DSCI 552 - Practice Quiz 1

Contents

Quiz

Solution Key

Quiz

Instructions

- Answer all open-ended questions in your own words when necessary.
- All multiple choice questions **ONLY** require your final answer.

Question 1

In four to six sentences, explain why and how parameter estimation in a population of interest is useful and give an example of a problem it would let you solve.

Question 2

In two or three sentences, compare and contrast the sample distribution versus the sampling distribution of your estimate. Discuss how they are related (or not).

Question 3

Fill in the blanks `(...)` in the code below to correctly calculate the percentile-type 90% confidence interval from `10000` bootstrap samples for the mean `mpg` in the `mtcars` data set.

> **Note:** `mtcars` is a base data set in `R`.

```
...(123)

mtcars %>%
  ...(size = ...(mtcars), ... = ..., ... = 10000) %>%
  ...(stat = ...(mpg)) %>%
  ...(... = 0.90, ... = ...)
```

Question 4

In two or three sentences, define confidence intervals and explain why they are essential for estimation.

Question 5

In a hypothesis test that uses permutation between two treatments to generate the distribution expected under the null hypothesis, how is permutation different from bootstrapping?

Answer in one or two sentences.

Question 6

Why should a bootstrap sample be the same size as your original sample? What would happen if the bootstrap sample size were larger/smaller?

Answer in two or three sentences.

Question 7

You are trying to perform a hypothesis test to determine whether providing chicks different feed (i.e., `horsebean` or `casein`) affects their weight. You have set up the following hypotheses and plan to carry out a difference in means test using permutation (with `10000` replicates) to answer this question.

The hypotheses are the following:

H_0: the population mean weight of chicks fed with `horsebean` is equal to that of chicks fed with `casein`.

H_a: the population mean weight of chicks fed with `horsebean` is not equal to that of chicks fed with `casein`.

Fill in the blanks (`...`) in the code below so that it will correctly generate the distribution you would expect under the null hypothesis, H_0. The data set `chickwts` has two columns:

1. `weight`: the weight of the chick.
2. `feed`: the type of feed (`horsebean` or `casein`).

   Note: `chickwts` is a base data set in `R`.

```
...(552)

horsebean_casein <- chickwts %>%
  filter(feed == "horsebean" | feed == "casein")

horsebean_casein_null_distribution <- horsebean_casein %>%
  specify(...) %>%
  hypothesize(...) %>%
  generate(...) %>%
  calculate(...)
```

Question 8

To answer this question, take Question 7 as the context.

Fill in the blanks (`...`) in the code below so that it will correctly get the p-value (`soybean_casein_p_value`) for the hypothesis test using `horsebean_casein_null_distribution` and the **observed** test statistic `delta_star`.

```
delta_star <- horsebean_casein %>%
  ...(feed) %>%
  ...(stat = ...(...)) %>%
  pull(stat) %>%
  diff()

soybean_casein_p_value <- ... %>%
  ...(delta_star, ... = ...)
```

Question 9

To answer this question, take Questions 7 and 8 as the context.

If you specify the significance level to be 0.05, what would you conclude about the null and alternative hypotheses given a p-value < 0.001?

Answer in one or two sentences.

Solution Key

Question 1

Parameter estimation is useful for obtaining crucial information from a population of interest, such as the mean or variance. Nonetheless, assume we do not have census data to work with. Hence, we collect a sample that provides its corresponding estimates for these parameters. For example, suppose we want to know the real proportion of smartphone users who use Android over iOS in North Vancouver for an app we are developing in that area. However, we do not have access to all the carrier information there. Hence, we will need to collect a proper sample of smartphone users in North Van to estimate this population proportion.

Question 2

A sample distribution describes how a single random sample of n elements is distributed over its corresponding ranges, e.g., the histogram distribution. We can compute a given estimate from this sample (e.g., the sample mean). This computation can be done over m different random samples of the same size n, and then plot the corresponding sampling distribution of the m estimates (as in a histogram), which might differ from the single random sample distribution.

Question 3

```r
set.seed(123)

mtcars %>%
  rep_sample_n(size = nrow(mtcars), replace = TRUE, reps = 10000) %>%
  summarize(stat = mean(mpg)) %>%
  get_confidence_interval(level = 0.90, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     18.4     21.8
```

Question 4

A confidence interval is an interval estimate that allows us to account for the sampling error (due to uncertainty) associated with the population parameters' estimation. They are essential since they provide a range of values for which the true population parameter is likely to be contained, given a confidence level.

Question 5

The treatment labels are randomly shuffled in permutation and then applied without replacement. On the other hand, in bootstrapping, we would use the observed data (both treatment labels and outcome of interest) as a base sample to generate bootstrap samples (of the same size n) via sampling with replacement.

Question 6

The bootstrap sample must have the same size n as the base sample, so we will not obtain a misleading spread in our bootstrap sampling distribution. If we draw m bootstrap samples of size smaller than n, then we would overestimate the bootstrap sampling variability of the m samples. Conversely, if the bootstrap sample size is larger than n, we would underestimate the bootstrap sampling variability.

Question 7

```
set.seed(552)

horsebean_casein <- chickwts %>%
  filter(feed == "horsebean" | feed == "casein")

horsebean_casein_null_distribution <- horsebean_casein %>%
  specify(formula = weight ~ feed) %>%
  hypothesize(null = "independence") %>%
  generate(10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("horsebean", "casein"))

head(horsebean_casein_null_distribution)
```

```
## Response: weight (numeric)
## Explanatory: feed (factor)
## Null Hypothesis: independence
## # A tibble: 6 x 2
##   replicate   stat
##       <int>  <dbl>
## 1         1  33.3
## 2         2 -42.6
## 3         3 -20.0
## 4         4  83.6
## 5         5   2.35
## 6         6  13.4
```

Question 8

```
delta_star <- horsebean_casein %>%
  group_by(feed) %>%
  summarise(stat = mean(weight)) %>%
  pull(stat) %>%
  diff()

soybean_casein_p_value <- horsebean_casein_null_distribution %>%
  get_p_value(delta_star, direction = "both")
soybean_casein_p_value
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0002
```

Question 9

Given that p-value < 0.05, we have enough statistical evidence to state that the population mean weight of chicks fed with `horsebean` is not equal to that of chicks fed with `casein` (i.e., we reject H_0 in favour of H_a).