Week 6 Day 5

Stat140-04

Case studies

Case 1: Olympic Swim Suits

In the 2000 Olympics, was the use of a new wetsuit design responsible for an observed increase in swim velocities? In a study designed to investigate this question, twelve competitive swimmers swam 1500 meters at maximal speed, once wearing a new wetsuit and once wearing a regular swimsuit [1]. The order of wetsuit versus swimsuit was randomized for each of the 12 swimmers.

The following code reads in the data set:

(a) What is each observational unit?

Each observational unit was a competitive swimmer.

(b) What is the population parameter of interest?

The average difference in swimmer's velocity when a competitive swimmer is wearing a wet suit and when they are wearing a regular swim suit. Denoted by μ.

(c) To make inference about the population parameter, you will need to add a new variable to the data frame that is calculated as the difference between `wet_suit_velocity` and `swim_suit_velocity`. You could call this new variables something like `velocity_difference`.

```
swim <- swim %>%
  mutate(
    velocity_difference = wet_suit_velocity - swim_suit_velocity
  )
```

(d) State the null and alternative hypotheses for a hypothesis test of whether the new wetsuit design led to an increase in swim velocities for competitive swimmers.

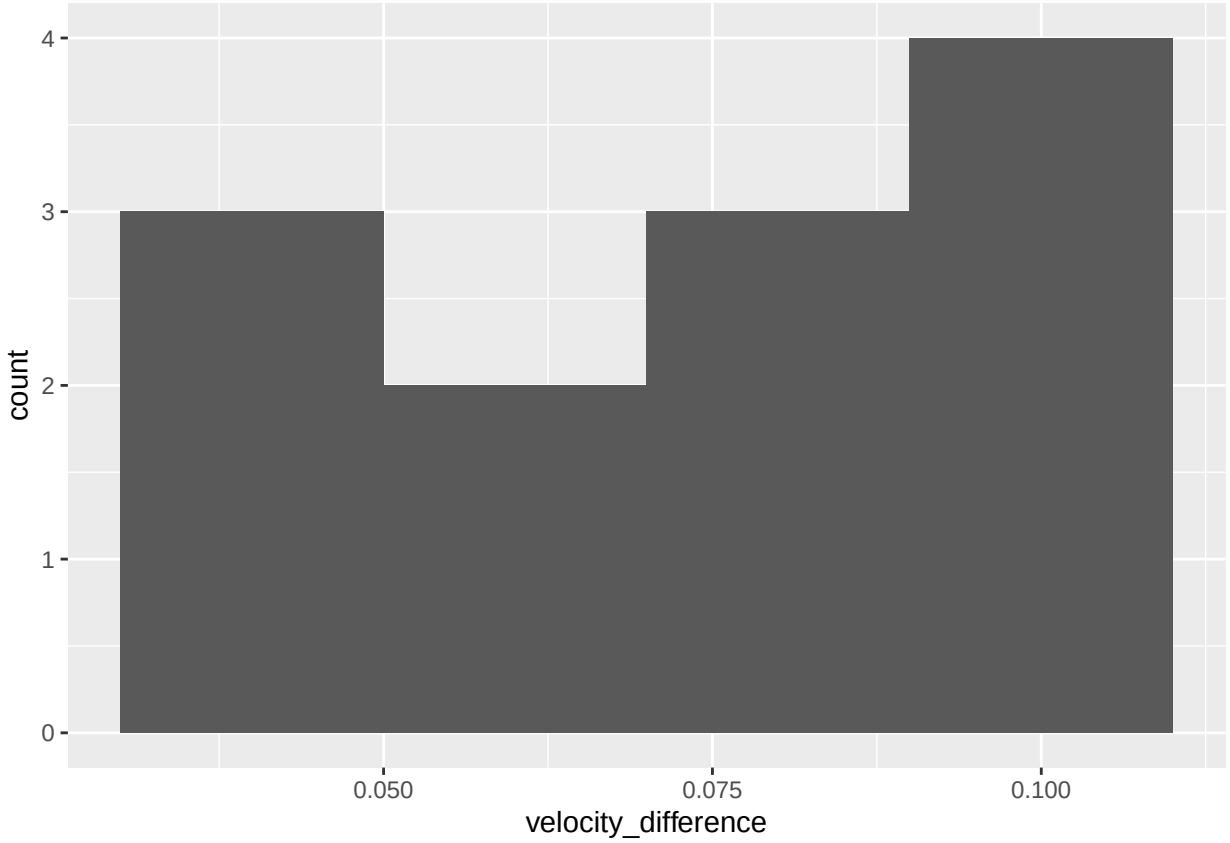
H_0: $\mu = 0$. There is no difference in mean swimsuit velocities between when the swimmer is wearing the new wetsuit design and when the swimmer is wearing the regular swim suit.

H_A: $\mu > 0$. On average, the swimmers are swimming faster when wearing the wet suit than when wearing the swim suit.

(e) Check all of the conditions required for using the Central Limit Theorem on this data set to make inference about the population parameter. For any conditions that you can't check based on the given information, note that and explain why.

[1] De Lucas et. al, The effects of wetsuits on physiological and biomechanical indices during swimming. 2000; 3(1): 1-8

```
ggplot(data = swim, mapping = aes(x = velocity_difference)) + geom_histogram(binwidth = 0.02)
```



Independence: There is no information about how these 12 swimmers were selected, so it is difficult to assess the conditions of bias and independence. We would need to know that the swimmers were representative of all competitive swimmers, and that there was no explicit connection between the different swimmers in our sample. We do know that the order of swimsuits was randomized, so that is not a source of bias. We recorded a quantitative variable for each swimmer: the difference in their velocity wearing the wet suit and their velocity wearing the regular swim suit.

Nearly normal: According to the nearly normal condition, if the sample size is less than 15, the data should follow a Normal model pretty closely. The histogram we plotted above is relatively symmetric, but it is not unimodal or bell-shaped. So the nearly normal condition does not meet.

(f) Regardless of the conditions, let's proceed to the testing for some practice. Run the necessary R code to conduct the hypothesis test you set up in part (e). Do this using a both a built-in R function (t.test) and a manual procedure (pt).

```
t.test(swim$velocity_difference, alternative = "greater", mu = 0)
```

```
##
##  One Sample t-test
##
## data:  swim$velocity_difference
## t = 12.318, df = 11, p-value = 4.443e-08
## alternative hypothesis: true mean is greater than 0
```

```
## 95 percent confidence interval:
##  0.06620114        Inf
## sample estimates:
## mean of x
##    0.0775
```

```
swim %>%
  summarize(
    mean = mean(velocity_difference),
    sd = sd(velocity_difference)
  )
```

```
## # A tibble: 1 x 2
##     mean          sd
##    <dbl>       <dbl>
## 1 0.0775 0.021794495
```

```
0.0775/(0.02179449/sqrt(12))
```

```
## [1] 12.31815
```

```
pt(12.318, df = 11, lower.tail = FALSE)
```

```
## [1] 4.443268e-08
```

(g) What is the conclusion of the hypothesis test at the $\alpha = 0.01$ significance level? Explain what it means in the context of the problem.

The p-value of 4.44326825488507e08 is less than 0.01, so we reject the null hypothesis at the $\alpha = 0.01$ significance level. We conclude that on average, competitive swimmers will swim faster when wearing the new wetsuit design than when wearing a regular swim suit.

(h) What would a Type I error be in this example? Is it possible that a Type I Error was made in part (g)?

The hypothesis test falsely identifies the swimmers are swimming faster when wearing the wet suit than when wearing the swim suit. In reality, there is no difference. Yes, the probability of making a Type I error is 1%.

(i) What would a Type II error be in this example? Is it possible that a Type II Error was made in part (g)?

The hypothesis test falsely identifies there is no difference between the swimmers when wearing the wet suit and wearing the swim suit, when in reality, there is indeed a difference. Yes, it is possitible to make a Type II error especially when the significance level is low ($\alpha = 0.01$), and the sample size is small ($n = 12$).

Case 2: Vehicle inspection

A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Many states require that cars be tested annually to be sure they meet these standards. Suppose state regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop's license if they find significant evidence that the shop is certifying vehicles that do not meet standards. A car shop is criticized because 3 of 40 certified cars are not meeting the standards in the double check. Is this statistical significant?

(a) Put this in the context of a hypothesis test. What is the population parameter, and what are the hypotheses?

Let p denote the proportion of certified cars in the shop that are not meeting the standards

$H_0 : p = 0$. The shop is meeting the standards

$H_a : p > 0$. The shop does not meet the standards

(b) What would a Type I error be?

It is decided that the shop is not meeting standards when it is

(c) What would a Type II error be?

The shop is certified as meeting standards when it is not

(d) Which type of error would the shop's owner consider more serious?

Type I error

(e) Which type of error might environmentalists consider more serious?

Type II error

(f) Check all of the conditions required for using the Central Limit Theorem on this data set to make inference about the population parameter. For any conditions that you can't check based on the given information, note that and explain why.

Independence condition: the data come from a random sample.

Success/Failure condition: note that since $n \times p_0 = 0 < 10$ regardless of the sample size, we cannot use CLT based method to do hypothesis testing.

(g) Try using the built-in R function (prop.test) and see if you can get any p-value.

```
prop.test(x = 3, n = 40, p = 0, alternative = "greater")
```

This will produce the following error: 'Error in prop.test(x = 3, n = 40, p = 0, alternative = "greater") : elements of 'p' must be in (0,1)'

(h) This approximation based on the central limit theorem becomes unreliable when the sample size is small or the success probability is close to 0 or 1 ($p = 0$ or $p = 1$). Alternatively, you could use confidence intervals to conduct hypothesis tests: If a 95% CI doesn't contain the value from H_0, you can reject H_0 at significance level $\alpha = 0.05$. Use the bootstrap method to create a 95% CI and use the confidence interval above to test whether this claim is plausible given the data.

Using statkey the bootstrap CI with 5000 simulation is 0.0175 ± 0.01 which contains H_0. Therefore we fail to reject H_0 and we don't have statisitcal significance to claim the shop does not meet the standards.

Case 3: Second job

Using confidence intervals to conduct hypothesis tests: If a 95% CI doesn't contain the value from H_0, you can reject H_0 at significance level $\alpha = 0.05$

In June 2010, a random poll of 800 working men found that 72 of them had taken on a second job to help pay the bills. Let's assume that this sample was representative of working men. Based on these data, a 95% confidence interval for the proportion of working men who had taken on a second job is [0.071, 0.112].

(a) A pundit on a TV news show claimed that 6% of working men had a second job. Use the confidence interval above to test whether this claim is plausible given the poll data. What is the significance level of your test?

Since the confidence interval [0.071, 0.112] does not contain 6% = 0.06, we reject the null hypothesis and conclude that the claim by the TV news is not plausible. The significance level is $alpha = 0.05$

(b) For what proportion of samples would the confidence interval from the problem statement contain the true proportion of working men who had taken on a second job?

95%