Introduction, Random Sampling, Review of R programming

STAT 135: Concepts of Statistics

GSI: Andy Shen

UC Berkeley, Fall 2022

About Me

- Andy Shen
- Second year PhD student in Statistics
- Email: aashen@berkeley.edu
- You may email me with any administrative/private questions or concerns (please include **Stat 135 102/105** in the subject).
- 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.

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.

Office Hours (Tenative)

- Mondays: 4:00 - 5:00pm
- Fridays 1:00 - 2:00pm (immediately after 11am lab session)
- All office hours will be in **Evans Hall 345**
- Course announcements will be sent to your email via bCourses (Canvas). Please check your emails regularly for such announcements.

Section 1

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 2

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.