STA130 W1

Monday, October 28, 2019 8:51 PM

Data science / statistics is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.

Data -> Appropriate methods/summarises -> Insights/conclusions

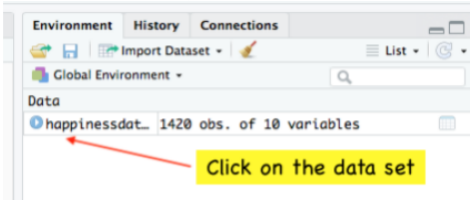
The officers were making the assumption that the planes that came back were a random sample of all the planes (i.e. that they were representative of all planes in combat)
But they are planes that **survived**.
In statistical lingo, the rate of survival and location of bullet holes are correlated.
This underlying statistical phenomena is often called **survivorship bias**.

- think about where the data **came from**
- think about the **assumptions** you are making

R Console: Executes each line of code as you go; does not save code for later use
R Script: Saves code and comments in a file so you can select some or all of the code in a script file to run; does not include output
R Notebook: A file which combines text and chunks of R code (which can be executed independently). This allows you to see output without "knitting" the whole file.

Read data into R
We'll be using the **read_csv** function which is in the **readr** package
`install.packages("readr")` Install the "readr" package
`library(readr)` Open the readr package
`happinessdata_2017 <- read_csv("happinessdata_2017.csv")` Read in the data

View the data
Method 1: Click on the Environment tab in the upper right hand corner (Environment, History, Connections pane). Then click on the data set



Method 2: Add an R code chunk and type **`glimpse(happinessdata_2017)`**

Rows:	**Observation** - For each country in a particular year
Columns:	**Variables** - Measured for each **observation**

*can tell from using **glimpse**

The file happinessdata_2017.csv contains the average happiness score for each country in different years
The *life_ladder* variable is an example of a **numerical (quantitative) variable**.

Numerical (quantitative) Variable

- A quantitative variable takes numerical values that are ordered and differences are meaningful.

The **distribution of a variable** tells us:

- **what** values it takes and how **often** it takes these values.

Histogram: examine the distribution of a **numerical variable**

```
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram(colour = "black", fill = "grey")
```

- Histogram displays distribution of the variable's values
- **Bins** defined by their **lower** bounds (inclusive); upper bound of one bin is the lower bound of the next bin
- Horizontal axis is **numerical** (no gaps)
- Vertical axis gives **the number of values** that lie in each bin

Features of the **distribution of a quantitative variable**
Shape: describes the pattern of values of the variable
Skewness (note: skew is to the direction of the longer tail)

- Symmetric

- Left-skewed
- Right-skewed

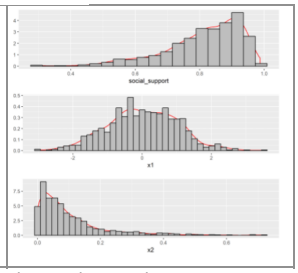
Number of **modes (peaks)**:

- Unimodal
- Bimodal
- Multimodal
- Uniform

Unusual observations

- When there is a data more/less comparing to the trend

Centre: describes a 'typical' value of the variable
Rough center distribution
Spread: describes how concentrated the values of the variable are (variation in the values)
Even if they have the same centre, data may **scattered** too

Left Right	

We'll use the *ggplot2* package in R to construct our graphs(included in the *tidyverse* package)
R studio: A new phone
R packages(e.g. tidy verse): Apps
You need to make sure a package is DOWNLOADED before you can OPEN it (and use it)

Downloading and opening packages
Download the tidyverse package:
Type install.packages("tidyverse") in console (lower left)

Open the tidyverse package:
Type library(tidyverse) in the console

ggplot2
the structure of the code to produce most plots is:

ggplot(data=[*datset*], aes(x=[var1], y=[var2])) + geom_xxx( ) + other options	**datset:** name of the data set **aesthetic:** mapping between a variable and where it will be represented on the graph (e.g., x axis, colour-coding, etc.) **geometry:** what are you plotting (e.g., points , lines, histogram, etc.) Every plot must have **at least one geometry** and there is no upper limit You add a geometry to a plot using +

```
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram(color="black", fill="gray")
```

*Histogram just need **one aesthetic (x)** because it plots the distribution of **only one variable**

Distribution of **a categorical variable: bar plot**
Categorical variable:
Takes a discrete number of values that are often **not ordered** (e.g. country, continent, etc.)
Sometimes these may be coded as numbers in the data (e.g. male = 1, female = 0), but the numerical differences are **not important**.
Bar plot
Displays the distribution of **a categorical variable**, the frequency of its different values
Heights (or lengths) of bars are **proportional to** the percent of individuals
Bars have arbitrary (but equal) widths and spacings

```
ggplot(data = happinessdata_2017,
       aes(x = continent)) +
  geom_bar() +
  theme(text=element_text(size=20))
```

```
ggplot(data = happinessdata_2017,
       aes(x = continent)) +
  geom_bar() + coord_flip() +
  theme(text=element_text(size=20))
```

Geom_bar: create a bar plot
Coord_flip: flip the x- & y-axis if u can't read the label

Relationship between **two numerical variables: Scatterplot** using **geom_point()**

```
ggplot(data = happinessdata_2017) +
 aes(x = logGDP, y = life_ladder) +
 geom_point() +
 theme(text=element_text(size=20))
```

How many aesthetics do we need to create this scatterplot?

- Each point represents a country in a specific year

A scatterplot of life_ladder versus logGDP **consists of points representing a countries** with values of both life_ladder and logGDP.

Features of associations between **quantitative variables**:

Form:
describes the pattern that the two variables follow together (e.g. linear, nonlinear, quadratic, etc.)

Direction:
positive association (values of one variable tend to increase as the other's increases)
negative association (values of one variable tend to **decrease** as the other's **increases**)

Strength:
describes how **concentrated** the values of the variable are around the pattern

*The association btw logGDP and life-ladder is a **positive linear pattern** with **strong**(moderate) **association** btw the 2 variables.

Relationship between 3 variables:

```
ggplot(data = happinessdata_2017, ) +
  aes(x = logGDP, y = life_ladder, colour = continent) +
  geom_point()
```

```
ggplot(data = happinessdata_2017, ) +
  aes(x = logGDP, y = life_ladder) +
  geom_point() +
  facet_wrap(~continent)
```

facet_wrap: separate data into several graphs depends on the variable in ()

Data visualization: Presents data by yourself using various methods, Graphs, diagrams (i.e. histogram, plots)

Vocabulary/ terms:

Bar graphs, histograms:

Where are the data centered (towards the left, right, middle)
How much **spread** (how spread out, how concentrated the data is)
Shape: symmetric, left-skewed, right-skewed
The tails of the distribution (heavy-tailed or thin-tailed)
Modes: where, how many, unimodal, bimodal, multimodal, uniform
Outliers, extreme values
Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

Scatterplots (bivariate or pairwise scatterplots):

Strong / weak relationship
Linear / nonlinear relationship
Direction of association (positive or negative)
Outliers (deviation from what?)
Any visible clusters forming(a group of similar things or people positioned or occurring closely together.)
Each dot represents …