Math 300 Lesson 32 Notes

Case Study

YOUR NAME HERE

July, 2022

Contents

Objectives

1. Conduct and interpret a hypothesis test for the difference of two means.
2. Calculate and interpret a confidence interval for the difference of two means.

Reading

Chapter 9.5

Lesson

Work through the learning checks LC 9.9 - LC 9.15.

- This is a two-sided test. The p-value calculation is a little different.
- Interpret the p-value in context of the problem.

Libraries

```
library(tidyverse)
library(infer)
library(moderndive)
library(ggplot2movies)
```

LC 9.9 (Objective 1)

(LC 9.9) Conduct the same analysis comparing action movies versus romantic movies using the median rating instead of the mean rating. What was different and what was the same?

Solution:

```
# Complete the code
set.seed(2511)
# In calculate() step replace "diff in means" with "diff in medians"
# null_distribution_movies_median <- __________ %>%
#   specify(formula = rating ~ ________) %>%
#   hypothesize(null = "______________") %>%
#   generate(reps = 1000, type = "___________") %>%
#   calculate(stat = "_______________", order = c("Action", "Romance"))
```

```
# Complete the code
# compute observed "diff in medians"
# obs_diff_medians <- ____________ %>%
#   specify(formula = rating ~ ________) %>%
#   calculate(stat = "____________", order = c("Action", "Romance"))
# obs_diff_medians
```

```
# Complete the code
# Visualize p-value. Observing this difference in medians under H0
# is very unlikely! Suggesting H0 is false, similarly to when we used
# "diff in means" as the test statistic.
# visualize(null_distribution_movies_median, bins = 10) +
#   shade_p_value(obs_stat = obs_diff_medians, direction = "__________")
```

```
# Complete the code
# p-value is very small, just like when we used "diff in means"
# as the test statistic.
# null_distribution_movies_median %>%
#   get_p_value(obs_stat = obs_diff_medians, direction = "__________")
```

- Confidence interval

```
# Complete the code
set.seed(2511)
# In calculate() step replace "diff in means" with "diff in medians"
# boot_distribution_movies_median <- movies_sample %>%
#   specify(formula = rating ~ __________) %>%
#   generate(reps = 1000, type = "_________") %>%
#   calculate(stat = "____________", order = c("Action", "Romance"))
```

```
#visualize(boot_distribution_movies_median, bins = 10)
```

```
# Complete the code
#boot_distribution_movies_median %>%
#  get_ci(level=_______,type="______________")
```

LC 9.10 (Objective 1)

(LC 9.10) What conclusions can you make from viewing the faceted histogram looking at `rating` versus `genre` that you couldn't see when looking at the boxplot?

Solution:

```
# Complete the code
# ggplot(data = movies_sample, aes(x = __________)) +
#   geom_histogram(bins=8) +
#   facet_wrap(~_______)+
#   labs(x = "IMDb rating")
```

LC 9.11 (Objective 1)

(LC 9.11) Describe in a paragraph how we used Allen Downey's diagram to conclude if a statistical difference existed between mean movie ratings for action and romance movies.

Solution:

LC 9.12 (Objective 1)

(LC 9.12) Why are we relatively confident that the distributions of the sample ratings will be good approximations of the population distributions of ratings for the two genres?

Solution:

LC 9.13 (Objective 1)

(LC 9.13) Using the definition of *p*-value, write in words what the *p*-value represents for the hypothesis test comparing the mean rating of romance to action movies.

Solution:

LC 9.14 (Objective 1)

(LC 9.14) What is the value of the *p*-value for the hypothesis test comparing the mean rating of romance to action movies?

Solution:

LC 9.15 (Not testable)

(LC 9.15) Test your data wrangling knowledge and EDA skills:

- Use `dplyr` and `tidyr` to create the necessary data frame focused on only action and romance movies (but not both) from the `movies` data frame in the `ggplot2movies` package.
- Make a boxplot and a faceted histogram of this population data comparing ratings of action and romance movies from IMDb.
- Discuss how these plots compare to the similar plots produced for the `movies_sample` data.

Solution:

- Use `dplyr` and `tidyr` to create the necessary data frame focused on only action and romance movies (but not both) from the `movies` data frame in the `ggplot2movies` package.

```
action_romance <- movies %>%
  select(title,year,rating,votes,Action,Romance) %>%
# Get rid of movies that are both
  filter(!(Action == 1 & Romance == 1)) %>%
  filter(Action == 1 | Romance == 1) %>%
   mutate(genre = case_when(
    Action == 1 ~ "Action",
    Romance == 1 ~ "Romance",
    TRUE ~ "Neither"
  )) %>%
  select(-Action,-Romance)
```

Summary stats

```
# Complete code
# action_romance %>%
#   group_by(_______) %>%
#   summarize(n = n(), mean_rating = mean(_______), std_dev = sd(_______))
```

- Make a boxplot and a faceted histogram of this population data comparing ratings of action and romance movies from IMDb.

```
# Complete code
# ggplot(data = action_romance, aes(x = __________)) +
#   geom_histogram(bins=8) +
#   facet_wrap(~__________)+
#   labs(x = "IMDb rating")
```

```
# Complete code
# ggplot(data = action_romance, aes(x = ________)) +
#   geom_density() +
#   facet_wrap(~________)+
#   labs(x = "IMDb rating") +
#   theme_classic()
```

```
# Complete code
# ggplot(data = action_romance, aes(x = _________,y=______________)) +
#   geom_boxplot() +
#   labs(y = "IMDb rating") +
#   theme_classic()
```

Documenting software

- File creation date: 2022-07-11
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1

- `moderndive` package version: 0.5.4
- `infer` package version: 1.0.2
- `ggplot2movies` package version: 0.0.1