# 140.615.HW.4.Delahanty.Jeremy

Jeremy Delahanty

2024-03-06

## 1

**Suppose I measure some treatment response on a set of 10 mice from strain A and measure the same response on a set of 5 mice from strain B. Calculate a 95% and 99% confidence interval for the difference in mean treatment responses of strains A and B assuming that the standard deviations within the two groups are the same. What is the p-value for the hypothesis test that the means are the same?**

In this question, we have a situation where allthough we assume the standard deviation within the two groups is the same, we have two different sizes of our sample. The sample sizes are quite small, so we cannot assume normality and instead will use a `T Distribution` that is built into `R`. We will effectively state that each strain is approximately independent and identifcally distributed (IID) with respective means $\mu$ and standard deviation $\sigma$. The null hypothesis is that our two means are equivalent, where $\mu_A = \mu_B$. The alternative hypothesis is that the two means are different, where $\mu_A \neq \mu_B$. We will depend upon the values returned for our Test Statistic $T$ which is calculated as follows:

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

*Sp* in this case refers to the pooled sample variance of our two strains. Remember, the pooled sample variance is given by:

$$S_p = \sqrt{\frac{S_A^2(n-1) \ + \ S_B^2(m-1)}{n+m-2}}$$

We can use the pooled value because we are assuming that the variances of each group are equal and that each sample is IID. We will use 13 degrees of freedom.

First, for the 95% confidence interval:

```
# For calculating confidence intervals between the means, we have to
# calculate the difference in the mean, the pooled standard error,
# and t-intervals for the differencein means.

# Calculate statistics for strain A
strain_a <- c(84, 106, 99, 101, 100, 99, 127, 105, 101, 108)
# Calculate mean and variance of strain B
mean_a <- mean(strain_a)
variance_a <- var(strain_a)
```

```r
# Calculate statistics for strain B
strain_b <- c(56, 62, 67, 81, 69)
# Calculate mean and variance of strain B
mean_b <- mean(strain_b)
variance_b <- var(strain_b)

# Calculate degrees of freedom: Sum of sample sizes minus groups (2)
df <- length(strain_a) + length(strain_b) - 2

# Calculate the pooled standard deviation because we assume the
# underlying standard deviation between the two populations is the same
sd_pooled <- sqrt((variance_a*(length(strain_a) - 1) +
                   variance_b*(length(strain_b) - 1))/df)

# Use the sd_pooled value to calculate the se_pooled value
se_pooled <- sd_pooled*(sqrt(1/length(strain_a) + 1/length(strain_b)))

# We now have enough information for our later calculation of
# the confidence intervals of these differences in means! All we need
# now is the t-statistic. We can calculate this from the qt function.
# For a 95% confidence interval, each tail makes up 2.5% of the distribution
# and we will use our calculated degrees of freedom
t_statistic <- qt(0.975, df)

# Finally, we can determine the confidence interval bounds by putting all
# these values together!
ci_difference_95 <- round(c((mean_a - mean_b) - (t_statistic * se_pooled),
                      (mean_a - mean_b) + (t_statistic * se_pooled)), 4)

# Display the interval
cat("The 95% CI for difference in means is:", ci_difference_95)
```

```
## The 95% CI for difference in means is: 23.8528 48.1472
```

Now, for the 99% CI. We can use the calculated values above for our pooled standard error, but have to calculate new T-statistics and upper/lower bounds of the interval:

```r
# No need to recalculate lots of stuff, just get a new t-statistic first
# This time, our tails must have 0.005 in each tail so both tails together sum
# to 1%. Use same degrees of freedom from above
t_statistic <- qt(0.995, df)

# Calculate the interval bounds using our values
ci_difference_99 <- round(c((mean_a - mean_b) - (t_statistic * se_pooled),
                            (mean_a - mean_b) + (t_statistic * se_pooled)), 4)

# Display the interval
cat("The 95% CI for difference in means is:", ci_difference_99)
```

```
## The 95% CI for difference in means is: 19.0627 52.9373
```

Now, we can finally calculate a p-value four our test of a difference in means. The `pt` function returns

the cumulative density function of a Student T distribution with random variables and specified degrees of freedom. This is all built into R!

```
# First calculate the t-statistic for significance
t_statistic <- (mean_a - mean_b)/se_pooled

# Use pt and multiply the value by 2 because we are interested in both sides
# of the distribution. We will supply the t-statistic, degrees of freedom, and
# specify that the lower.tail=FALSE because we would like to know the
# probability of observing a value above a given t-statistic, or the upper tail of
# it.
p_value <- 2*pt(q=t_statistic, df=df, lower.tail = FALSE)

# Display the value
cat("The p-value for our difference in means is:", p_value)
```

```
## The p-value for our difference in means is: 2.333983e-05
```

# 2

**Suppose we wish to estimate the concentration of a specific dose of ampicillin in urine. We recruit 25 volunteers and find that they have a specific sample mean and standard deviation. Find a 90% confidence interval for the population mean concentration, assuming that the underlying population distribution of concentrations is normal.**

With a sample mean of 7.0 $\mu g/mL$ and standard deviation of 3.0 $\mu g/mL$ and an assumed normal distribution of concentrations in the population, we can use the T-distribution to calculate our confidence interval. We cannot uses something like the normal distribution with such small sample sizes like 25 subjects.

```
# Define number of individuals sampled
n <- 25

# Calculate degrees of freedom for one sample
df <- n - 1

# Define sample mean and standard deviation
s_mean <- 7.0
s_sd <- 3.0

# We need a t-statistic where 0.05% of the distribution is on each tail. Use
# the qt function for determining the t-statistic.
t_statistic <- qt(p = 0.95, df = df)

# We also need the standard error of the samples, which is calculated using the
# standard deviation and the square root of the sample size
se <- s_sd/sqrt(n)

# Finally, we can calculate the confidence intervals!
ci_90 <- round(c(s_mean - (t_statistic*se),
                 s_mean + (t_statistic*se)
```

```
            ), 4)

# Display the value
cat("The 90% CI for ampicillin concentration is:", ci_90)
```

```
## The 90% CI for ampicillin concentration is: 5.9735 8.0265
```

# 3

## Consider data for some measurement on 6 mice before and after treatment. Does the treatment have an effect?

### No difference in means

The question does not define a specific directionality for a treatment effect (i.e. it does not specify if the values should increase or decrease), so we will not assume one in our testing. In other words, we will use both tails of the distribution found in testing. Tests will be performed using the built in `t.test()` function in R.

Our null hypothesis in this case will be that the means are the same following treatment ($\mu_{Before} = \mu_{After}$) while the alternative is that these means are not equal ($\mu_{Before} \neq \mu_{After}$).

```
# First, define the data for before and after
before <- c(81, 101, 76, 67, 125, 144)
after <- c(138, 210, 162, 105, 259, 319)

# Define level of confidence, and therefore value of alpha, for our t-test
alpha <- 0.05

# Perform the t-test using paired = TRUE because each value of before/after
# is paired together. We want to determine if there's a difference after the
# treatment, so the order of testing is important to use after, before
# Don't store as a variable, just print out to the terminal
t.test(after, before, paired = TRUE, conf.level = 1 - alpha)
```

```
##
##  Paired t-test
##
## data:  after and before
## t = 4.8425, df = 5, p-value = 0.004705
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   46.8377 152.8290
## sample estimates:
## mean difference
##        99.83333
```

We see here that our p-value is quite small at 0.004705 and thus $p < \alpha$! We can reject the null hypothesis that there is no difference in means and state that, with 95% confidence, there is an effect of treatment.

### Difference in means: Increase in value following treatment

If we wanted to test whether there is an increase in response, where the null hypothesis is that $\mu_{Before} = \mu_{After}$ and the alternative is that $\mu_{Before} < \mu_{After}$, we can again use the `t.test()` function in `R` but state our alternative as `greater` in the function.

```
# Use same alpha as before, but specify our alternative hypothesis as
# "greater" so the difference in means after is greater than before treatment
t.test(after, before, paired = TRUE, alternative = "greater",
       conf.level = 1 - alpha)
```

```
##
##  Paired t-test
##
## data:  after and before
## t = 4.8425, df = 5, p-value = 0.002352
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  58.29069      Inf
## sample estimates:
## mean difference
##        99.83333
```

We can see that our p value is quite low at 0.002352 and thus can reject our null hypothesis that there is no difference in means and conclude at 95% confidence that the mean after treatment is greater than the mean before treatment!

### Difference in Means: Decrease in value following treatment

To confirm that it is indeed a unidirectional treatment effect, we can alter our alternative hypothesis so that $\mu_{Before} > \mu_{After}$. We can again use the `t.test()` function in `R`.

```
# Use same alpha as before, but specify our alternative hypothesis as
# "less" so the difference in means after is less than before treatment
t.test(after, before, paired = TRUE, alternative = "less",
       conf.level = 1 - alpha)
```

```
##
##  Paired t-test
##
## data:  after and before
## t = 4.8425, df = 5, p-value = 0.9976
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##     -Inf 141.376
## sample estimates:
## mean difference
##        99.83333
```

We see our p value of 0.9976 is much greater than our alpha of 0.05 and thus fail to reject the null hypothesis in this direction!

# 4

**E.canis infection is a tick-borne disease of dogs that is sometimes contracted by humans. Among infected humans, the distribution of white blood cell counts has an unknown mean and an unknown standard deviation. In the general population, the mean white blood cell count is 7250/mm3. It is believed that persons infected with E.canis must on average have a lower white blood cell count.**

**4a) What are the null and alternative hypotheses for a one-sided test?**

Our Null hypothesis $H_0$ is that there is no difference between the two means of those infected with *E.canis* and those who are not. In other words, the mean white blood cell count (WBC) $\mu_{Infected} = \mu_{Not\ Infected}$. The alternative hypothesis $H_A$ is that those with an infection have a lower white blood cell count on average than those without. In other words, the mean WBC can be represented as $\mu_{Infected} < \mu_{Not\ Infected}$.

**4b) For a sample of 15 infected persons, the mean WBC is $\bar{x} = 4767/mm^3$ and the standard deviation is $s = 3204/mm^3$. Carry out the one-sided test using a 0.05 significance level.**

We can use R's built in functionality for this question again using the `qt` function.

```
# Define values for alpha, sample mean/sd, etc...
alpha <- 0.05
sample_mean <- 4767
population_mean <- 7250
sample_sd <- 3204
std_err <- sample_sd/sqrt(n)
n <- 15
df <- n - 1

# Determine a t-statistic for the critical value where the entire probability of
# significant values can be found in one tail. Use the qt function with our
# specified alpha and degrees of freedom.
t_critical <- qt(p = alpha, df = df)

# Calculate the value for the t-statistic
t_statistic <- (sample_mean - population_mean)/std_err

# Gather the probability with pt to determine the lower tail of probability
# for our distribution
p_value <- pt(q = t_statistic, df = df, lower.tail = TRUE)

# Display the critical T-score:
cat("The critical T value is:", round(t_critical, 4))
```

```
## The critical T value is: -1.7613
```

```
# Display the t-statistic
cat("\n The T-statistic value is:", round(t_statistic, 4))
```

```
##
##  The T-statistic value is: -3.8748
```

```r
# Display the p value
cat("\n The p value for this statistic is:", round(p_value, 6))
```

```
##
##  The p value for this statistic is: 0.000842
```

**4c) What do you conclude?**

We can make the following conclusion: We can reject the null hypothesis that there is no difference between the means and conclude at a level of 5% confidence that, in our sample, those infected with *E.canis* do indeed have lower WBC than those without infection.

# 5

## A researcher in your department asks for some statistical advice.

**5a) She is very concerned about falsely rejecting her null hypothesis. As a result, she decides to increase the sample size of her study. Explain why she is on the wrong track. What should she do instead?**

While increasing the sample size will improve the precision of her estimates for her study, differences in sample size will not change the likelihood of False Positives (Type I Errors) where she mistakenly rejects the null hypothesis.

The best way to be restrictive about the statistics she uses when evaluating whether or not her hypothesis is true is through the changing of her significance level, alpha $\alpha$. She should remember that all a p-value represents is the probability of observing an effect at least as large as the one she truly observed in her sample data assuming the null hypothesis is true. Another way of stating this, while trying to reference Type I errors, $\alpha$ demonstrates the probability of making an incorrect decision given that the null hypothesis is true.

Instead of collecting additional data points, a straightforward way to be extra careful about these types of errors is to simply use a lower value for $\alpha$ such as 0.01 instead of the commonly used 0.05, for example, when designing the tests for her data.

**5b) She also wants to know if a one-sample t-test for a mean with the dat recorded in pounds always agrees with the same test conducted on the same data recorded in kilograms. Explain.**

The difference between means represented in kilograms and pounds is that of scaling between the two units. As demonstrated in Homework 2, both the mean and median values of the data are changed when a multiplicative scalar is applied. However, this does *not* mean that the underlying statistical conclusions between scaled data would be any different!

This can be best demonstrated through an example between a dataset in pounds and the same dataset in kilograms.

```r
# Define seed for reproducibility
set.seed(4)

# First, define random dataset of pounds with rnorm
pounds <- rnorm(10, mean = 50, sd = 5)
```

```r
# Convert measurements into kilograms through a scaling factor, there are
# about 2.2046 lbs in one kilogram
kg <- pounds/2.2046

# Do one sample t-test for each dataset using the default values
t.test(pounds)
```

```
##
##  One Sample t-test
##
## data:  pounds
## t = 31.904, df = 9, p-value = 1.434e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  49.08649 56.57880
## sample estimates:
## mean of x
##  52.83265
```

```r
t.test(kg)
```

```
##
##  One Sample t-test
##
## data:  kg
## t = 31.904, df = 9, p-value = 1.434e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  22.26549 25.66398
## sample estimates:
## mean of x
##  23.96473
```

We see that, except for the mean, standard deviation, and associated confidence intervals, our values of the t-statistic, degrees of freedom, and most importantly the p-values are all the same between the two datasets!