# Nessie: A Monstrous Tutorial

Green are answers/talking points.

Imagine we have a dataset of reports of the Loch Ness Monster. They come from a variety of sources:books, newspapers, direct interviews etc. In each case different features (estimated length, colour etc.) of the monster and the report are recorded, date, time, what was seen, etc.

***Qs. Can the plural of anecdote be data?***

At this stage ask the students their initial views on the value of the above data set.

***What do you think “anecdote” means? Does it constitute data?***

The Oxford English Dictionary defines anecdote as “A short account of an amusing, interesting, or telling incident or experience; sometimes with implications of superficiality or unreliability.”

***Q. Outline the steps in a creation of the above dataset from witness to analyst?***

Sighting-> Memory -> Transmitted-> Collated-> analysed and at every stage of that process some bias can occur.

***Q. What statistical population does the above data set a sample of?***

Loch Ness monsters?

Sightings of Loch Ness monsters?

Memories of Loch Ness Monsters?

Collated reports of Loch Ness monsters?

How does a sighting of the Loch Ness monster become an analysed report?

Sighting-> Memory -> Transmitted-> Collated-> analysed and at every stage of that process some bias can occur.

So ultimately the statistical population that the reports are an unbiased sample of is collated reports of the Loch Ness Monster.

Hint: what population would the above sample be a biased sample of?

Some of the datums are from multiple reports of the same event by the same witness or from multiple witnesses of the same event.

***Q. Can the reports as described above be treated as independent?***

No, clearly if a single event can produce multiple samples they are not independent. If they were treated as independent this would be an example of pseudoreplication.

However there may be a more subtle violation of independence. Perhaps previous reports generate new reports or how the Loch Ness Monster has been reported in the immediate past influences the way it is reported now etc.

***Q. If there are multiple reports from a single event, how might they be sampled?***

A variety of processes are possible. One could randomly choose one report from each event or choose the report with the most information.

The reports can be first (i.e. direct quotes) or second (“Mr X saw the Loch Ness Monster and described is a 20 ft long” hand.

***Q. What might be the distinction between first and second hand reports of the Loch ness Monster?***

One might expect first hand reports to have on average more detailed information that secondhand reports. Secondhand reports (“friend of a friend tales”) might be more exaggerated compared to the original report with the monster larger or closer etc.

***Q. There are a lot of potential biases in a dataset like this. How representative do you think the recorded Nessie reports discussed here are compared to all reports (unreported and recorded) Nessie reports?***

The recorded reports might be unrepresentative in that only the more dramatic or memorable reports may get remembered or recorded

***Q. Consider variables in the reports such as length, reported nearest distance and duration length,... Do you think there is bias in these variables and if yes, why?***

Like in the in the general case, more extreme values of the variables might be recorded. A long duration, a long length etc.

## Exploratory analysis of the Loch Ness Monster data

Consider some basic features of the Loch Ness Monster reports

**Produce a histogram of the month of sighting.**

Nessie <- read.table("Nessiereducedforeducationpaper.csv", header = T, sep = ",") *# n=1452*

***########### By Month***  
  
Month <- tapply(Nessie$EncounterMonthNumeric, Nessie$Event.ID, mean, na.rm = T)  
par(mfrow = c(1, 1))  
Month2 <- hist(as.numeric(Month), main = "Month", breaks = seq(0, 12), xlab = "Month")

**Produce a histogram of the time of sighting**

***############# Time of Day***  
  
TimeSI <- tapply(Nessie$TimeSI, Nessie$Event.ID, mean, na.rm = T)  
TimeSI <- TimeSI[is.na(TimeSI) == F]  
hist(TimeSI, breaks = seq(0, 24), xlab = "Time of Day", main = "Time of day", col = "white")

***Q. Describe the patterns you see in the plots***

Nessie seem to be more active in the summer and less active at lunchtime…or perhaps there are other explanations!

***Q. Do you think the patterns seen represent a feature of Loch Ness Monsters or is it a function of the reporting process?***

Perhaps Nessie is more active in the summer months but there are also more people to report Nessie in the summer.

Likewise Nessie might be less active at human lunchtimes or something else is going on.

***Q. Does the suitability (i.e. bias) of a sample, vary dependent upon the statistical population under study?***

Yes!

***Q. One final mystery. What do you think is the median duration of a Loch Ness Monster report?***

***A few seconds? A few minutes?***

***############# Duration***  
  
median (Nessie$Duration, na.rm = T)

The median duration is a staggering 4.5 mins.