Biostatistics [SBE304] (Fall 2019)
Tutorial 1: Introduction to R and Probabilities

Prof. Ayman M. Eldeib Asem Alaa

Tuesday 1st October, 2019

1 Tutorial facts

The problems in this tutorial comprises:

(A) Programming Works:

- **1** programming in-class demos.
- **1** programming homework.
- **1** self-practicing programming works.

(B) Problem Set:

- **8** problems to be solved in-class.
- **8** problems homework.
- **9** self-practicing problems.

Join this GitHub assignment page to create a repository for your submissions: `https://classroom.github.com/a/CGREt6r6`

2 Programming Language

2.1 Learning Resources

2.1.1 Courses and interactive tutorials

- swirl: "swirl teaches you R programming and data science interactively, at your own pace, and right in the R console"
- R exercises: a nice and light practicing through the basic R.
- DataCamp's free Introduction to R class: 4 hours & 62 exercises on the common data structures used in R (vectors, matrices, factors, data frames, lists).
- DataCamp's free Data Manipulation with **dplyr** in R class: 4 hours & 46 exercises on manipulating data using **dplyr** library.

2.1.2 References

- Quick-R: quick online reference for data input, basic statistics and plots
- R reference card (PDF) by Tom Short (more can be found under Short Documents and Reference Cards)
- Advanced R by Hadley Wickham: "for R users who want to improve their programming skills and understanding of the language"

2.2 R skills needed for this course

The intended R topics to learn is ordered as following:

1. Milestone 1: Setting the R environment & basic data structures (vectors, matrices, factors, data frames, lists);
2. Milestone 2: R simple functions, data manipulation, & visualization;
3. Milestone 3: Simulations & common statistical libraries.

2.3 Programming Problems

1. Problem

Learning R Basic Data Structures Choose one of the excellent introductory courses/tutorials from 2.1.1 to walk through the basic R data structures. Make sure to attempt the exercises as much as you can.

2. Problem

Pollutant Mean The data in the directory **data/pollution_data/specdata** of the demo repository (clone or pull) [1] contains measurements of several pollutant substances in air. Using R script, implement a function that loads multiple **csv** files into a single table and computes the mean value of a given **pollutant**.

```r
# install.packages("data.table")
library("data.table")

pollutantmean <- function(directory, pollutant, id = 1:332) {

  # Format number with fixed width and then append .csv to number

  # Reading in all files and making a large data.table

}

# Example usage
m = pollutantmean(directory = 'data/pollution_data/specdata', pollutant =
    'sulfate', id = 20)
```

The demo file **pollutantMean.R** will be uploaded/updated on the following directory **week1** of the demo repository.

3. Problem

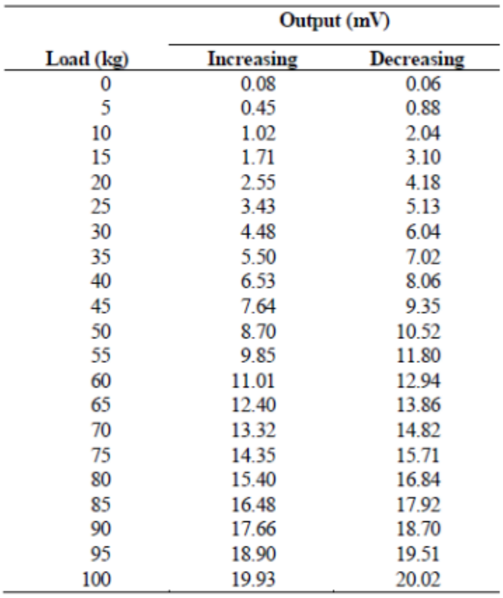
Measurement errors (Measurements SBE206, Fall 2018) A load cell is a sensor used to measure weight. A calibration record table is given below. Determine the maximum error (as a percentage of the full-scale output y_{FSO}) for:

a. accuracy $\varepsilon_a = \frac{y_{\text{measured}} - y_{\text{true}}}{y_{\text{FSO}}} \times 100\%$

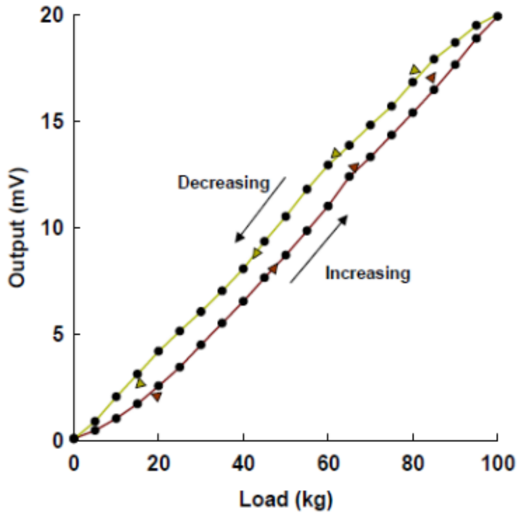
b. hysteresis $\varepsilon_h = \frac{y_{\text{decreasing}} - y_{\text{increasing}}}{y_{\text{FSO}}} \times 100\%$

c. linearity $\varepsilon_l = \frac{y_{\text{measured}} - y_{\text{L}}}{y_{\text{FSO}}} \times 100\%$

The equation of the best-fit line is $y_L(x) = a_0 + a_1 x$, where $a_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$, $a_0 = \frac{1}{n}(\Sigma y - a_1 \Sigma x)$, n is the number of data points. Assume that the true or expected output has a linear relationship with the input. In addition, the expected outputs are 0 mV at 0 kg load and 20 mV at 50 kg load.

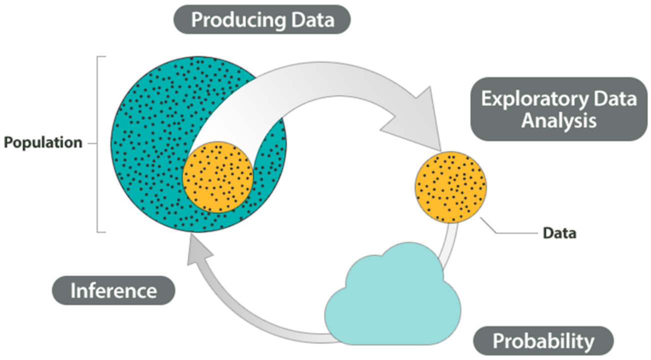


	Output (mV)	
Load (kg)	Increasing	Decreasing
0	0.08	0.06
5	0.45	0.88
10	1.02	2.04
15	1.71	3.10
20	2.55	4.18
25	3.43	5.13
30	4.48	6.04
35	5.50	7.02
40	6.53	8.06
45	7.64	9.35
50	8.70	10.52
55	9.85	11.80
60	11.01	12.94
65	12.40	13.86
70	13.32	14.82
75	14.35	15.71
80	15.40	16.84
85	16.48	17.92
90	17.66	18.70
95	18.90	19.51
100	19.93	20.02



[1] https://github.com/sbme-tutorials/biostatistics-sbe304.git

3 Introduction to Probabilities (and Statistics)



3.1 Pre-class reading

1. Lecture notes of "Probability" by Prof. Ayman M. Eldeib
2. Chapter 2 (pp. 17-41) of Montgomery's textbook

3.2 Chapter overview

3.2.1 Sample Space and Events

Experiment	activity with uncertain outcome
Sample Space($\mathcal{S}$)	the set of all possible outcomes
Event	any collection of outcomes in $\mathcal{S}$

3.2.2 Axioms, Interpretations and Properties of Probability

Given an experiment and a sample space $\mathcal{S}$, the objective probability is to assign to each event A a number $P(A)$, called the probability of event A, which will give a precise measure of the chance that A will occur. Behaves very much like norm.

3.2.3 Axioms & Properties of Probability:

1. $\forall A \in \mathcal{S}, 0 \leq P(A) \leq 1$
2. $P(\mathcal{S}) = 1$
3. If $A_1, A_2, \ldots$ is an infinite collection of disjoint events, $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$
4. $P(\emptyset) = 0$
5. $\forall A, P(A) + P(A') = 1$ from which $P(A) = 1 - P(A')$
6. For any two events $A, B \in \mathcal{S}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
7. For any three events $A, B, C \in \mathcal{S}$, $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Equally Likely Outcomes : $P(A) = \frac{N(A)}{N}$

3.2.4 Counting Techniques

Product Rule for Ordered k-Tuples: If the first element can be selected in n_1 ways, the second in n_2 ways and so on, then there are $n_1 n_2 \cdots n_k$ possible k-tuples.
Permutations: An ordered subset. The number of permutations of size k that can be formed from a set of n elements is $P_{k,n}$
$P_{k,n} = (n)(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$
Combinations: An unordered subset.
$\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$

3.2.5 Conditional Probability

$P(A|B)$ is the conditional probability of A given that the event B has occurred. B is the conditioning event.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication Rule: $P(A \cap B) = P(A|B) \cdot P(B)$

Baye's Theorem Let $A_1, A_2, \ldots, A_k$ be disjoint and exhaustive events (that partition the sample space). Then for any other event B $P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$
$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$

3.2.6 Independence

Two events A and B are **independent** if $P(A|B) = P(A)$ and are **dependent** otherwise.

A and B are **independent** iff $P(A \cap B) = P(A) \cdot P(B)$ and this can be generalized to the case of n mutually independent events.

3.3 Problem Set

1. Problem

A blood test indicates the presence of a particular disease 95% of the time when the disease is actually present. The same test indicates the presence of the disease 0.5% of the time when the disease is not present. One percent of the population actually has the disease. Calculate the probability that a person has the disease given that the test indicates the presence of the disease.

1. Solution

2. Problem

An insurance company issues life insurance policies in three separate categories: standard, preferred, and ultra-preferred. Of the company's policyholders, 50% are standard, 40% are preferred, and 10% are ultra-preferred. Each standard policyholder has probability 0.010 of dying in the next year, each preferred policyholder has probability 0.005 of dying in the next year, and each ultra-preferred policyholder has probability 0.001 of dying in the next year. A policyholder dies in the next year. What is the probability that the deceased policyholder was ultra-preferred?

2. Solution

3. Problem

An auto insurance company has 10,000 policyholders. Each policyholder is classified as:

- young or old,
- male or female, and
- married or single.

Of these policyholders, 3000 are young, 4600 are male, and 7000 are married. The policyholders can also be classified as 1320 young males, 3010 married males, and 1400 young married persons. Finally, 600 of the policyholders are young married males. How many of the company's policyholders are young, female, and single?

3. Solution

4. Problem

a. What is the probability of preferring a glass bottle?

b. What is the probability of preferring a glass bottle given that one lives in Region B?

c. What is the probability of preferring a can given that one lives in Region C?

•	Region A	Region B	Region C
Prefer Can	300	190	60
Prefer Glass Bottle	200	110	40

4. Solution

5. Problem

A clinical engineer receives a lot of 20 syringe pumps. Unknown to him that five of these are defective. He picks 2 pumps at random and tests them. What is the probability that the first is satisfactory and the second is defective?

5. Solution

6. Problem

Two Hemodialysis machines are operating independently. The probability that a specific machine is available when needed is 0.99. Find the probability that

a. A machine is available when needed?

b. Neither is available when needed?

6. Solution

7. Problem

A company manufacturing laptops installs one of three operating systems on each laptop produced. It is known from product testing that during an hour of web browsing the probability a laptop with system 1 installed will crash is 0.15, the probability of a laptop with operating system 2 crashing is 0.08 and the probability of a laptop with system 3 crashing is 0.1. Operating systems 1 and 3 are installed on the same number of laptops while system 2 is installed on twice as many laptops as system 1 (or system 3).

a. What is the probability that a randomly selected laptop crashes during an hour of web browsing?

b. If a laptop selected at random crashed during an hour of browsing the web, what is the probability it has operating system 2 installed?

c. If a laptop selected at random does not crash during an hour of web browsing what is the probability it has operating system 1 or operating system 2 installed?

7. Solution

8. Problem

A company has three plants at which it produces a certain item. 30% are produced at Plant A, 50% at Plant B, and 20% at Plant C. Suppose that 1%, 4% and 3% of the items produced at Plants A, B and C respectively are defective.

a. If an item is selected at random from all those produced, what is the probability that the item was produced at Plant C and is defective?

b. If an item is selected at random from all those produced, what is the probability that the item is not defective?

c. If an item is selected at random, and it is found that the part is not defective what is the probability that the part is from plant A?

8. Solution

9. Problem

A machine produces defective parts with three different probabilities depending on its state of repair. If the machine is in good working order, it produces defective parts with probability 0.02. If is wearing down, it produces defective parts with probability 0.1. If it needs maintenance, it produces defective parts with probability 0.3. The probability that the machine is in good working order is 0.8, the probability that it is wearing down is 0.1, and the probability that it needs maintenance is 0.1. Compute the probability that a randomly selected part is defective.

9. Solution

10. Problem

An Urn contains 2 red balls, and 3 black balls. Two balls are chosen in succession. The first ball is returned to the urn before the second ball is chosen. Each ball is chosen at random which means that each ball is equally likely to be chosen, what is the probability of choosing first a black ball followed by a black ball (in two different ways)?

10. Solution

11. Problem

In a class of k students, what is the probability that at least two students share a common birthday? (Assume all students are born in the same year)

11. Solution

12. Problem

A person answers each of two multiple choice questions at random. If there are four possible choices on each question, what is the conditional probability that both answers are correct given that at least one is correct?

12. Solution

13. Problem

A class in advanced physics is comprised of 10 juniors, 30 seniors, and 10 graduate students. The final grades show that 3 of the juniors, 10 of the seniors, and 5 of the graduate students received an A for the course. If a student is chosen at random from this class and is found to have earned an A, what is the probability that he or she is a senior?

13. Solution

14. Problem

The local area network (LAN) for the College of Business computing system at a large university is temporarily shutdown for repairs. Previous shutdowns have been due to hardware failure, software failure, or power failure. Maintenance engineers have determined that the probabilities of hardware, software, and power problems are .01, .05, and .02, respectively. They have also determined that if the system experiences hardware problems, it shuts down 73% of the time. Similarly, if software problems occur, the system shuts down 12% of the time; and, if power failure occurs, the system shuts down 88% of the time. What is the probability that the current shutdown of the LAN is due to hardware failure? Software failure? Power failure?

14. Solution

15. Problem

A manufacturing operation utilizes two production lines to assemble electronic fuses. Both lines produce fuses at the same rate and generally produce 2.5% defective fuses. However, production line 1 recently suffered mechanical difficulty and produced 6.0% defectives during a 3-week period. This situation was not known until several lots of electronic fuses produced in this period were shipped to customers. If one of two fuses tested by a customer was found to be defective, what is the probability that the lot from which it came was produced on malfunctioning line 1? (Assume all the fuses in the lot were produced on the same line.)

15. Solution

16. Problem

Heart failures are due to either natural occurrences (87%) or outside factors (13%). Outside factors are related to induced substances (73%) or foreign objects (27%). Natural occurrences are caused by arterial blockage (56%), disease (27%), and infection (e.g., staph infection) (17%)

a. Determine the probability that a failure is due to an induced substance.

b. Determine the probability that a failure is due to disease or infection.

16. Solution

17. Problem

A test of a printed circuit board uses a random test pattern with an array of 10 bits and each is equally likely to be 0 or 1. Assume the bits are independent.

a. What is the probability that all bits are 1s?

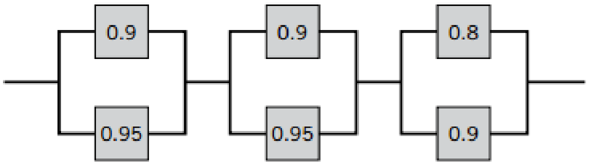
b. What is the probability that all bits are 0s?

c. What is the probability that exactly 5 bits are 1s and 5 bits are 0s?

17. Solution

18. Problem

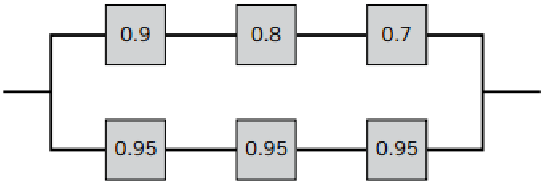
The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?



18. Solution

19. Problem

The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?



19. Solution

20. Problem

An e-mail filter is planned to separate valid e-mails from spam. The word free occurs in 60% of the spam messages and only 4% of the valid messages. Also, 20% of the messages are spam. Determine the following probabilities:

a. The message contains free.

b. The message is spam given that it contains free.

c. The message is valid given that it does not contain free.

20. Solution

21. Problem

Two Web colors are used for a site advertisement. If a site visitor arrives from an affiliate, the probabilities of the blue or green colors being used in the advertisement are 0.8 and 0.2, respectively. If the site visitor arrives from a search site, the probabilities of blue and green colors in the advertisement are 0.4 and 0.6, respectively. The proportions of visitors from affiliates and search sites are 0.3 and 0.7, respectively. What is the probability that a visitor is from a search site given that the blue ad was viewed?

21. Solution

22. Problem

An article in Genome Research ["An Assessment of Gene Prediction Accuracy in Large DNA Sequences" (2000, Vol. 10, pp. 1631–1642)] considered the accuracy of commercial software to predict nucleotides in gene sequences. The following table shows the number of sequences for which the programs produced predictions and the number of nucleotides correctly predicted (computed globally from the total number of prediction successes and failures on all sequences).

•	Number of Sequences	Correct prediction
GenScan	177	0.93
Blastx default	175	0.91
Blastx topcomboN	174	0.97
Blastx 2 stages	175	0.90
GeneWise	177	0.98
Procrustes	177	0.93

Assume the prediction successes and failures are independent among the programs.

a. What is the probability that all programs predict a nucleotide correctly?

b. What is the probability that all programs predict a nucleotide incorrectly?

c. What is the probability that at least one Blastx program predicts a nucleotide correctly?

22. Solution

🏠

23. Problem

A batch contains 36 bacteria cells. Assume that 12 of the cells are not capable of cellular replication. Of the cells, 6 are selected at random, without replacement, to be checked for replication.

a. What is the probability that all 6 of the selected cells are able to replicate?

b. What is the probability that at least 1 of the selected cells is not capable of replication?

23. Solution

🏠

24. Problem

A computer system uses passwords that are exactly seven characters, and each character is one of the 26 letters (a–z) or 10 integers (0–9). Uppercase letters are not used.

a. How many passwords are possible?

b. If a password consists of exactly 6 letters and 1 number, how many passwords are possible?

c. If a password consists of 5 letters followed by 2 numbers, how many passwords are possible?

24. Solution

25. Problem

A person has received the result of his medical test and realized that his diagnosis was positive (affected by the disease). However, the lab report stated that this kind of test has false positive probability of 0.06 (i.e., diagnosing a healthy person, H, as affected, D) and that the probability of false negative is 0.038 (i.e., diagnosing an affected person as healthy). Therefore, while this news was devastating, there is a chance that he was misdiagnosed. After some research, he found out that the probability of this disease in the population is P(D) = 0.02. Find the probability that he is actually affected by the disease given the positive lab result.

25. Solution