# 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?
(b) What is the population parameter of interest?
(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`.
(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.
(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.
(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).
(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.
(h) What would a Type I error be in this example? Is it possible that a Type I Error was made in part (g)?
(i) What would a Type II error be in this example? Is it possible that a Type II Error was made in part (g)?

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

## 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?
(b) What would a Type I error be?
(c) What would a Type II error be?
(d) Which type of error would the shop's owner consider more serious?
(e) Which type of error might environmentalists consider more serious?
(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.
(g) Try using the built-in R function (prop.test) and see if you can get any $p$-value.
(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.

## 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?
(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?