Introduction, R Programming, Random Sampling

STAT 135: Concepts of Statistics

GSI: Andy Shen

UC Berkeley, Fall 2022

Section 1

Highlights from the Syllabus

Logistics

- GSI: Andy Shen
- Email: aashen@berkeley.edu
- You may email me with any administrative/private questions or concerns (please include **Stat 135 102/105** in the subject).
- Course announcements will be sent to your email via bCourses (Canvas). Please check your emails regularly for such announcements.
- Any questions about course content (lectures, homework, exams) should be posted **on Piazza** (join code XYZ)

My Lab Sections

- **Lab 102:** 11:00am - 1:00pm in Evans Hall 342
- **Lab 105:** 3:00 - 5:00pm in Evans Hall 332

To get credit for in-class quizzes, you must attend the section you are registered for.

Office Hours

- Mondays: 1:00 - 2:00 PM
- Thursdays: 10:30 - 11:30 AM
- All office hours will be in **Evans Hall 345**

Homework

- HW1 is posted on bCourses and is due **next Friday, Sept 2**
- Homework will be assigned weekly for a total of 10 assignments
- Late assignments will be accepted with a 50% penalty **until solutions are posted.** HW submitted after that will receive no credit.
- Please submit homework to Gradescope. **You must tag your questions correctly.** Improperly tagged questions will not receive credit since the grader cannot see it!
- Is everyone registered on Gradescope?

Quizzes

There will be four 50 minute quizzes during lab sessions to test your understanding of most recent lectures and homeworks. The dates of the quizzes are:

Quiz 1: September 9

Quiz 2: September 23

Quiz 3: November 4

Quiz 4: November 18

We will drop the lowest quiz score from the final grade. For this reason, we will not accommodate make-up quizzes unless under unusual and unexpected circumstances.

Section 2

Review of R Programming (courtesy Miles Chen and Mike Tsiang)

Overview

- Base R commands
  - Good for fast and simple tasks
  - Easy to learn, no need to install packages
- `tidyverse` commands
  - Good for more complex tasks
  - You can generally do more with tidyverse commands
  - Syntax is slightly different from Base R (but not much)
  - `dplyr`: data manipulation and wrangling
  - `ggplot2`: making graphics (looks much nicer than Base R)
  - `tidyr`: reshaping data frames
  - ...and much more!

This class

- You can most likely stick to Base R for most of your commands
- For plotting, I suggest using `ggplot2` because the plots are generally easier to see and interpret but it is up to you. Just make sure your plots are readable.

Section 3

Base R

Object assignment

Use <- to assign an object to a value (or an existing object).

```r
x <- 2
y <- x + 3
```

```r
x
```

```
## [1] 2
```

```r
y
```

```
## [1] 5
```

Vectors

- Vectors are the most important family in R. They form the basis for matrices, data frames, and linear algebra computations
- Create a vector using `c()`
- A string of consecutive integers can be created using the colon operator: `:`

```r
x <- c(1, 2, 3)
x
```

```
## [1] 1 2 3
```

```r
1:9
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

Vector Recycling

When applying arithmetic operations to two vectors of different lengths, R will automatically recycle, or repeat, the shorter vector until it is long enough to match the longer vector.

Question: What is the output of the following commands?

```
c(1, 3, 5) + c(5, 7, 0, 2, 9, 11)
c(1, 3, 5) + c(5, 7, 0, 2, 9)
```

```r
c(1, 3, 5) + c(5, 7, 0, 2, 9, 11)
```

```
## [1]  6 10  5  3 12 16
```

```r
c(1, 3, 5) + c(5, 7, 0, 2, 9)
```

```
## [1]  6 10  5  3 12
```

Matrices

- A **matrix** in R is a vector, but with a dimension attribute of length 2 (rows and columns)
- You can create a matrix using `matrix()`, `rbind()` or `cbind()`

```
M1 <- matrix(1:9, nrow = 3, ncol = 3)
M2 <- cbind(c(1, 2, 3), c(4, 5, 6), c(7, 8, 9))
M3 <- rbind(c(1, 4, 7), c(2, 5, 8), c(3, 6, 9))
```

```
all(M1 == M2)
```

```
## [1] TRUE
```

```
all(M1 == M3)
```

```
## [1] TRUE
```

You can also create a matrix row-wise using `byrow = TRUE` in the `matrix()` command:

```
matrix(1:9, nrow = 3, ncol = 3, byrow = TRUE)
```

Section 4

Review of Probability

Random Variables

A **random variable** is set of possible values that come from some random phenomenon. A **discrete random variable** takes on a countable number of distinct values (such as the number of heads found in n trials). A **continuous random variable** can take on an infinite number of values in an interval (such as the amount of water in a lake).

The **probability mass function (PMF)** $P(X = x)$ is the probability that a *discrete random variable* takes on a specific value. We refer to the support of a probability mass function as the set of values that the discrete random variable takes on.

Some properties of the PMF of a random variable include:

- $\mathbb{P}\left(X = x\right) \geq 0$
- $\sum_x P(X = x) = 1$ (discrete)

The **probability density function (PDF)** $f(x)$ is the probability that a continuous random variable takes on a specific range.

Similarly, some properties of the PDF of a random variable include:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

The **cumulative distribution function (CDF)** of a random variable, X is defined to be:

$$F_X(x) = P(X \leq x);\, x \in (-\infty, \infty)$$

Some properties of the **CDF** of a random variable include:

- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- CDF is non-decreasing (if $x \leq y$, then $F(x) \leq F(y)$) and right-continuous.

Expected Value

The **expectation of a discrete random variable** X with probability mass function $(P(X = x))$ is:

$$\mathbb{E}[X] = \sum_x x \ \mathbb{P}(X = x)$$

The **expectation of a continuous random variable** X with probability density function $(f(x))$ is:

$$\mathbb{E}[X] = \int_x x f(x) dx$$

Variance

The **variance of a random variable** X is the expectation of the squared deviation of the random variable from its mean.

$$\mathsf{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Exercise

Show that the expectation and variance of a Gamma(α, β) distribution is $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively.

The pdf of the Gamma(α, β) distribution is:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad \text{for } x > 0, \quad \alpha, \beta > 0.$$

Hint: $\Gamma(\alpha) = \frac{\Gamma(\alpha+1)}{\alpha}$.

Solution

Section 5

Estimation of Parameters

Parameter Estimation

- Imagine that we have a population and we are interested in some characteristics of that population
  - Example: the sleep time of all UC Berkeley undergraduates majoring in statistics
- In a perfect world, we may want to perform a census and calculate the characteristic (or parameter) of interest.
- However, it is unfeasible and sometimes impossible to find this parameter of interest. As a result, we obtain a subset of this population (a sample), and under certain sampling conditions, we can obtain estimates these characteristics.

Parameters, Statistics, Estimators

- A **parameter** is some constant (usually unknown) that is a characteristic of the population.
- A **statistic** is a random variable that is a function of the observed data.
  - It is important to note that statistics *are not* a function of the parameter of interest.
- An **estimator** is a statistic related to some quantity of the population characteristic.
- In order to fit a probability law to data, we have to estimate parameters associated with the probability law from the data.
- For instance, the normal/Gaussian distribution involves two parameters, μ (mean) and σ (standard deviation), so if we believed that our data followed a normal distribution and either or both are unknown, we would need to provide some estimator for μ and σ.