Math 300 Lesson 33 Notes

More on Hypothesis Tests

YOUR NAME HERE

July, 2022

Contents

Objectives

1. Standardize a variable.
2. Find percentile and probabilities for the t distribution; this includes determining the correct degrees of freedom.
3. Use the `infer` package to conduct a theory-based hypothesis test.
4. Know and verify the assumptions of the two-sample t-test.

Reading

Chapter 9.6

Lesson

There are no learning checks for this section.

- The material in this section is the traditional small sample normal-based inference.

These traditional theory-based methods have been used for decades mostly because researchers didn't have access to computers that could run thousands of calculations quickly and efficiently. Now that computing power is much cheaper and more accessible, simulation-based methods are much more feasible. However, researchers in many fields continue to use theory-based methods. Hence, we make it a point to include an example here.

- Often, when using traditional methods, we have to *standardize* a variable to get a mathematical solution for the sampling distribution of that variable (i.e., difference in means, the t distribution). Standardization can be helpful in other ways as well, such as comparisons between two quantities on different scales (ACT vs SAT scores). Note that we don't have to standardize when using computational methods.
- Spend time on the problem with p-values. Read the papers and point out key ideas from the reading.

Libraries

```
library(tidyverse)
library(infer)
library(moderndive)
library(nycflights13)
library(ggplot2movies)
```

Problem

Let's use the 2013 airline data to test the hypothesis that United Airlines flights arrive at their destination later than Delta flights on average. We will take a sample even though we have the computational power to look at all the flights.

Data

Let's get a sample.

```
# Complete code
set.seed(90)
# flight_sample <- ______ %>%
#   select(arr_delay, carrier,origin) %>%
#   filter(carrier%in%c("________","UA")) %>%
#   group_by(________) %>%
#   slice_sample(n=50) %>%
#   ungroup()
```

EDA

```
# Complete code
# Density plot
# ggplot(data = flight_sample, aes(x = _________)) +
#   geom_density(fill="cyan") +
#   facet_wrap(~__________)+
#   labs(x = "Arrival Delay") +
#   theme_classic()
# Complete code
# Boxplots
# ggplot(data = flight_sample, aes(y = __________, x= _________)) +
#   geom_boxplot(fill="cyan") +
#   labs(x = "Arrival Delay") +
#   theme_classic()
```

```
# Complete code
# Summary stats
# flight_sample %>%
#   group_by(_________) %>%
#   summarize(n = n(), mean_delay = mean(__________), std_dev = sd(__________))
```

Summary of findings:

Hypothesis Test Theory-based

```
# Complete code
# Get the null distribution
# null_dist <- _________ %>%
#   specify(formula = __________ ~ ___________) %>%
#   assume("t")
```

```
# Complete code
# Find observed difference
# obs_flights_mean_diff <- ___________ %>%
#   specify(formula = __________ ~ ___________) %>%
#   hypothesize(null = "_________________") %>%
#   calculate(stat = "t", order = c("DL", "________"))
```

```
# Complete code
# visualize(null_dist) +
#   theme_classic() +
#   shade_p_value(obs_stat = _____________, direction = "both")
```

```
# Complete code
#Get p-value
# null_dist %>%
#   get_p_value(obs_stat = ___________________, direction = "both")
```

Conclusion:

Hypothesis Test Permutation

This method does not assume normality.

```
# Complete code
# Get the null distribution
# null_dist_permute <- _________ %>%
#   specify(formula = ___________ ~ ______________) %>%
#   hypothesize(null = "_______________") %>%
#   generate(reps = 1000, type = "______________") %>%
#   calculate(stat = "diff in means", order = c("DL", "__________"))
```

```
# Complete code
# Find observed difference
# obs_flights_mean_diff_permute <- _________________ %>%
#   specify(formula = ______________ ~ ____________) %>%
#   calculate(stat = "diff in means", order = c("DL", "____________"))
```

```
# Complete code
# visualize(_________________) +
#   theme_classic() +
#   shade_p_value(obs_stat = ____________________, direction = "both")
```

```
# Complete code
#Get p-value
# ________________________ %>%
#   get_p_value(obs_stat = ____________________________, direction = "both")
```

Conclusion:

Confidence Interval

```
# Complete code
# _____________________ %>%
#   get_ci(level=________________,type="__________________")
```

We could use medians instead of means in the permutation test. It will not be as sensitive to the outliers.

Confidence Interval for Medians

```
# Complete code
# Get the null distribution
# null_dist_permute_median <- flight_sample %>%
#   specify(formula = ________________ ~ ________________) %>%
#   hypothesize(null = "__________________") %>%
#   generate(reps = 1000, type = "__________________") %>%
#   calculate(stat = "diff in medians", order = c("DL", "____________"))
```

```
# Complete code
# ______________________ %>%
#   get_ci(level=_____________,type="percentile")
```

Conclusion:

Documenting software

- File creation date: 2022-07-11
- R version 4.1.1 (2021-08-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4
- `infer` package version: 1.0.2
- `nycflights13` package version: 1.0.2
- `ggplot2movies` package version: 0.0.1