PS630 Homework 2

Haohan Chen

September 11, 2018

1 Concept (3 points)

1. y is a random variable generated from a Bernoulli distribution with $p = 0.6$. Calculate its expected value $E(y)$ and variance $Var(y)$. $y \sim Bern(0.6)$ (0.5pt)
2. y is a random variable generated from a Binomial distribution with $p = 0.7, n = 23$. Calculate its expected value $E(y)$ and variance $Var(y)$. $y \sim Binom(p = 0.7, n = 23)$ (0.5pt)
3. y is a random variable generated from a Poisson distribution with $\lambda = 10$. $y \sim Pois(10)$
   a. Calculate its expected value $E(y)$ and variance $Var(y)$ (0.5pt).
   b. Discuss: What special property does a random variable generated from a Poisson distribution have? (0.5pt)
4. y is a random variable generated from a Normal distribution with $\mu = 10, \sigma^2 = 5$. $y \sim N(10, 5)$
   a. When $y = 5$, what is the probability density? When $y = 15$, what is the probability density? (0.5pt)
   b. Calculate $Pr(y > 15)$. (1) write the mathematical formula and (2) use R for the calculation. Hint: `rnorm, pnorm, qnorm`. One of these three functions does the job. (0.5pt)

2 Programming (7 points)

In the programming task of this homework assignment is a mini data analysis project. You will run t-test on a toy datasetusing what we have learned from Lecture 3 and Lab 1 & 2. You will (1) load the dataset, (2) perform some exploratory analysis, (3) run statistical analyses and (4) interpret the results.

The dataset, named `data.csv`, contains two columns: The column `y` is the outcome variable of interest. The column `treatment` indicates whether the data point is in the treatment group (`treatment = 1`) or the control group (`treatment = 0`).

2.1 Load the dataset (0pt, just a preparation step)

Load the attached dataset. The dataset is in `.csv` format. Use the function `read.csv`. If you wonder how it works, type `?read.csv` in the console and learn from information in the `Help` panel.

2.2 Describe the dataset (2pt)

The first step to take after obtaining any dataset is an exploratory data analysis (aka EDA). For this homework, you will re-do the following tasks similar to Homework 1

1. Get summary statistics of `y` and `treatment`. Discuss if `y` appears different in the two groups. (1pt) [Hint: Produce a table similar to the slide of Lecture 3, page 40.]
2. Plot the density and the histogram of `y` of the whole dataset. What do you think is its distribution? (0.5pt)
3. Plot the density and the histogram of `y` by group. What do you think is the distribution of `y` in each group? (0.5pt)

Hint: Check the code of Lab 1. If you feel like trying `ggplot` for this task, here's a good tutorial .

2.3 One-sample T-test (2pt)

What is an one-sample t-test? (0.5pt) Perform one-sample T-tests to answer the following questions.

1. Is the `y` of the whole sample significantly different from 5.4 at 95% confidence level? (0.5pt)
2. Is the `y` of the whole sample significantly greater to 5.4 at 95% confidence level? (0.5pt)
3. Is the `y` of the whole sample significantly less than 5.4 at 95% confidence level? (0.5pt)

Run analysis in R and describe the results in text with reference to relevant stastics. Clearly state what your null hypotheses are.

Hint: Use the `t.test` function. Learn how to perform one-sided t-test from this webpage.

2.4 Two-sample T-test (2pt)

What is a two-sample t-test? (0.5pt) Perform two-sample T-tests to answer the following questions.

1. Is the `y` of the treatment group significantly different from that of the control group at 95% confidence level? (0.5pt)
2. Is the `y` of the treatment group significantly greater to that of the control group at 95% confidence level? (0.5pt)
3. Is the `y` of the treatment group significantly less than that of the control group at 95% confidence level? (0.5pt)

Run analysis in R and describe the results in text with reference to relevant statistics. Clearly state what your null hypotheses are.

Hint: Use the `t.test` function. Run a two-sample t-test. Perform 3 tests in total as you do in the above question. This may be a useful reference.

2.5 Reflection on the results (1pt)

1. What do you find in the results of one-sided and two-sided t-tests you ran above? When should you run one-sided and when should you run two-sided tests? (0.5pt)
2. What are Type I and Type II errors? Use one of the t-tests you run above as example to briefly explain the concept. (0.5pt)

3 Bonus for Nice Figures (1pt)

There is not a bonus question for this homework. Bonus points will be given if you draw clearly labeled figures.