Correlation using R

Alex Mutuku

Plotting the data

Data Visualization and Summary Statistics

1. Define clearly the scientific question to answer,
2. Select a set of representative members from the population of interest and collect data (either through observational studies or randomized experiments)
3. The next major step we usually begin our analysis with **data exploration**.- use graphs and summary statistics to explore the distribution of individual variables.

Objectives of data exploration

1. Develop high-level understanding of the data.
2. Learn about the possible values of each characteristic.
3. Find out how a characteristic varies among individuals/subjects in our sample.

The data exploration methods allow us to reduce the amount of information so that we can focus on the key aspects of the data.

Data exploration technique depends on the type of **variables**.

Data Exploration

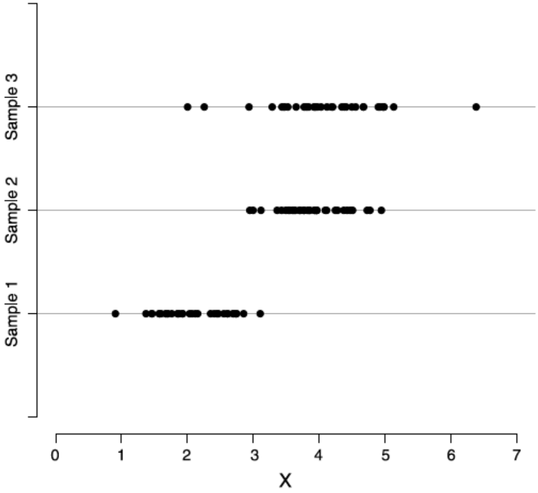
Exploring categorical variables

1. Summary statistics - Frequencies, Proportions and percentages
2. Graphical explorations - Bar Graphs, Pie chart, etc

Data Exploration

Exploring Numerical Variables

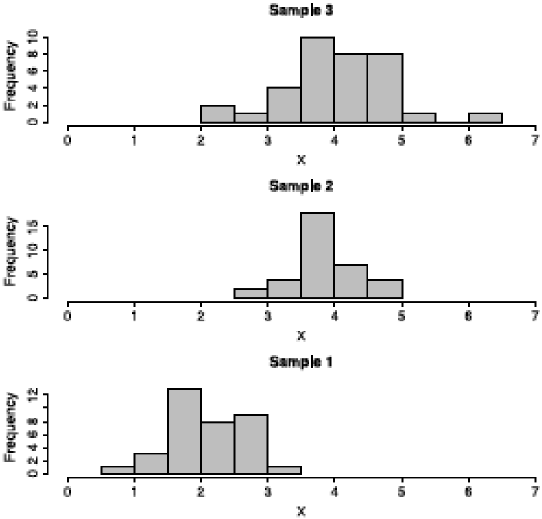
Fig. 2.8 Three separate samples for variable X. Observations in Sample 1 are gathered around 2, whereas observations in Sample 2 and Sample 3 are gathered around 4. Observations in Sample 3 are more dispersed compared to those in Sample 1 and Sample 2



Interested two key aspects of the distribution: **its location and its spread**. In Fig. 2.8, we can see that the observed values in Sample 1 are gathered around X = 2; whereas, the observations in Sample 2 and Sample 3 are gathered around X = 4. Therefore, Sample 2 and Sample 4 have roughly the same location.

On the other hand, Sample 1 and Sample 2 have roughly the same spread, which is smaller than the spread in Sample 3. The individual observations in Sample 3 tend to be further away from the location compared to those in Sample 1 and Sample 2.

Data exploration



Data exploration

1. Means and median
2. Variance and standard deviation
3. Quantiles
4. Boxplots

Exploring Relationships

Session dedicated to using graphs and summary statistics to investigate relationships between two or more variables.

Our objective is to develop a high-level understanding of the type and strength of relationships between variables.

NB: We are not making formal conclusions regarding the existence of relationship or whether the relationship, if exists, is strong or not.

Relationships between Two Numerical Random Variables

Simple way is to visualize the relationship - scatterplot (Graphical)

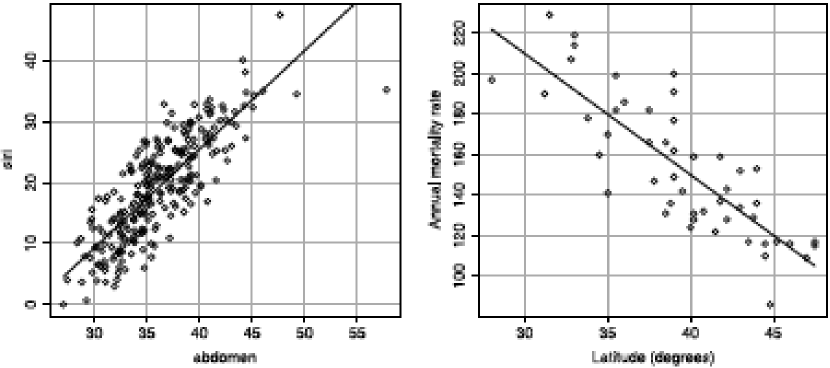


Fig. 3.1 *Left panel*: The scatterplot of percent body fat by abdomen circumference. There is a clear positive linear relationship between the two variables. *Right panel*: The scatterplot of annual mortality rate (per 100,000,000 population) and latitude in degrees. There is a clear negative linear relationship between the two variables

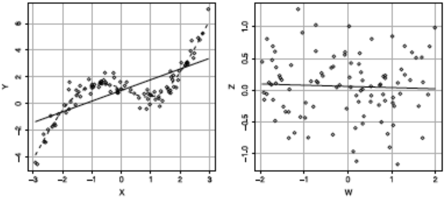


Fig. 3.2 *Left panel*: Scatterplot for two numerical variables with nonlinear relationship. *Right panel*: Scatterplot for two numerical variables that seem to be unrelated

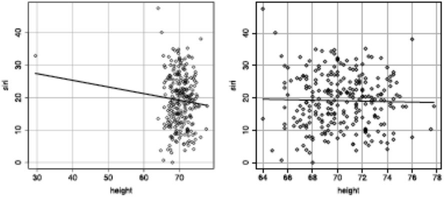


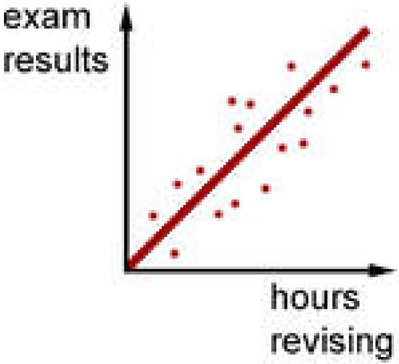
Fig. 3.3 *Left panel*: The scatterplot of percent body fat by height from the `bodyfat` data set. The isolated point at the left of the graph is an outlier, which has a drastic influence on the overall pattern. *Right panel*: The scatterplot of percent body fat by height after removing the outlier. The two variables seem to be unrelated

Correlation- Statistical procedure used to measure and describe the relationship between two variables, ranges between -1 and 1

Positive when the values increase together

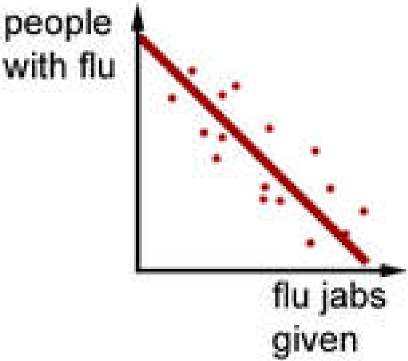
Negative when one value decreases as the other increases

"+1 is a perfect positive correlation, 0 is no correlation (independence) & -1 is a perfect negative correlation"



POSITIVE CORRELATION

- people who do more revision get higher exam results.
- revising increases success.

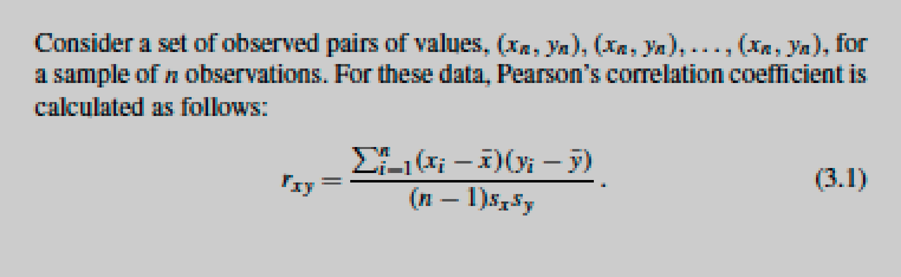


NEGATIVE CORRELATION

- when more jabs are given the number of people with flu falls.
- flu jabs prevent flu.

Correlation Types

1. Pearson product-moment correlation -When both variables, X and Y, are continuous
2. Point bi-serial correlation - When 1 variable is continuous and 1 is dichotomous
3. Phi coefficient - When both variables are dichotomous
4. Spearman rank correlation - When both variables are ordinal (ranked data)



Consider a set of observed pairs of values, $(x_n, y_n), (x_n, y_n), \ldots, (x_n, y_n)$, for a sample of n observations. For these data, Pearson's correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}. \qquad (3.1)$$

Calculation of Correlation

$$r = S_{xy} / \sqrt{S_{xx} S_{yy}}.$$

$$S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 \ \{(\text{variance of x})\}$$

$$S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \ \{(\text{covariance of x and y})\}$$

Deskwork

Input the data below

```
temp <- c(14.2,16.4,11.9,15.2,18.5,22.1,19.4,25.1,23.4,18.1,22.6,17.2)
icecream <- c(215,325, 185, 332, 406, 522, 412,614, 544, 421, 445, 408)
df <- data.frame(temp=temp,icecream=icecream)
table <- list(df) print(df)
```

Correlation computations

df$deviationTemp < -df$temp- mean(df$temp$)
dfdeviationIce <- df$icecream - mean(df$icecream)
df$SSxy < -(df$deviationTemp$dfdeviationIce$)
df SSxx <- (dfdeviationTemp $*$ *df deviationTemp)*
$dfSSyy < -(df deviationIcedfdeviationIce$)
$sum.SSxy < -sum(df$SSxy)
sum.SSxx <- sum(df$SSxx$)
$sum.SSyy < -sum(df$SSyy)

Correlation in R

cor(df*temp*, *df* icecream)
cor.test(df*temp*, *df* icecream)

Diff btwn cor and cor.test

The cor.test output also includes the point estimate reported by cor,
Cor.test has p-value and also CI

4
Thank You