childsmath $\int_M d\omega = \int_{\partial M} \omega$

R Tutorials

Creating a Grouped Frequency Distribution on R

We will now create a frequency table for the numeric variable `IQ` in the databank data set.

Watch this Video First!!

1. Open the `databank.RData` workspace that you created in the introductory lab using `File->Load Workspace` (on a **Mac:** `workspace > load workspace file`)

2. Extract the IQ column and remove any `NA`'s:

```
IQ<-databank$IQ
IQ<-IQ[!is.na(IQ)]
IQ
```

```
[1] 118 105 115 108 106 111  99 106 111 109 115 114 101 109  98 103 103  99 110
 [20] 108 112 106 116 121 122 103 119 116 105 123 113  99 118 100 106 121 129 109
 [39] 121 117 102 123 103 111 114 107 106  95 115 100 116 103 112 119 131 100 109
 [58] 126 122 121  98 105 111 127 129 103 121  98 113  99 101 128 104 103 103 114
 [77] 126 123 121 131 105 119 116 127 114 119 117  88  98  96 102 102 100 101 105
 [96] 103  95  99 113 116
```

3. We use the `range` function to determine the range of values for the `IQ` variable:

```
range(IQ)
```

```
[1]  88 131
```

Since the values range from 88 to 131, we will make a frequency table using the following groups,

80-90, 90-100, 100-110, 110-120, 120-130, 130-140.

4. Now we create the desired grouped frequency table by first grouping the observations in the `IQ` variable into intervals using the `cut()` function:

```
group=cut(IQ,breaks=c(80,90,100,110,120,130,140),right=FALSE)
group
```

```
[1] [110,120) [100,110) [110,120) [100,110) [100,110) [110,120) [90,100)
  [8] [100,110) [110,120) [100,110) [110,120) [110,120) [100,110) [100,110)
...
```

By using the option `right=FALSE` we have instructed R to NOT include the right endpoint in each interval. The above output tells us that the first observation in the `IQ` variable is in the interval `[110,120)` (and that interval does NOT include 120), the second observation is in the interval `[100,110)` (which does NOT include 110), etc.

5. Then we create the frequency table using the `table()` function:

```
count<-table(group)
count
```

```
  [80,90)   [90,100) [100,110) [110,120) [120,130) [130,140)
        1         12         37         30         18          2
```

Since there are no decimals in the `IQ` variable, the above output tells us that 1 person in the data set has an IQ between 80 and 89 (inclusive), 12 people have an IQ between 90 and 99 (inclusive), etc.

6. Now we find the percentage of observations in each group:

```
n<-length(IQ)
percent<-count/n*100
percent
```

```
  [80,90)   [90,100) [100,110) [110,120) [120,130) [130,140)
        1         12         37         30         18          2
```

The above output gives us the percentage of observations in each interval. But in our case the IQ variable has 100 observations, so the percentage in each interval is the same as the number of observations in each interval.

7. To find the cumulative percents, we use the `cumsum` command :

```
cumpct<-cumsum(percent)
cumpct
```

```
  [80,90)   [90,100) [100,110) [110,120) [120,130) [130,140)
        1         13         50         80         98        100
```

The above output tells us, for example, that 80% of the observations in the `IQ` variable are less than 120.

8. Finally we print the complete frequency table in column format using the `cbind()` function:

```
cbind(count,percent,cumpct)
```

```
          count percent cumpct
[80,90)       1       1      1
[90,100)     12      12     13
[100,110)    37      37     50
[110,120)    30      30     80
[120,130)    18      18     98
[130,140)     2       2    100
```