STA 711 Homework 1

Due: Friday, January 26, 11:00am on Canvas.

Instructions: Submit your work as a single PDF. Your document should be created using LaTeX; see the course website for a homework template file and instructions on getting started with LaTeX and Overleaf.

1 Probability review questions

The purpose of these questions is to review some key concepts which will be useful in STA 711. Review the material in Casella & Berger, chapters 1, 2, 3.1–3.3, and 4.1–4.5. Then choose **4** of the following questions to submit (I suggest you practice topics you feel less comfortable with).

1. Suppose that $X \sim Poisson(\lambda)$.

   (a) Calculate $\mathbb{E}[X]$ and $Var(X)$ (without using the mgf).

   (b) Find the mgf of X, $M_X(t)$.

   (c) Use the mgf $M_X(t)$ to derive $\mathbb{E}[X]$ and $Var(X)$, and confirm that your answers match part (a).

2. Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent.

   (a) Show that $X + Y \sim Poisson(\lambda + \mu)$.

   (b) Show that $X|(X + Y = n) \sim Binomial\left(n, \dfrac{\lambda}{\lambda + \mu}\right)$.

3. Let X and Y be continuous random variables with joint pdf

$$f(x, y) = \begin{cases} c(x + 2y) & 0 < y < 1, 0 < x < 2 \\ 0 & \text{else} \end{cases}$$

   where $c > 0$ is a constant.

   (a) Find c.

   (b) Find the marginal pdf of X.

   (c) Find the pdf of $Z = \dfrac{9}{(X + 1)^2}$.

4. The Gamma distribution for random variable Y has probability density function

$$f(y) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} e^{-\frac{y}{\theta}}$$

   where $k > 0$ is the shape parameter, $\theta > 0$ is the scale parameter, and

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$$

   is the Gamma function evaluated at k.

(a) Without using the mgf, derive $\mathbb{E}(Y) = k\theta$. *(Hint: integration by parts)*

(b) Without using the mgf, derive $Var(Y) = k\theta^2$. *(Hint: integration by parts)*

(c) Find the mgf of Y.

(d) Suppose $Y_1, ..., Y_n$ are independent, identically distributed Gamma($k = 1/2, \theta = 2$) random variables. What distribution does $\sum Y_i$ follow? Prove the result using moment generating functions. What is the expected value and variance of this distribution?

5. Consider the following *hierarchical* model, in which the distribution of Y depends on a parameter λ which itself is a random variable:

$$Y|\lambda \sim Poisson(\lambda) \qquad \lambda \sim Gamma\left(\frac{1}{\psi}, \mu\psi\right).$$

(a) Use the law of iterated expectation to show that $\mathbb{E}[Y] = \mu$.

(b) Use the law of total variance to find $Var(Y)$.

(c) Show that the marginal distribution of Y is

$$Y \sim NB\left(\frac{1}{\psi}, \frac{\mu}{\mu + 1/\psi}\right).$$

(Recall that a negative binomial random variable $Y \sim NB(r, p)$ takes values $y = 0, 1, 2, 3, ...$ with probabilities

$$P(Y = y) = \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)}(1 - p)^r p^y,$$

where $r > 0$ and $p \in [0, 1]$.)

6. Let $U \sim Uniform(0, 1)$, and let F_X be a monotonic, continuous cdf.

(a) Show that the random variable

$$X = F_X^{-1}(U)$$

has cdf F_X.

(b) In R, generate 1000 $Uniform(0, 1)$ samples. Then use the inverse cdf method (part (a)) to convert these 1000 uniform samples into 1000 samples from an *Exponential*(1) distribution. (The inverse cdf method is one way of generating random samples from a distribution!)

7. Let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be random variables, and let $a_1, ..., a_m$ and $b_1, ..., b_n$ be constants.

(a) Show that $Cov\left(\sum\limits_{i=1}^{m} a_i X_i, \sum\limits_{j=1}^{n} b_j Y_j\right) = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} a_i b_j Cov(X_i, Y_j)$.

(b) Use (a) to show that

$$Var\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} Var(X_i) + 2 \sum_{1 \le i < j \le m} Cov(X_i, X_j).$$

8. Let $X_1, ..., X_n \overset{iid}{\sim} Uniform(0, 1)$, and let $Y = \max\{X_1, ..., X_n\}$.

(a) Find the cdf and pdf of Y.

(b) Find $\mathbb{E}[Y]$.

(c) More generally, the *order statistics* of the sample $X_1, ..., X_n$ are denoted $X_{(1)}, ..., X_{(n)}$, where $X_{(k)}$ denotes the kth smallest value in the sample. So, e.g., $X_{(1)} = \min\{X_1, ..., X_n\}$ and $X_{(n)} = \max\{X_1, ..., X_n\}$. Find the pdf of the kth order statistic $X_{(k)}$, when $X_1, ..., X_n \overset{iid}{\sim} Uniform(0, 1)$.

2 Logistic regression practice

In this section, you will practice fitting and interpreting a logistic regression model.

Data

The RMS Titanic was a huge, luxury passenger liner designed and built in the early 20th century. Despite the fact that the ship was believed to be unsinkable, during her maiden voyage on April 15, 1912, the Titanic collided with an iceberg and sank. Of all the passengers and crew, less than half survived. Part of the reason why so few people survived has been attributed to the fact that the Titanic did not carry enough lifeboats for its passengers and crew. This meant that there was competition for space in the boats, and not everyone was able to make it aboard. Communication errors, stress and shock...there were a great many factors that contributed to this tragedy.

The loss of life during the Titanic tragedy was enormous, but there were survivors. Was it random chance that these particular people survived? Or were there some specific characteristics of these people that led to their positions in the life boats? Let's investigate.

We have observations on 12 different variables, some categorical and some numeric:

- `Passenger`: A unique ID number for each passenger.
- `Survived`: An indicator for whether the passenger survived (1) or perished (0) during the disaster.
- `Pclass`: Indicator for the class of the ticket held by this passengers; 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
- `Name`: The name of the passenger.
- `Sex`: Binary indicator for the sex of the passenger.
- `Age`: Age of the passenger in years; Age is fractional if the passenger was less than 1 year old.
- `SibSp`: number of siblings/spouses the passenger had aboard the Titanic. Here, siblings are defined as brother, sister, stepbrother, and stepsister. Spouses are defined as husband and wife.
- `Parch`: number of parents/children the passenger had aboard the Titanic. Here, parent is defined as mother/father and child is defined as daughter,son, stepdaughter or stepson. NOTE: Some children traveled only with a nanny, therefore parch=0 for them. There were no parents aboard for these children.
- `Ticket`: The unique ticket number for each passenger.
- `Fare`: How much the ticket cost in US dollars.

- `Cabin`: The cabin number assigned to each passenger. Some cabins hold more than one passenger.
- `Embarked`: Port where the passenger boarded the ship; C = Cherbourg, Q = Queenstown, S = Southampton

Goal: Our goal is to predict the probability that a passenger survives the Titanic disaster.

Loading the data

The `titanic` data can be loaded into R with the following command:

```
titanic <- read.csv("https://sta711-s24.github.io/homework/Titanic.csv")
```

Questions

9. We want to fit a logistic regression model to predict the probability that a patient survives the disaster. For this question, use passenger class, sex, and age as the explanatory variables.

   (a) Write down the population logistic regression model, making sure to include both the random and systematic components. Use proper notation and include all subscripts. Note: passenger class is a categorical variable with more than two levels!

   (b) In R, fit your logistic regression model, and write down the equation of the fitted model.

   (c) Interpret each estimated coefficient in terms of the log odds.

   (d) Interpret each estimated coefficient in terms of the odds.

10. Finally, let's use the logistic regression model to make some predictions!

   (a) Suppose we have a 20 year old male passenger, but we don't know their passenger class. Which passenger class (first, second, or third) would have the highest chance of survival? You should answer the question without doing any calculations, and explain your reasoning.

   (b) Suppose we have a 30 year old passenger, but we don't know their passenger class or sex. Which combination of class and sex would have the highest chance of survival? You should answer the question without doing any calculations, and explain your reasoning.

   (c) *In the observed data*, what is the highest predicted probability of survival for a female first-class passenger?

3 Simulations

A simulation study allows us investigate statistical questions by simulating data under a variety of different conditions, and seeing how the statistical methods behave under these different conditions. We will use simulations pretty regularly in STA 711.

The paper "Using simulation studies to evaluate statistical methods" (Morris et al. 2019) provides a good overview of the important steps in designing a simulation study. Read sections 1 (Introduction) and 3 (Planning simulation studies), and then answer the following questions.

11. What are some reasons researchers use simulation studies?

12. According to the paper, what are the five components (abbreviated ADEMP) involved in planning a simulation study? Summarize each of the five components.