

Consulting Homework 4

Tim Vigers

September 30, 2018

Data examination steps:

- Looked at histograms and boxplots (I know these generally aren't detailed enough, but it helps me spot outliers) of height, weight, and bmi, which raises a lot of questions:
- First, weight ranges from 41 to 288, which makes me think that some of these weights were measured in kilograms and some were in pounds (it's unlikely that anyone in this cohort weighs just 41 pounds). Also, you can see in the histogram that there's an unexpected bump in the 50-100 range. So it might be worth confirming with the data manager that units are consistent between hospitals. Based on the weight boxplot by hospital, it looks like hospitals 1-16 are different from the rest.
- The height histogram and boxplots are much more regular, and there don't appear to be any noticeable differences between hospitals. However, the range is 45 - 70, so I'm not sure what units they're using. It's clearly not meters, but if it's inches then the tallest person in the cohort is 5'10" and the shortest is 3'9", which seems very unlikely. So either way, I'd like to confirm the height values and units with the data manager.
- BMI looks overall like it's what you'd expect (mostly in the 20-30 range, which is normal and slightly overweight), and there isn't an obvious difference by hospital. However, there are 3 obvious outliers with BMIs of 3, 72, and 75. These seem unlikely to be correct, although BMIs in the 70s are technically possible. Also, there are quite a few in the 40-60 range, which is very high. Since this is a CVD study, that might be correct, but worth double checking. The most concerning thing about the BMI scores though, is that if you calculate them yourself based on the weight and height values, the calculated values don't match the reported ones (whether you calculate assuming pounds and inches or kilograms and centimeters).

Continuous variable ranges:

```
range(va$height)
```

```
## [1] 45 70
```

```
range(va$weight,na.rm = TRUE)
```

```
## [1] 41 288
```

```
range(va$bmi,na.rm = TRUE)
```

```
## [1] 3 75
```

BMI calculation check:

```
va$bmi_kg_calc <- round(va$weight / ((va$height/100)^2),0)
va$bmi_lbs_calc <- round(703 * (va$weight / (va$height^2)),0)
va$bmi_kg_calc[which(va$bmi_kg_calc == va$bmi)]
```

```
## numeric(0)
```

```
va$bmi_lbs_calc[which(va$bmi_lbs_calc == va$bmi)]
```

```
## numeric(0)
```

Check missing continuous data:

- There are no missing height values, and the participants with missing weight values are also missing bmi. So that's a good sign!

```
length(which(is.na(va$weight)))
```

```
## [1] 104
```

```
length(which(is.na(va$height)))
```

```
## [1] 0
```

```
length(which(is.na(va$bmi)))
```

```
## [1] 104
```

```
FALSE %in% (which(is.na(va$weight)) == which(is.na(va$bmi)))
```

```
## [1] FALSE
```

Check categorical variables:

- Make sure the levels in categorical variables make sense. The hospital codes look good, but we are missing data for six month period 38. It also looks like there are about twice as many values for six month period 37, so I'm wondering if 37 and 38 were accidentally combined. Also, there are two procedures called "2", and I'm not sure what that means, since it should just be 0 or 1. Finally, a huge proportion of the participants have an ASA score of 4. Again, this might be expected for this cohort, but is probably worth double checking.

```
table(va$hospcode)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
## 654 612 614 588 574 577 615 585 606 592 574 581 586 631 571 595 600 620
## 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
## 572 599 619 584 602 621 622 650 608 614 575 558 602 575 567 606 581 557
## 37 38 39 40 41 42 43 44
## 641 566 599 619 617 576 575 575
```

```
summary(va[,c(1:3,7)])
```

```
##      hospcode      sixmonth      proced      asa
## 1      : 654    34:4444    0 : 5128    1 : 15
## 26     : 650    35:4393    1 :21019   2 : 1144
## 37     : 641    36:4399    2 : 2      3 : 5137
## 14     : 631    37:8679   NA's: 106   4 :19744
## 25     : 622    39:4340           5 : 66
## 24     : 621           NA's: 149
## (Other):22436
```

Check for missing categorical data:

- There are 106 missing procedures and 149 missing ASA scores. It might be worth asking PIs about how to handle missing procedure and ASA data. My guess is that they would want to exclude anyone without procedure data, but keep in those without ASA data (since ASA is subjective anyway, and not a huge part of the analysis as far as I know).

Check everything by procedure code:

- Overall the two procedures look pretty similar to me, so the issues with the data don't appear to be related to the procedure. There are a lot more with procedure 1, but I think that's to be expected.

```
proc.0 <- as.data.frame(split.data.frame(va,va$proced)[1])
proc.1 <- as.data.frame(split.data.frame(va,va$proced)[2])
summary(proc.0)
```

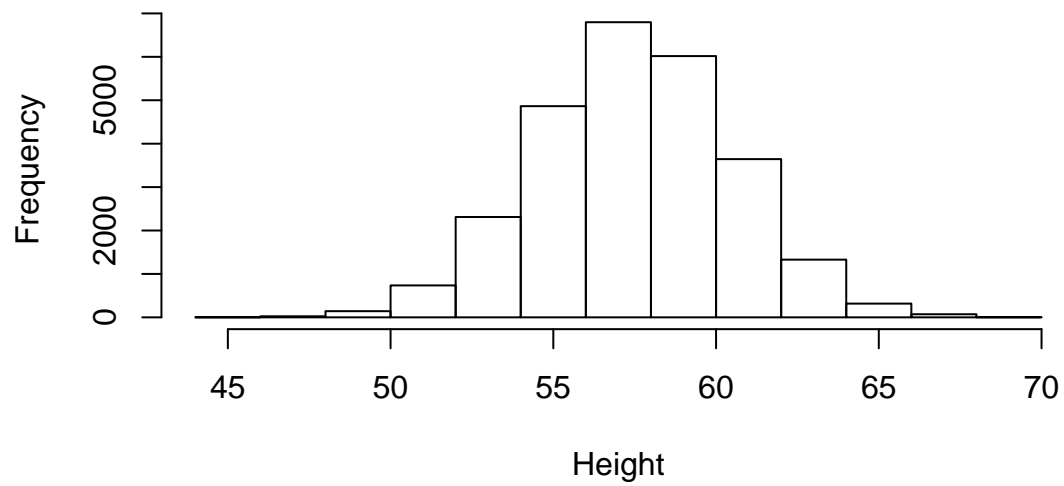
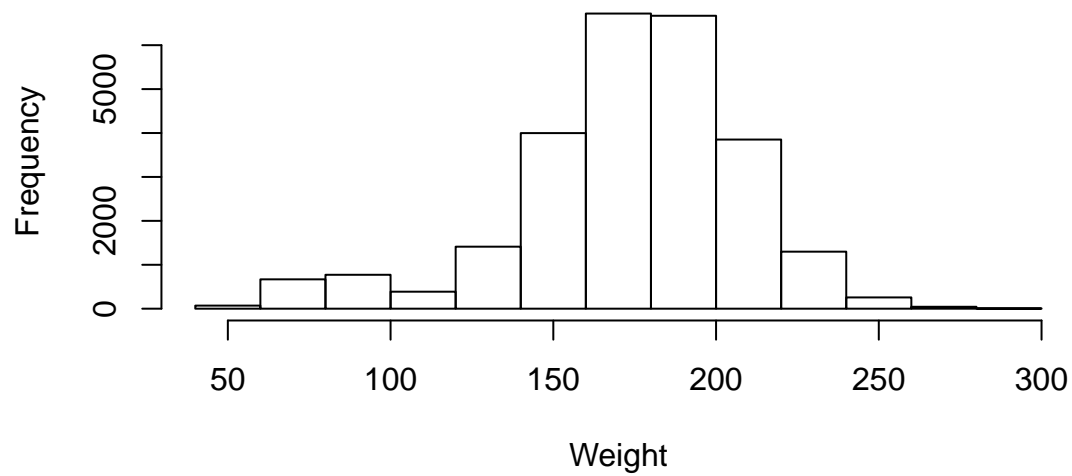
```
##      X0.hospcode      X0.sixmonth X0.proced      X0.weight      X0.height
## 26      : 143    34: 896      0:5128    Min.      : 50    Min.      :45.00
## 13      : 138    35: 829      1: 0      1st Qu.:159    1st Qu.:56.00
## 23      : 136    36: 827      2: 0      Median :178    Median :58.00
## 41      : 131    37:1734           Mean  :175    Mean   :58.01
## 16      : 130    39: 842           3rd Qu.:197    3rd Qu.:60.00
## 18      : 129           Max.   :283    Max.   :68.00
## (Other):4321           NA's   :24
##      X0.bmi      X0.asa      X0.bmi_kg_calc      X0.bmi_lbs_calc
## Min.      :13.00    1 : 1      Min.      :130.0    Min.      : 9.00
## 1st Qu.:24.00    2 : 237    1st Qu.: 459.0    1st Qu.:32.00
## Median :27.00    3 :1002    Median : 526.0    Median :37.00
## Mean   :27.29    4 :3851    Mean   : 524.3    Mean   :36.87
```

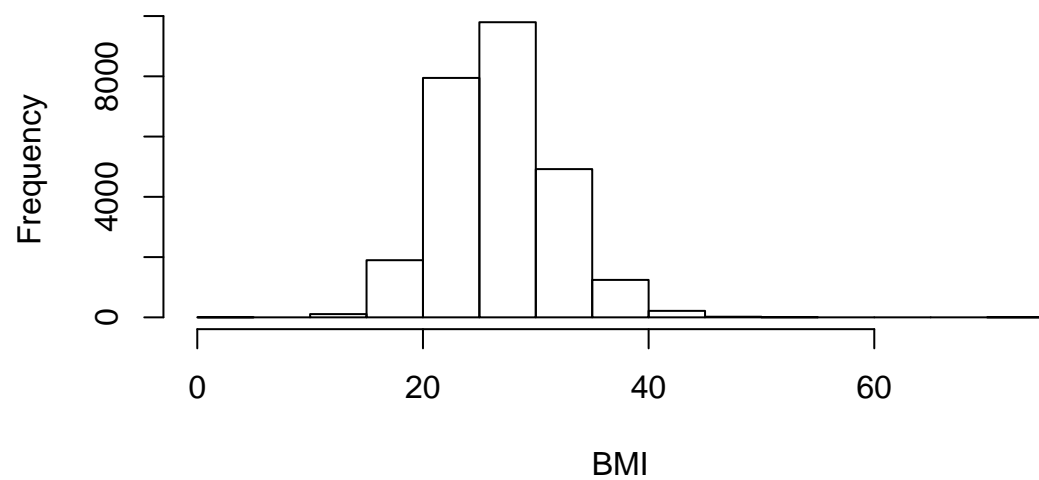
```
## 3rd Qu.:30.00 5 : 15 3rd Qu.: 597.0 3rd Qu.:42.00
## Max. :53.00 NA's: 22 Max. :1081.0 Max. :76.00
## NA's :24 NA's :24 NA's :24
```

```
summary(proc.1)
```

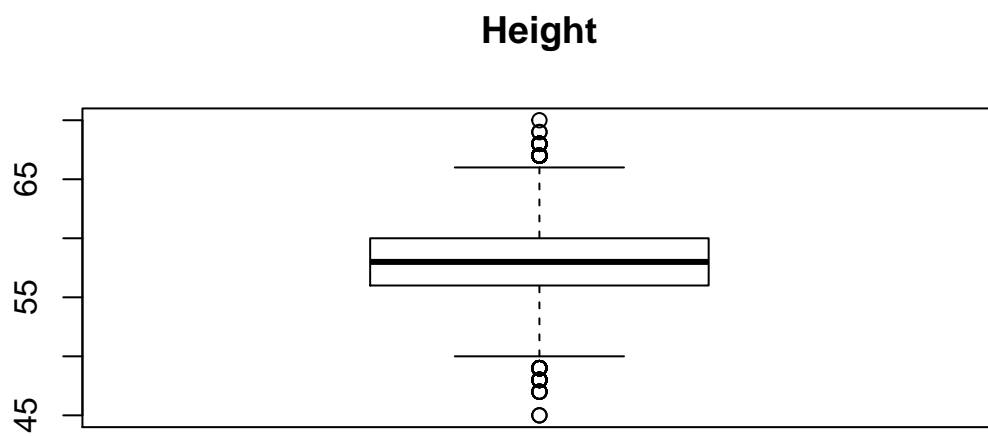
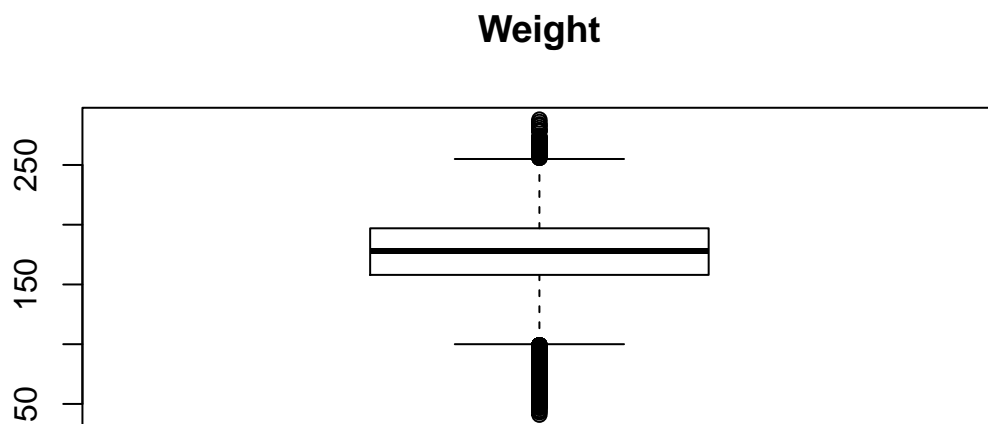
```
## X1.hospcode X1.sixmonth X1.proced X1.weight X1.height
## 1 : 540 34:3534 0: 0 Min. : 41.0 Min. :45
## 37 : 511 35:3548 1:21019 1st Qu.:157.0 1st Qu.:56
## 7 : 510 36:3557 2: 0 Median :178.0 Median :58
## 21 : 507 37:6907 Mean :173.8 Mean :58
## 40 : 507 39:3473 3rd Qu.:197.0 3rd Qu.:60
## 14 : 506 Max. :288.0 Max. :70
## (Other):17938 NA's :80
## X1.bmi X1.asa X1.bmi_kg_calc X1.bmi_lbs_calc
## Min. : 3.00 1 : 14 Min. : 118.0 Min. : 8.00
## 1st Qu.:24.00 2 : 900 1st Qu.: 456.0 1st Qu.:32.00
## Median :27.00 3 : 4114 Median : 526.0 Median :37.00
## Mean :27.25 4 :15815 Mean : 521.1 Mean :36.64
## 3rd Qu.:30.00 5 : 51 3rd Qu.: 597.0 3rd Qu.:42.00
## Max. :75.00 NA's: 125 Max. :1057.0 Max. :74.00
## NA's :80 NA's :80 NA's :80
```

Continuous variable histograms:

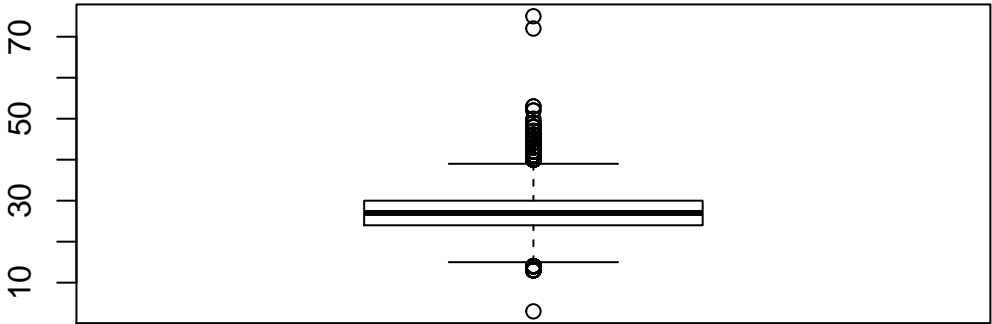




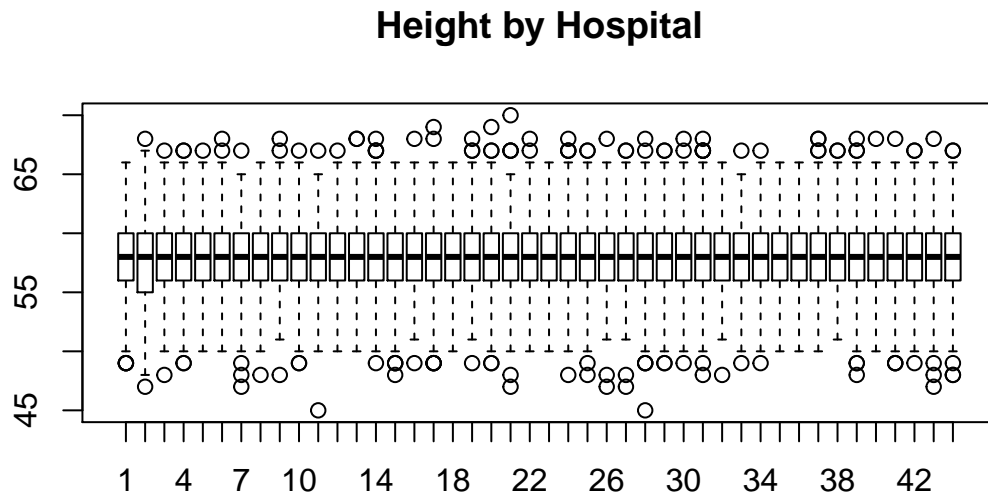
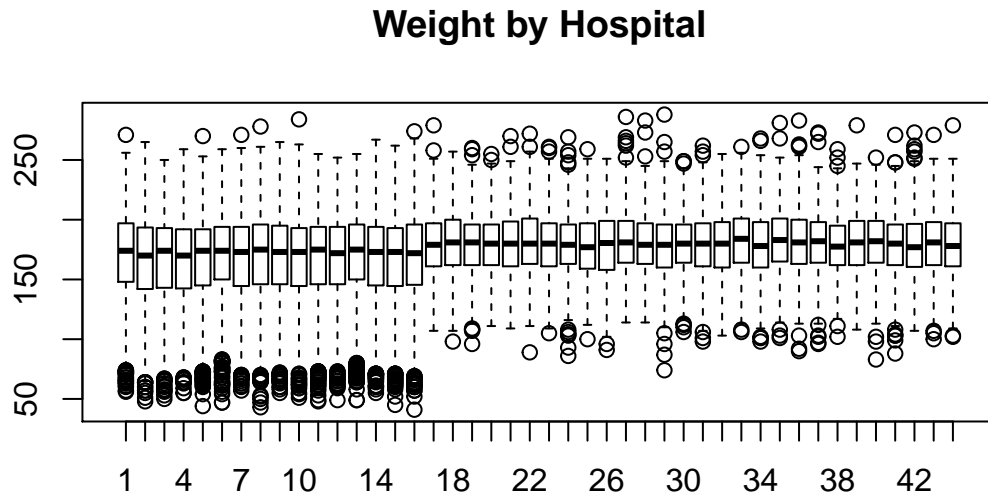
Continuous variable boxplots:



BMI



Continuous variable boxplots by hospital:



BMI by Hospital

