TakeHomeMidterm

Vikas Sanil

Due 3/11 11:59 pm

» Do not distribute outside of ITMD/ITMS/STAT 514, SPRING 2022, Illinois Tech. «

Instructions & Rules

Use this Markdown document as your working copy of the exam, and edit it. Please use output option `pdf_document` or `html_document`.

- *Questions and Points*

This test has **FOUR** questions. Attempt them all. The maximum number of points is **45**.

- *Some standard writing considerations:*
  - Replace comments that instruct you to put code with your own code.
  - Ensure your plots and output are visible and readable.
  - Ensure you've typed up an explanation of your answers wherever required.
- *Format*: delete comments and replace with your answers and code. Do not just place code, execute it, and expect the reader to be able to interpret the answer for themselves. Type a sentence saying what you just computed and what the reader should understand.
- *Name*: do not forget to put your name on the exam under the 'author' heading.
- *Submission*: your submission must consist of your copy of this Markdown document *and* a knitted `pdf` file (or save a knitted `html` as a `pdf`). Any other type of submission will receive no credit and no opportunity for a re-submission. Late submissions are not accepted.

Honor code

> I will not give or receive information to or from any other persons during this midterm. This document was edited and PDF knitted by me alone.

[Vikas Sanil]

Getting started

Load the packages you will need for your code to run. Probably you need at least these two, but add others if needed.

(These were used on previous homework assignments, so you should not have to run the command `install.packages("....")`, but do run that first if `library` does not load.)

```
# library("ggplot2")
library("tidyverse") # includes tibbles, ggplot2, dyplr, and more.
library("scales")
```

In addition, I'd like to ask `R` to print decimal numbers with 2 digits:

```
options(scipen=2)
```

Obtaining and Understanding the Data

For this exam, we will be using the cybersecurity breach report data downloaded 2015-02-26 from the US Health and Human Services.

To understand what the data represents, here is some information from the *Office for Civil Rights* of the *U.S. Department of Health and Human Services*:

- "As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals.
- "Since October 2009 organizations in the U.S. that store data on human health are required to report any incident that compromises the confidentiality of 500 or more patients / human subjects (45 C.F.R. 164.408). These reports are publicly available. Our data set was downloaded from the Office for Civil Rights of the U.S. Department of Health and Human Services, 2015-02-26."

Load this data set and save it as `cyberData`, using the following code:

```
cyberData<-read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBreac
```

Data Exploration

Question 1. (5 points)

Check the structure of the data using the `str` command. What type of object is `cyberData`? How many observations are recorded? How many variables are recorded? List all of the types of random variables that are recorded based on the output (i.e. int/float etc.).

```
str(cyberData)
```

```
## 'data.frame':    1151 obs. of  10 variables:
##  $ X                                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name.of.Covered.Entity           : chr  "Brooke Army Medical Center" "Mid America Kidney Stone Ass
##  $ State                            : chr  "TX" "MO" "AK" "DC" ...
##  $ Covered.Entity.Type              : chr  "Healthcare Provider" "Healthcare Provider" "Healthcare Pr
##  $ Individuals.Affected           : int  1000 1000 501 3800 5257 857 6145 952 5166 5900 ...
##  $ Breach.Submission.Date         : chr  "2009-10-21" "2009-10-28" "2009-10-30" "2009-11-17" ...
##  $ Type.of.Breach                 : chr  "Theft" "Theft" "Theft" "Loss" ...
##  $ Location.of.Breached.Information: chr  "Paper/Films" "Network Server" "Other, Other Portable Elec
##  $ Business.Associate.Present     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Web.Description                : chr  "A binder containing the protected health information (PHI
```

Answer: *The data set is stored as a ‘data.frame’ with 1151 rows and 10 columns. There are 1151 observations of 10 variables. All of the types of random variables are integer, character and logical.*

Question 2. (20 points)

Let us compare the number of affected individuals across some states.

- Extract the subset of the data for Kansas and Arkansas; in other words, the subset of the data for which `State` column equals `"KS"` or`"AR"`. Add a third state to the dataframe, say, Illinois (i.e., where `State == "IL"`). Name the new **dataframe** `threeStates`.

```
cyberData_StateKS<-data.frame(cyberData[cyberData$State=="KS",])
cyberData_StateAR<-data.frame(cyberData[cyberData$State=="AR",])
cyberData_StateKSnAR<-rbind(cyberData_StateKS, cyberData_StateAR)
str(cyberData_StateKSnAR)
```

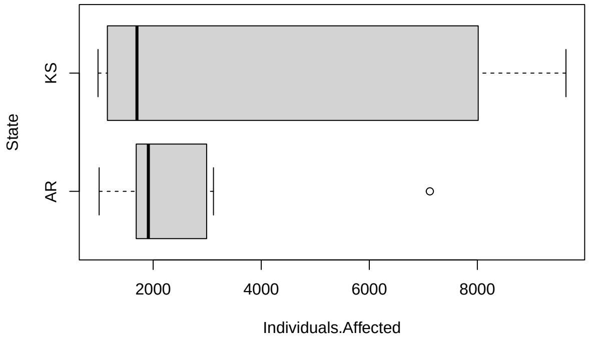
```
## 'data.frame':    14 obs. of  10 variables:
##  $ X                              : int  90 159 389 414 865 935 980 178 341 372 ...
##  $ Name.of.Covered.Entity         : chr  "Occupational Health Partners" "Matthew H. Conrad, M.D., P
##  $ State                          : chr  "KS" "KS" "KS" "KS" ...
##  $ Covered.Entity.Type            : chr  "Healthcare Provider" "Healthcare Provider" "Business Asso
##  $ Individuals.Affected           : int  1105 1200 8275 7757 1700 979 9640 1000 3116 1472 ...
##  $ Breach.Submission.Date         : chr  "2010-06-01" "2010-09-19" "2011-11-14" "2012-01-19" ...
##  $ Type.of.Breach                 : chr  "Theft" "Theft" "Theft" "Theft" ...
##  $ Location.of.Breached.Information: chr  "Laptop" "Laptop, Paper/Films" "Other" "Laptop" ...
##  $ Business.Associate.Present     : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ Web.Description                : chr  "\\N" "\\N" "The covered entity\032\032\032\032\032\032\03
```

```
cyberData_StateIL<-data.frame(cyberData[cyberData$State=="IL",])
threeStates<-rbind(cyberData_StateKSnAR,cyberData_StateIL)
str(threeStates)
```

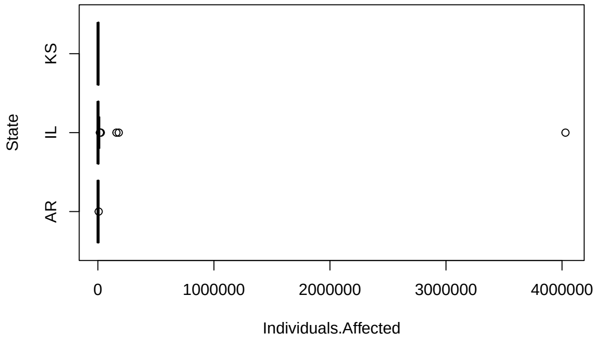
```
## 'data.frame':    71 obs. of  10 variables:
##  $ X                              : int  90 159 389 414 865 935 980 178 341 372 ...
##  $ Name.of.Covered.Entity         : chr  "Occupational Health Partners" "Matthew H. Conrad, M.D., P
##  $ State                          : chr  "KS" "KS" "KS" "KS" ...
##  $ Covered.Entity.Type            : chr  "Healthcare Provider" "Healthcare Provider" "Business Asso
##  $ Individuals.Affected           : int  1105 1200 8275 7757 1700 979 9640 1000 3116 1472 ...
##  $ Breach.Submission.Date         : chr  "2010-06-01" "2010-09-19" "2011-11-14" "2012-01-19" ...
##  $ Type.of.Breach                 : chr  "Theft" "Theft" "Theft" "Theft" ...
##  $ Location.of.Breached.Information: chr  "Laptop" "Laptop, Paper/Films" "Other" "Laptop" ...
##  $ Business.Associate.Present     : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
##  $ Web.Description                : chr  "\\N" "\\N" "The covered entity\032\032\032\032\032\032\03
```

- Create a boxplot of `Individuals.Affected` split across the three states. What conclusion can you draw from it?

```
boxplot(Individuals.Affected ~ State, data=cyberData_StateKSnAR, horizontal = TRUE )
```



```
boxplot(Individuals.Affected ~ State, data=threeStates, horizontal = TRUE )
```



Answer: *From above plots we can see State IL has an outlier observation of Individuals.Affected of value 4029530, which is way too larger than maximum value reported in State KS(9640) and AR(7121).*

The above plot should leave you wondering if Illinois is special, in that it contains some really large data breaches. Let's investigate:

- How many observations in `cyberData` represent a cyber security breach that affected 100,000 individuals or more?

```
str(cyberData[cyberData$Individuals.Affected >=100000,])
```

```
## 'data.frame':    40 obs. of  10 variables:
##  $ X                            : int  62 75 91 93 112 115 180 182 185 216 ...
##  $ Name.of.Covered.Entity       : chr  "Affinity Health Plan, Inc." "Millennium Medical Managemen
##  $ State                           : chr  "NY" "IL" "FL" "PA" ...
##  $ Covered.Entity.Type             : chr  "Health Plan" "Business Associate" "Health Plan" "Business
##  $ Individuals.Affected            : int  344579 180111 1220000 130495 105470 800000 1023209 475000 
##  $ Breach.Submission.Date          : chr  "2010-04-14" "2010-04-29" "2010-06-03" "2010-06-04" ...
##  $ Type.of.Breach                  : chr  "Theft" "Theft" "Theft" "Theft" ...
##  $ Location.of.Breached.Information: chr  "Other" "Other, Other Portable Electronic Device" "Laptop"
##  $ Business.Associate.Present      : logi  FALSE TRUE FALSE TRUE TRUE TRUE ...
##  $ Web.Description                 : chr  "Under a settlement with the U.S. Department of Health and
```

> **Answer:** *40 observations in `cyberData` represent a cyber security breach that affected 100,000 individuals or more.*

- How many of those are in Illinois?

```
str(cyberData_StateIL[cyberData_StateIL$Individuals.Affected >=100000,])
```

```
## 'data.frame':    3 obs. of  10 variables:
##  $ X                               : int  75 746 1126
##  $ Name.of.Covered.Entity          : chr  "Millennium Medical Management Resources, Inc." "Advocate 
##  $ State                           : chr  "IL" "IL" "IL"
##  $ Covered.Entity.Type             : chr  "Business Associate" "Healthcare Provider" "Healthcare Pro
##  $ Individuals.Affected            : int  180111 4029530 160000
##  $ Breach.Submission.Date          : chr  "2010-04-29" "2013-08-23" "2014-12-15"
##  $ Type.of.Breach                  : chr  "Theft" "Theft" "Other"
##  $ Location.of.Breached.Information: chr  "Other, Other Portable Electronic Device" "Desktop Compute
##  $ Business.Associate.Present      : logi  TRUE FALSE FALSE
##  $ Web.Description                 : chr  "\\N" "\\N" "\\N"
```

> **Answer:**: *3 observations are found in State IL where cyber security breach affected 100,000 individuals or more.*

Small analyses across time

Let us now compare attacks before and after 2013. The goal is to see if there is a significant difference in mean number of affected individuals.

Question 3. (10 points)

Check the type of the `Breach.Submission.Date` column: is it a numeric? What type is it?

```
typeof(cyberData$Breach.Submission.Date)
```

```
## [1] "character"
```

```
is.numeric(cyberData$Breach.Submission.Date)
```

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

> **Answer:** *Breach.Submission.Date column is not numeric. The type of the `Breach.Submission.Date` column is character.*

Let us change it to a numeric and extract *the year only*. The code that does this is `as.numeric(format(as.Date(.....),"%Y"`
Let us use this code to break up the data to before and after 2013, like this:

```
before2013 <- subset(cyberData, as.numeric(format(as.Date(Breach.Submission.Date),"%Y")) <=2013 )
after2013  <- subset(cyberData, as.numeric(format(as.Date(Breach.Submission.Date),"%Y")) > 2013 )
```

How many observations are in each subset of the population?

> **Answer:** *Number of observations of breach before 2013 is 848. And number of observations of breach after 2013 is 303.*

Specific type of security breaches

Question 4. (10 points)

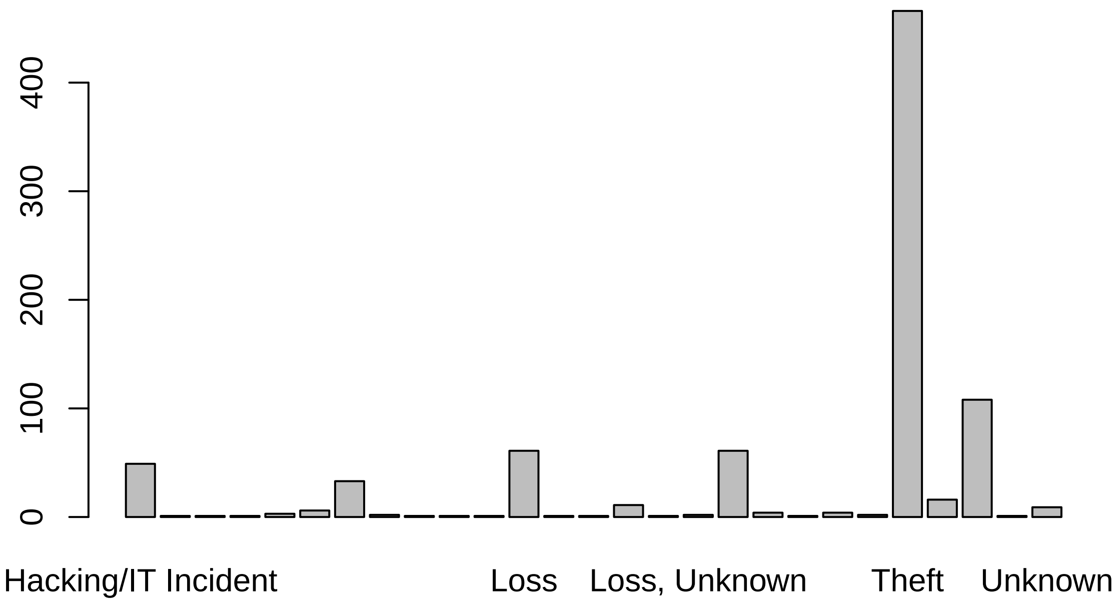
- What proportion of data entries in `cyberData` have `Type.of.Breach == "Hacking/IT Incident"` ?

```
HIT_incident1<-nrow(cyberData[cyberData$Type.of.Breach== "Hacking/IT Incident",])
Overall_incident1<-nrow(cyberData)
proportion1<-percent(HIT_incident1/Overall_incident1, accuracy = 0.01)
```

> **Answer:** *The proportion of data entries in* `cyberData` *which has* `Type.of.Breach == "Hacking/IT Incident"` *is 6.69%.*

- What proportion of data entries in `before2013` have `Type.of.Breach == "Hacking/IT Incident"` ?

```
HIT_incident2<-nrow(before2013[before2013$Type.of.Breach== "Hacking/IT Incident",])
Overall_incident2<-nrow(before2013)
proportion2<-percent(HIT_incident2/Overall_incident2, accuracy = 0.01)
barplot(table(before2013$Type.of.Breach))
```



Answer: *The proportion of data entries before 2013 which has* `Type.of.Breach == "Hacking/IT Incident"` *is 5.78%.*

- What proportion of data entries in `after2013` have `Type.of.Breach == "Hacking/IT Incident"` ?

```
HIT_incident3<-nrow(after2013[after2013$Type.of.Breach== "Hacking/IT Incident",])
Overall_incident3<-nrow(after2013)
proportion3<-percent(HIT_incident3/Overall_incident3, accuracy = 0.01)
```

Answer: *The proportion of data entries after 2013 which has* `Type.of.Breach == "Hacking/IT Incident"` *is 9.24%.*

End of midterm, congratulations!

» Do not distribute outside of ITMD/ITMS/STAT 514, SPRING 2022, Illinois Tech. «
