Contents

CSI 2300: Introduction to Data Science

Lecture 06: Exploratory Data Analysis

Today's Topics

Different types of variables

- continuous
- categorical, factors w/levels, or coded numerically

Numerical summaries of data

- mean
- median
- standard deviation
- proportion

Variable Types

The type of variable that you have helps to determine how you summarize and look at it.

Quantitative/Continuous: These values are recorded on a continuous scale and can theoretically take an infinite number of decimal places. These are values that make sense to average. They can be discrete (taking only a few numeric values) or continuous (taking many numeric values).

Are these quantitative variables? If so, are they continuous or discrete?

- Ambient air temperature
- Time to run one mile
- Sale price of a new Honda Accord EX

- Zip code
- Grades on an exam from 0 to 100
- Number of children a woman has in her lifetime
- Your age

Categorical: The outcomes belong to groups. The outcomes could be ordered, such as class rank (freshman, sophomore, junior, senior), or unordered.

- Make of a car
- Country of birth
- Citzenship
- Academic major
- Decade of your age
- Type of cooling system of a home

Other characteristics of particular variable can also make it unique, such as

- **Time index:** observations can be equally spaced or unequally spaced through time
  - *Equally spaced*–observations recorded every 5 minutes
  - *Unequally spaced*–the height and weight of a baby since birth recorded at each check-up. These tend to be recorded at 0 days, 7 days, 30 days, 60 days, 360 days, but even then, they will not occur on the same day of life for each baby.
- **Spatial location:** the coordinates in 1, 2, or 3-dimensional space could be recorded.
- **Univariate:** just one measurement is taken per individual.
- **Multivariate:** multiple measurements are taken per individual.

Example, Whale Ear Wax: Earplugs from baleen whales form in layers like tree rings do. These long-lived whales can serve as sentinels of the ocean with both environmental and individual chemical signatures stored in their ear wax. Progesterone concentration was measured in earplugs of several species of baleen whales, and a rapid increase in progesterone can indicate that a whale is pregnant. The first few rows of data are shown below.[1] Can you identify the types of variables? Are there any other special features of this data?

```
data <- read.csv(file = "dat/whale_ear_wax_progesterone.csv", header=T)
class(data)
# [1] "data.frame"
dim(data)
# [1] 809  10
head(data)
#   WhaleID Species    Sex    Location Age   Year Progesterone DeltaProgesterone
# 1    1006     Fin Female Netherlands 1.0 1944.0     1.025606                NA
# 2    1006     Fin Female Netherlands 1.5 1944.5     1.182777         15.324726
# 3    1006     Fin Female Netherlands 2.0 1945.0     1.081985         -8.521662
# 4    1006     Fin Female Netherlands 2.5 1945.5     1.219564         12.715444
# 5    1006     Fin Female Netherlands 3.0 1946.0     1.060499        -13.042727
# 6    1006     Fin Female Netherlands 3.5 1946.5     1.226772         15.678701
#   DeltaProgZscore Pregnancy
# 1              NA         0
# 2       1.1634162         1
# 3      -0.7253173         0
# 4       0.9567501         0
# 5      -1.0834045         0
# 6       1.1914525         1
```

---

[1]This dataset is provided courtesy of Stephen Trumble (Biology) and Sascha Usenko (Environmental Science), who are both faculty at Baylor University.

Numerical Summaries

Summarizing quantitative variables can be done in many ways, and different types of summaries or plots will reveal different information about the variables. This lecture focuses on conceptual understanding and not on precise definitions.

Measures of Center:

Measures of center give an idea of where the majority of values fall.

- **Mean:** The average of the observations, denoted by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

  which is just the sum divided by the number of observations

- **Median:** The observation in the center of the values when they are sorted from smallest to largest.
- **Quantiles:** Divides the data (when sorted from smallest to largest) between the lower ηth value and the $(1-\eta)$th, where $\eta \in (0,1)$. For example, if you score in the 30th quantile on an exam, then 30% of people scored worse than you, and 70% scored better than you.
- **Mode:** Usually used for quantitative discrete observations, the mode is the value that occurs the most frequently in a dataset.

Measures of Spread

Measures of spread give an idea of how spread out the values are. For example, say the mean score for an exam is 80 points. If a measure of spread is 2 points, then most people scored within 2 points of 80 (from 78 to 82). If the associated measure of spread is 20 points, then most people scored within 20 points of 80 (from 60 to 100).

- **Range:** the maximum minus the minimum
- **Standard deviation:** the most commonly used measure, it gives the square root of the average of the squared deviations of each observation from the mean, calculated as

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- **Variance:** This is the standard deviation squared, or s^2.
- **IQR:** stands for the interquartile range. This is the difference between the 75th quantile and the 25th quantile. It is the range of the middle 50% of the data.

Categorical Variables

- **Proportion:** The primary tool for summarizing categorical variables is the number of times a particular category occurs divided by the number of observations. Quantitative variables can also be turned into categorical variables, so this is a useful tool for both types of variables.

Example, Whale Ear Wax: Use the dataset loaded earlier to answer the following questions:

- How many different species of whales are there?
- How many measurements of each type of whale are there?
- In what locations do these whales live?
- What is the mean progesterone of each type of whale?
- What is the standard deviation of progesterone of each type of whale?
- What proportion of the progesterone values are above 2?

```
summary(data)
unique(data$Species)
table(data$Species)
sort(table(data$Species))

unique(data$Location)

mean(data$Progesterone[data$Species=="Fin"])
mean(data$Progesterone[data$Species=="Humpback"])
mean(data$Progesterone[data$Species=="Blue"])
mean(data$Progesterone[data$Species=="Minke"])

sd(data$Progesterone[data$Species=="Fin"])
sd(data$Progesterone[data$Species=="Humpback"])
sd(data$Progesterone[data$Species=="Blue"])
sd(data$Progesterone[data$Species=="Minke"])

length(which(data$Progesterone>2))/length(data$Progesterone)

#Here are commands illustrating other commands in the lecture
fin_prog <- data$Progesterone[data$Species=="Fin"]
length(fin_prog)

mean(fin_prog)
median(fin_prog)
```

```r
mode(fin_prog) #What's the problem here?

IQR(fin_prog)
var(fin_prog)
quantile(fin_prog, seq(0.1, 0.9, by = 0.1))
quantile(fin_prog, c(0.25, 0.50, 0.75))
range(fin_prog)
max(fin_prog)-min(fin_prog)
```
