

Sample Answer Lab 02:

Handout date: Wednesday, September 2, 2020

Due date: Monday, September 14, 2020 via **SUBMITLAB02** link in eLearning

Task 1: Identify and justify the measurement levels of several statistical variables (1 point)

Justify your selection of the measurement scale. You may want to look up Wikipedia for some of the variables

- a. Give an example each for nominal, ordinal, interval and ratio scaled measurement. Do not use examples from the lecture or lab. (0.4 points)
 - Credit score (300-850) (Interval scaled)
 - Number of break-ins in a neighborhood (Ratio scaled)
- b. The longitude and latitude in degrees on the earth's spherical surface. Be cautious in your arguments with respect to the origins of the coordinate system. That is, are distances between two longitudes constant all over the globe? (0.2 points)
 - a. Longitude: interval scaled only at a given latitude. Because longitudes and latitudes are circular measures, the selection of the origin is arbitrary. Therefore, they cannot be ratio scaled. Distances between two longitudes dependent on the latitude at which they are measured, using the cosine of the latitude as adjustment factor. The distance adjustment factor at the equator is $\cos(0^\circ) = 1$ whereas at both poles it is $\cos(90^\circ) = 0$. Distances between longitudes at a given latitude are comparable, however, distances between longitudes at different latitudes do not correspond with each other.
 - b. Latitude: Interval scaled. Because it is a circular measure, it does not have natural zero and the difference between two latitudes are almost identical with a slight elliptical distortion due the rotational forces exerted onto the earth. The table below show the distances between 1 degrees of latitude and longitude for an elliptical model (e.g., WGS84) of the earth surface. Therefore, the distances between 1° latitudes are not perfectly constant:


ϕ	Δ_{lat}^1	Δ_{long}^1
0°	110.574 km	111.320 km
15°	110.649 km	107.550 km
30°	110.852 km	96.486 km
45°	111.132 km	78.847 km
60°	111.412 km	55.800 km
75°	111.618 km	28.902 km
90°	111.694 km	0.000 km

The $1/\cos(\text{latitude})$ adjusted longitudinal distances lead to the Mercator map projection.

- c. The wind direction in degrees. (0.1 points)
Interval scaled. Because it is also a circular measure without a natural origin.
- d. Grouping of census block groups into neighborhoods. (0.1 points)
Ordinal scaled. The *smaller* census enumeration units are embedded in larger units without constant distance measure: Block < Block Groups < Tract < County < State < Region
- e. Income brackets for taxation purposes (0.1 points)
Ordinal scaled. The higher income class need to pay more tax, but the increments are not constant.
- f. Elevation above sea-level at a fixed point in time. (0.1 points)
Ratio scaled. It has meaningful zero. However, over time this reference points shifts ranging from tides to climate associated sea-level changes.

Task 2: Working with Data (2 points)

For all tasks below show your properly formatted code. You find the necessary code for the examples in Lander and the online help. Only if asked show also the output.

Import the SPSS data-file **Concord1.sav** in the **WEEK03** channel as **data-frame** into the  environment by using a function from the library **foreign**. Make sure to name your data-frame properly.

```
library(foreign)
```

```
concord <- read.spss('Concord1.sav', to.data.frame = TRUE)
```

- a. Discuss the summary statistics for the water consumption: How did the *average* water consumption change from 1979 to 1981? (0.1 points)

```
Summary(concord[, 2:4])
```

```
water81 water80 water79
```

```
2296.214 2704.900 2974.165
```

Average water consumption reduces every year from 2,974 to 2,298 in 1979 to 1981.

- b. Discuss the summary statistics: Which variable has *missing* observations? (0.1 points)

```
summary(Concord1)
```

The summary statistics indicate that **water79** variable has 47 missing observations.

- c. Discuss the summary statistics: Which variable is a factor? (0.1 points)

```
class(concord$retire)
```

```
[1] "factor"
```

Variable retire is a factor variable.

- d. List all *case numbers* (variable **case**), which have at least for one variable missing value in a variable. Show also the code. (0.2 points)

```
concord[is.na(concord$water79), ]$case
```

```
23 40 46 108 142 143 144 145 146 153 159 178 181 197 199 205
```

```
213 283 290 310 334 359 375 385 408 421 466 480 481 487 488 490
491 497 498 499 500 502 506 507 508 511 512 513 514 515 516
```

- e. Which **class** are the following data selections: [a] `Concord$retire`, [b] `Concord["retire"]`, [c] `Concord[, "retire"]`, and [d] `Concord[["retire"]]`? Show code and the output. (0.2 points)
- Use the **class** function to determine the type of each statement:
- ```
[a]: factor,
[b]: data.frame,
[c]: factor,
[d]: factor
```
- f. Calculate the **average** water consumption over the 3 years for each household and save it the new variable **meanWater** into the data-frame. Caution: also include households, which have **NAs** in the water consumption. Hint: look at the documentation of the function `mean( )`. (0.2 points)
- ```
meanWater <-
rowMeans(concord[c("water79", "water80", "water81")], na.rm=T)
```
- g. Use logical statements to identify those households (variable **case**), which have above average water consumption in 1981. Show the code and the household numbers (0.2 point)
- ```
above_mean <- concord$water81 > meanWater
concord$case[above_mean]
[1] 9 10 11 19 29 33 37 38 39 41 49 50 53 64 70
71 74 76 78 80 81 86 87 88 91
[26] 92 95 97 98 102 104 109 117 127 131 133 134 139 144 160
163 188 198 205 209 219 226 232 233 254
[51] 257 283 296 297 303 309 310 312 318 319 323 324 325 328 334
335 338 345 346 347 348 349 350 352 359
[76] 361 364 370 375 377 378 381 388 392 401 403 405 408 417 421
434 443 445 454 457 459 465 476 483 500
[101] 515
```
- h. Draw a sample of 10 households without repetitions. Show the household numbers (variable **case**) and the code. Hint: look at the documentation of the function `sample( )`. (0.2 points)
- ```
sample(concord$case, 10, replace = FALSE)
[1] 433 482 189 18 205 445 319 127 312 240
```
- i. Add a new variable **seqID** by labeling each record by its record number ranging from 1 to the number of observations. Show the code. (0.2 points)
- ```
concord$seqID <- 1:nrow(concord)
```
- j. **Bind** the two variables **peop80** and **peop81** together into a **matrix**. Show the code. (0.1 points)
- ```
ma <- as.matrix(concord[c("peop80", "peop81")])
```
- k. Give a code example of the use of the **ifelse** statement. (0.1 points)
- ```
ifelse(concord$water79 <= concord$water80, "increase", "decrease")
```
- l. Give an example of the **while** statement (0.1 points)

```

i<-1
while(i<10){
 print(Concord[i,])
 i <- i+1
}

```

- m. What are [a] positional, [b] named and [c] default arguments of a function? (0.3 points)

See the user function `pow( )`, which powers a base by an exponent and optional with an inverse exponent:

```

pow <- function(base, expo, inv=FALSE){
 if (inv==FALSE) result <- base^expo else
 result <- base^(1/expo)
 return(result)
} # end::pow

```

[a] A positional argument of a function matches arguments by their positions, it is the most

common and simplest one.

Example: `pow(8, 2)` : "8 raised to the power 2 is 64"

`pow(2, 8)` "2 raised to the power 8 is 256"

[b] A named argument of a function matches arguments by their names.

Example: `pow(base=8, expo=2)` : "8 raised to the power 2 is 64"

`pow(expo=2, base=8)` "8 raised to the power 2 is 64"

[c] Previous statements assumed the default argument `inv=FALSE` of the `pow( )` function. Overwriting the default argument with `inv=TRUE` calculates the inverse power.

Example: `pow(4, 2)` : squares the base 4

`pow(4, 2, inv=TRUE)` : takes the square-root of 4

### Task 3: Critical discussion of big data analyses (1 point)

Read the document **BIGDATAANDSTATISTICS.PDF** and **MATHMUSICSTATSLITERATURE.PDF** in the **WEEK03** channel.

- [a] Give the reasons why large samples may not necessary be better than small samples. (0.5 points)

Take the election as the example here. The reason of the election example is that samples did not reflect the true population so sampling error and sampling bias become the problems. when it comes to getting a representative sample, sample source is more important than sample size.

- [b] What makes mathematics different from statistics. List the main differences. (0.5 points)

Mathematics is an abstract science that can exist without a direct connection to real world events. Thus, real world experience by the mathematician is not needed. In contrast, statistics makes statement of an underlying data generating process and thus require substantive knowledge of this real world process from the researcher.

Math always follows a consistent definition-theorem-proof structure. In statistics, it is common to define things with intuition and examples, so “you know it when you see it”.

Take for example the concept of an “outlier” what exactly constitutes an outlier? Well, that depends on many criteria, like how many data points you have, how far it is from the rest of the points, and what kind of model you’re fitting.