# Problem Set #1: MVIS 5301

*C. Todd Lombardo*

*11 January 2015*

This work can also be found online at: http://rpubs.com/ctodd/PS1

---

**1. What is the level of measurement for each of these variables (from the 2002 General Social Survey) (nominal, ordinal, interval/ratio)? Briefly justify your answer. *Note: Ignore the response categories of NAP, DK, and NA for purposes of this question.***

**CHILDS – How many children have you ever had?** 0-6. Actual Number; 7. Seven or more; 9. DK or NA

| Response | TOTAL | % |
|---|---|---|
| 0) None | 799 | 28.9 |
| 1) One | 469 | 17 |
| 2) Two | 657 | 23.8 |
| 3) Three | 481 | 17.4 |
| 4) Four | 185 | 6.7 |
| 5) Five | 73 | 2.6 |
| 6) Six | 40 | 1.4 |
| 7) Seven | 22 | 0.8 |
| 8) Eight or more | 34 | 1.2 |
| 9) No answer | 5 | 0.2 |
| TOTAL | 2765 | 100 |

*Ordinal.* At first glance, one might think the levels of measurement with these variables is *interval/ratio*, since a zero means an absence of children and there are equal distances between each measure. There are no "half-children", and if you have four children you have twice as many than if you only have two. However, the level of measurement is actually *ordinal* for this question since there is a grouping of responses at 'Eight or more' children. There's no difference if you have 8, 9 or 15 children for this grouping, but there is a difference between this grouping and any response less than eight, so it would not be considered *nominal* either.

**EDUC – Highest year of school completed** 0-20. Actual number of years; 9. DK or NA

| Response | TOTAL | % |
|---|---|---|
| 0) No formal schooling | 5 | 0.2 |
| 1) 1st grade | 2 | 0.1 |
| 2) 2nd grade | 15 | 0.5 |

| Response | TOTAL | % |
|---|---|---|
| 3) 3rd grade | 2 | 0.1 |
| 4) 4th grade | 6 | 0.2 |
| 5) 5th grade | 8 | 0.3 |
| 6) 6th grade | 25 | 0.9 |
| 7) 7th grade | 14 | 0.5 |
| 8) 8th grade | 66 | 2.4 |
| 9) 9th grade | 60 | 2.2 |
| 10) 10th grade | 88 | 3.2 |
| 11) 11th grade | 137 | 5 |
| 12) 12th grade | 818 | 29.6 |
| 13) 1 year of college | 265 | 9.6 |
| 14) 2 years | 374 | 13.5 |
| 15) 3 years | 157 | 5.7 |
| 16) 4 years | 377 | 13.6 |
| 17) 5 years | 93 | 3.4 |
| 18) 6 years | 128 | 4.6 |
| 19) 7 years | 43 | 1.6 |
| 20) 8 years | 70 | 2.5 |
| 98) Don't know | 5 | 0.2 |
| 99) No answer | 7 | 0.3 |
| TOTAL | 2765 | 100 |

Levels of measurement here are *interval/ratio* as a zero level indicates no years of schooling and the intervals equal in length. Four years of schooling is twice as long as 2 years of schooling.

---

**2. A student discovers that his grade on a recent test was in the 72nd percentile. If 90 students took the test, then approximately how many students received a higher grade than he did?**

Approximately 25 students

```
90-(90 * .72)
```

```
## [1] 25.2
```

---

**3. A sample of 7 underweight babies was fed a special diet and the following weight gains (lbs) were observed at the end of three months.**

```
6.7 2.7 2.5 3.6 3.4 4.1 4.8
```

**Find the mean and standard deviation for these 7 babies. Show your calculations (credit will not be given if you simply provide the answers).**

First, we find the mean (u), then we substract each variable from the mean (x-u), then square each

```
x=c(6.7,2.7,2.5,3.6,3.4,4.1,4.8)        # Entering the Dataset into a Vector

u=sum(x)/length(x)                       # Calculate the Mean

u                                        # Print the Mean
```

```
## [1] 3.971429
```

| x | x-u | (x-u)^2 |
|---|---|---|
| 6.7 | 2.7 | 7.4 |
| 2.7 | -1.3 | 1.6 |
| 2.5 | -1.5 | 2.2 |
| 3.6 | -0.4 | 0.1 |
| 3.4 | -0.6 | 0.3 |
| 4.1 | 0.1 | 0.0 |
| 4.8 | 0.8 | 0.7 |

We then sum those up, divide by number of variables minus one to calculate the variance, and take the square root to get the standard deviation.

```
Variance = sum((x-u)^2)/(length(x)-1)     # Calculate Variance

sqrt(Variance)                            # Calculate and print Standard Deviation
```

```
## [1] 1.437259
```

---

**4. The Nielsen Company publishes information on the TV-viewing habits of Americans in Nielsen Report on Television. A sample of 20 people yielded the weekly viewing times, in hours, in the table below.**

```
set1=c(25,41,27,32,43,66,35,31,15,5,34,26,32,38,16,30,38,30,20,21) # Dataset into a Vector
n=length(set1)               # Obtain number of variables
u=sum(set1)/length(set1)     # Mean
sorted = sort(set1)          # Sort the data into a new vector
                             # The median is the average of the two middle values, 30 and 31
medianvalues = sorted[(length(set1)/2)] + (sorted[(length(set1)/2)]+1)
medianvalues / 2
```

**a) Determine the median of these data.**

```
## [1] 30.5
```

```
# Q1 = 25th Percentile
quantile(sorted, 0.25)
```

**b) The quartiles of these data.**

```
## 25%
##  24
```

```
# Q3 = 75th Percentile
quantile(sorted, 0.75)
```

```
##   75%
## 35.75
```

```
# IQR = Q3 - Q1   <-- The interquartile range
IQR(sorted)
```

```
## [1] 11.75
```

**c) Obtain the lower and upper limits.**

Lower Limit

```
24 - 1.5*(IQR(sorted))              # Lower Limit = Q1 - 1.5 × IQR
```

```
## [1] 6.375
```

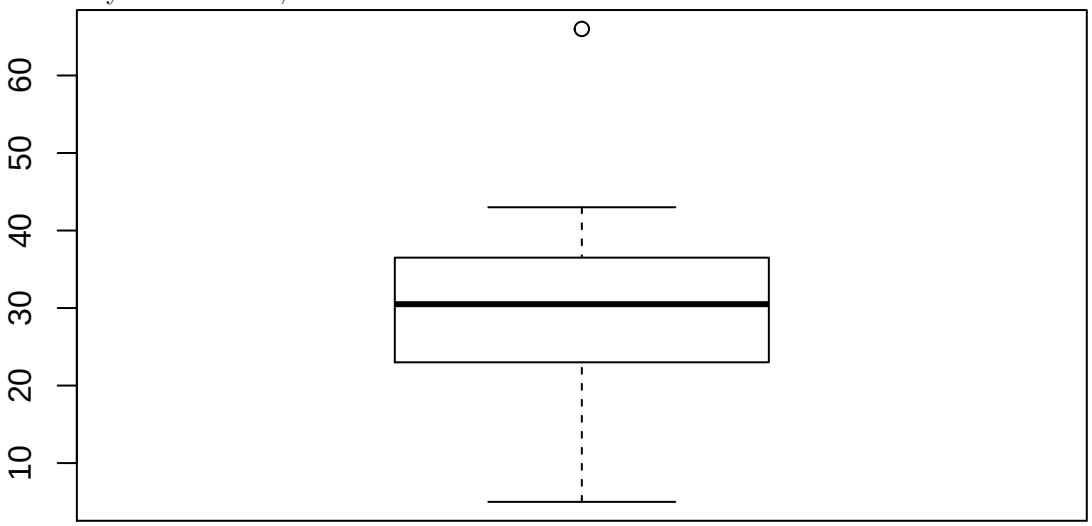
Upper Limit

```
35.75 + 1.5*(IQR(sorted))           # Upper Limit = Q3 + 1.5 × IQR
```

```
## [1] 53.375
```

**d) Determine potential outliers, if any.**

Potential outliers are 5 and 66 since they are outside the range of 6.375 and 53.375. A boxplot can easily visualize this, and shows that 66 is indeed an outlier.




---

**5. Use the Area under the standard normal curve table to obtain the shaded area under the standard normal curves below.**

**a. Area between z = -1.28 and 1.28**

The area under the curve for z = -1.28 is 0.1003. Multiply by two, then substract from 1: 1 - (2 *(0.1003)) = 0.7994 or 79.94% of the area under the normal distribution curve.

**b. Area between z = -1.64 and 1.64**

The area under the curve for z =-1.64 is 0.0505. Similar to case a: multiply by two, then substract from 1: 1 - (2 *(0.0505)) = 0.899 or 89.90% of the area under the normal distribution curve.

**c. Area below z = -1.96 and above 1.96**

The area under the curve for z = -1.96 is 0.0250, so the area would be double that: 0.05, or 5% of the total area under the normal distribution curve.

**d. Area below z = -2.33 and 2.33**

The area under the curve for z = -2.33 is 0.0102, so the area would be double that: 0.0204, or 2.04% of the total area under the normal distribution curve.