Accelerated TA session 8: Style and review

Accelerated Coding Lab

2022-09-08

Review

Doing math with vectors

T-tests are used to determine if two sample means are equal. The formula for a t-score is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where x_i is the mean of the first or second set of data, s_i is the sample standard deviation of the first or second set of data, and n_i is the sample size of the ith set of data.

We'll first create two data sets of random numbers following a normal distribution:

```
set.seed(1)
data_1 <- rnorm(1000, 3)
data_2 <- rnorm(100, 2)
```

1. What built-in functions do you need to calculate the variables in the formula for each `data_`i?
2. Calculate the t-score using the formula above?

```
# SOLUTIONS
n_1 <- length(data_1)
n_2 <- length(data_2)
sd_1 <- sd(data_1)
sd_2 <- sd(data_2)
x_bar_1 <- mean(data_1)
x_bar_2 <- mean(data_2)

(x_bar_1 - x_bar_2)/sqrt(sd_1^2/n_1 + sd_2^2/n_2)
```

What did you get for the t-score? Hint: You should have gotten 9.243, if not, double check your code!

Remark: As a rule of thumb, t-scores close to 0 imply that the means are not statistically distinguishable, and large t-scores (e.g. t $>$ 3) imply the data have different means. You'll learn more in stats 1!

Data manipulation with `[` and `dplyr`

Using `storms` which comes with `dplyr`. Do the following in base R and dplyr.

1. What category storms have a non-zero value for `hurricane_force_diameter`? (Once you subset the data you can use `distinct()` (tidyverse) or `unique()` base R to find the answer.)
2. Find all data from storms named "Ana".
3. What is the `maximum` `category` for storms named "Ana"? Have we ever had a hurricane named "Ana"?
4. Collect the columns that relate to the time and location of the storms.
5. Get the columns that measure the "force diameter" of tropical storms and hurricanes.
6. Create a column that is called `ratio` that is the ratio of pressure to wind.
7. What is the mean and sd of `ratio` for category 5 storms?
8. What is the mean and sd of `ratio` for category 1 storms?
9. What is the first year `tropicalstorm_force_diameter` is not `NA`?

```
# SOLUTIONS
# 1
storms %>%
  filter(hurricane_force_diameter > 0) %>%
  distinct(category)

storms[storms$hurricane_force_diameter > 0, "category"] %>% unique()

# 2 & 3
storms %>%
  filter(name == "Ana") %>%
  summarize(max_category = max(category))


ana <- storms[storms$name == "Ana", ]
max(ana$category)

# 4
select(storms, year:long)

storms[, 2:7]

# 5

select(storms, ends_with("diameter"))

storms[, c("tropicalstorm_force_diameter", "hurricane_force_diameter")]

# 6-8
new_storms <- storms %>%
  mutate(ratio = pressure/wind)

new_storms %>%
  filter(category == 5) %>%
  summarize(mean = mean(ratio),
            sd = sd(ratio))

new_storms %>%
```

```
  filter(category == 1) %>%
  summarize(mean = mean(ratio),
            sd = sd(ratio))


# next week!
storms %>%
  mutate(ratio = pressure/wind) %>%
  filter(category %in% c(1,5)) %>%
  group_by(category) %>%
  summarize(mean = mean(ratio),
            sd = sd(ratio))


storms$ratio <- storms$pressure / storms$wind

cat_five <- storms[storms$category == 5, ]
mean(cat_five$ratio)
sd(cat_five$ratio)
cat_one <- storms[storms$category == 1, ]
mean(cat_one$ratio)
sd(cat_one$ratio)

# 9
storms %>%
  filter(!is.na(tropicalstorm_force_diameter)) %>%
  summarize(min(year))
```

ifelse or case_when

1. Answer the question: What is the first year `tropicalstorm_force_diameter` is not `NA` using sorting and no filtering. You'll notice that when we sort `NA` goes to the end of the line / bottom of the data. This motivates creating an indicator column that is 1 if the data is missing and 0 otherwise.
2. Add a column to the data called `season` that takes names "winter", "spring", "summer" or "fall" depending on the month of the year. (You can pick the cut offs as you see fit.)
3. *challenge* using `case_when` in `mutate` make your season indicator depend on the month and day of the year. (E.g. Winter is roughly December 21st to March 20th.)

```
# SOLUTIONS
# 1
storms %>%
  mutate(is_missing = ifelse(is.na(tropicalstorm_force_diameter), 1, 0)) %>%
  arrange(is_missing, year) %>%
  select(year, is_missing, tropicalstorm_force_diameter)

storms$is_missing <- ifelse(is.na(storms$tropicalstorm_force_diameter), 1, 0)

storms[order(storms$is_missing, storms$year),]


# 2
```

```r
storms %>%
  mutate(season = case_when(month == 3 & day <= 20 ~ "winter"
                            between(month, 3, 5) ~ "spring",
                            month == 6 & day <= 21 ~ "spring",
                            between(month, 6, 8) ~ "summer",
                            month == 9 & day <= 21 ~ "summer",
                            between(month, 9, 11)  ~ "fall",
                            month == 12 & day <= 20 ~ "fall",
                            TRUE ~ "winter"))
```