2
Female	1.63	127	47.8
Female	1.55	115	48
Female	1.65	131	48.2
Female	1.55	109	45.3
Male	1.78	143	45.1
Female	1.65	127	46.6
Female	1.63	132	49.5
Female	1.7	152	52.6
Female	1.6	127	49.6
Female	1.5	106	47.2
Female	1.73	154	51.5
Female	1.6	116	45.4

The smallest BMI of 9.2 kg/m2 is very low, with a weight of 22.7 kg. We should check the recorded height and weight values against the original data (paper records, survey responses) if they were available. However, as a weight of 22.7kg is not impossible, this record will not be deleted. An alternative approach would be to analyse the data including the very low BMI and again excluding the very low BMI as a sensitivity analysis.

The largest BMI values are based on participants with large weights, and none of these seem biologically implausible. Therefore, no changes will be made to participants with small or large values of BMI.

We can use the `cut()` function to create the BMI categories. The WHO cutpoints are inclusive of the lower-bound, so we use right=FALSE. After creating the categories, it is good practice to check the resulting categories using `summary()`:

```
survey$bmi_cat <- cut(survey$bmi, c(0, 18.5, 25, 30, 35, 40, 100), right=FALSE)
summary(survey$bmi_cat)
```

```
[0,18.5) [18.5,25)    [25,30)    [30,35)    [35,40)  [40,100)
      18        362        411        201        101        47
```

Finally, we can create a two-way table using the contTables() function within the jmv package. We can define the rows by BMI category, and the columns by sex:

```
contTables(data=survey,
           rows = bmi_cat,
           cols = sex)
```

```
CONTINGENCY TABLES

Contingency Tables

  bmi_cat      Male    Female    Total

  [0,18.5)        6        12       18
  [18.5,25)     134       228      362
  [25,30)       216       195      411
  [30,35)        95       106      201
  [35,40)        46        55      101
  [40,100)       16        31       47
  Total         513       627     1140



 2 Tests

        Value       df    p

   2   22.49802      5    0.0004209
  N        1140
```

To assess whether there is a difference in BMI between males and females, we should look at the within-sex relative frequencies. In other words, column percents (for this table), by specifying pcCol = TRUE:

```
contTables(data=survey,
           rows = bmi_cat,
           cols = sex,
           pcCol = TRUE)
```

CONTINGENCY TABLES

Contingency Tables

bmi_cat		Male	Female	Total
[0,18.5)	Observed	6	12	18
	% within column	1.16959	1.91388	1.57895
[18.5,25)	Observed	134	228	362
	% within column	26.12086	36.36364	31.75439
[25,30)	Observed	216	195	411
	% within column	42.10526	31.10048	36.05263
[30,35)	Observed	95	106	201
	% within column	18.51852	16.90590	17.63158
[35,40)	Observed	46	55	101
	% within column	8.96686	8.77193	8.85965
[40,100)	Observed	16	31	47
	% within column	3.11891	4.94418	4.12281
Total	Observed	513	627	1140
	% within column	100.00000	100.00000	100.00000

^{2} Tests

	Value	df	p
^{2}	22.49802	5	0.0004209
N	1140		