140.615.HW.3.Delahanty.Jeremy

Jeremy Delahanty

2024-02-24

1

What critical values should be used for a confidence interval for the mean of the population in each of the following situations:

First, it will be useful to describe the concept of a confidence interval.

A constant problem in statistics is that, while we would like to take a census of all available subjects, events, data, etc, it is almost never the case that the entirety of the population can be measured. Therefore, we must rely upon samples that are randomly selected from the population of interest and estimate the true population parameters. The paramters we are interested in, much of the time, are the population mean and standard deviation. The populations we measure are often stochastic. In other words, there is some element of randomness. The evidence collected through sampling is often, therefore, the source of some kind of chance or probability.

Remember, the population mean from our sample is intended to measure the mean of the larger population being studied. This measure is called a *point estimate*. However, because of randomness, our measures have some sort of *uncertainty* attached to them. The confidence interval is one way of measuring uncertainty because it offers a range of values that is likely to contain a population parameter with a specific value for confidence, often called *alpha*, or α. Critically, we talk about confidence and *not* probabilities in confidence intervals. The confidence interval either contains the parameter we're estimating, or it does not.

This *significance level*, *alplha*, explicitly defines a threshold that is set before data collection for determining whether or not a *Null Hypothesis* is true. Specifically, the significance level corresponds to the area under the curve of a distribution that is precisley equal to alpha. Some textbooks call this region the *critical region* of the distribution.

A critical value is what defines the boundaries for that distribution's associated α. This is what question 1 is asking us. If a calculated test statistic is beyond this value, the data can be considered too inconsistent with what one can expect if the null hypothesis is true.

1a) A 90% confidence interval based on n=12 observations.

We are looking for the confidence interval of the mean of the population. Given that our sample size is quite small, we should use the *Student T Distribution* for our calculations.

`R` has tools built in for using this distribution and, for this question, we need to sue the `qt` function. This is the value of the *inverse cumulative density function* of that distribution from a random variable x and specific degrees of freedom. This will give us the t-score of the p-th quantile. Further, we do not know the mean or standard deviation of the data.

```r
# Find 90% confidence interval based on n=12 observations
# Degrees of freedom is n_observations - 1
df <- 12- 1

# Use qt function with p = 0.95, df = 11
# ncp: non-centrality parameter, not used (for now...)
# lower.tail = TRUE by default, no need to change it.
# We need 5% both to the right and left of our mean, so our p-value must
# be 0.95 as this gives us the value of our tail.
critical_value <- qt(p = 0.95, df = df)

# Display the value
cat("Critical Values: +", critical_value, "/-", critical_value)
```

```
## Critical Values: + 1.795885 /- 1.795885
```

1b) A 95% confidence interval based on n=30 observations

We can use the same procedure as above for this question. This sample size is still not sufficiently large (DEBATEABLE... lots of things I've read online say that $x \geq 30$ is sufficient) for us to use a normal distribution, so we will continue with `qt`.

```r
# Find 95% confidence interval based on n=30 observations
# Degrees of freedom is n_observations - 1
df = 30 - 1

# Use qt function with p = 0.975, df = 29
# We need 2.5% on both tails, so our p-value must be 0.975 to give us one
# of the tails.
critical_value <- qt(p = 0.975, df = df)

# Display the value
cat("Critical Values: +", critical_value, "/-", critical_value)
```

```
## Critical Values: + 2.04523 /- 2.04523
```

1c) An 80% confidence interval based on n=18 observations

We can again use the same procedure as above. The sample size is still not sufficiently large for us to use a normal distribution, so we will again use `qt`.

```r
# Find 80% confidence interval based on n=18 observations
df <- 18 - 1

# Use qt function with p = 0.90, df = 17
# We need 10% on both tails, so our p-value must be 0.90 to give us each tail.
critical_value <- qt(p = 0.90, df = df)

# Display the value
cat("Critical Values: +", critical_value, "/-", critical_value)
```

```
## Critical Values: + 1.333379 /- 1.333379
```

2

A study of 400 glaucoma patients yields a sample mean of 140 mm and a sample standard deviation of 25 mm for the the following summaries for the systolic blood pressure readings. Construct the 95% and 99% confidence intervals for μ, the population average systolic blood pressure for glaucoma patients.

In this example, we have a sample size (n) of 400 patients, a sample mean $\bar{x}$ to estimate a population mean blood pressure μ of 140mmHg, and a sample standard deviation σ of 25mmHg.

A study of 400 individuals means that we have a sufficiently large sample size to assume that the data is normal. This means that we can use statistics from the *Normal Distribution*.

We will thus rely upon the `qnorm` function in `R` which returns the value of the inverse cumulative density function of a normal distribution given a random variable `p`, population mean μ, and a population standard deviation σ.

```
# The confidence interval we are looking for is using the normal distribution
# because our sample size is sufficently large (n = 400) to assume normality.
n <- 400

# Sample mean is 140mmHg
mu <- 140

# Standard deviation is 25mmHg
sd <- 25

# The 95% confidence interval can be calculated with qnorm by finding the quantile
# where the test statistic has 2.5% of the area under the curve in each tail.
# Use qnorm with p = 0.975
quantile_95 <- qnorm(p = 0.975)

# With the quantile calculated, we can calculate the specific values of blood
# pressure that beyond the 95% confidence interval critical values.
lower_bound_95 = mu - quantile_95 * sd/sqrt(n)
upper_bound_95 = mu + quantile_95 * sd/sqrt(n)

# Turn this into a vector
confidence_interval_95 <- c(lower_bound_95, upper_bound_95)

# Display the interval
cat("The 95% confidence interval for blood pressure (mmHg) is:",
    confidence_interval_95, "\n")
```

```
## The 95% confidence interval for blood pressure (mmHg) is: 137.55 142.45
```

```
# We can repeat this procedure but for the 99% confidence interval using qnorm
# for finding the quantile where the test statistic has 0.05% of the area under
# the curve for each tail.
# Use qnorm with p = 0.995, mean = 140, sd = 25
quantile_99 <- qnorm(p = 0.995)

# Once calculated, calculate specific values of blood pressure that would go
```

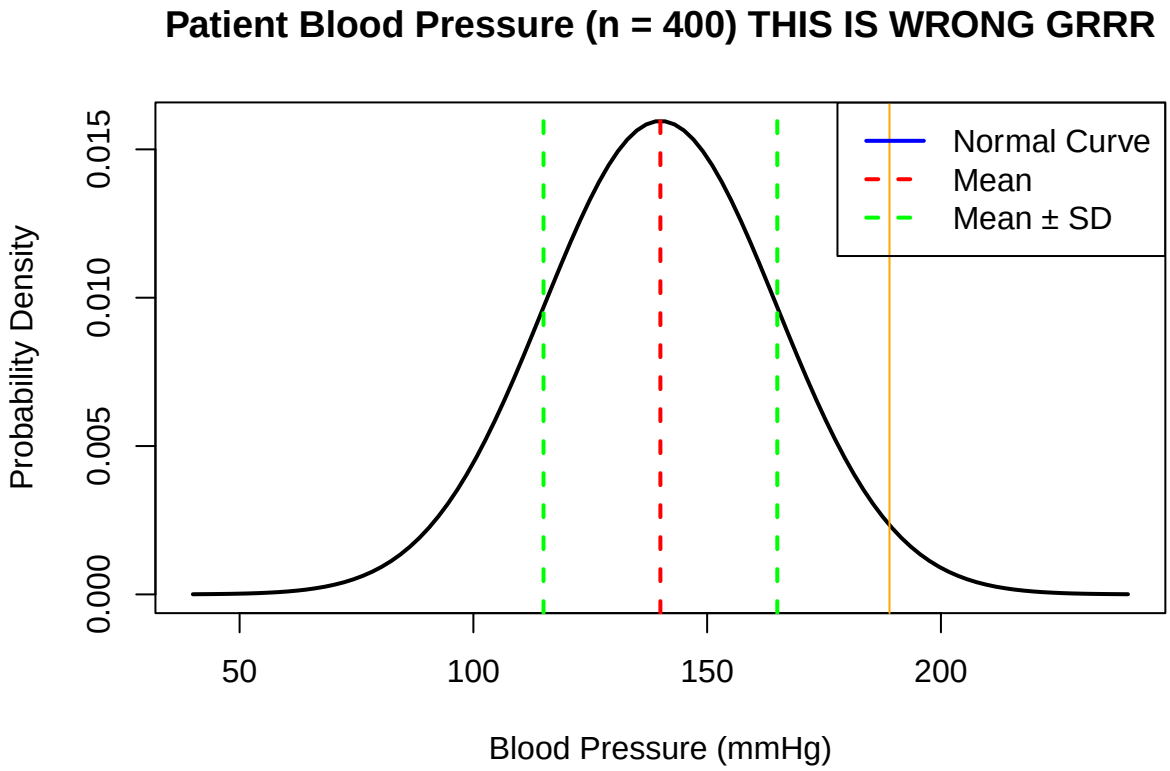
```r
# beyond the 95% confidence interval critical values.
lower_bound_99 <- mu - quantile_99 * sd/sqrt(n)
upper_bound_99 <- mu + quantile_99 * sd/sqrt(n)

# Turn this into a vector
confidence_interval_99 <- c(lower_bound_99, upper_bound_99)

# Display the interval
cat("The 99% confidence interval for blood pressure (mmHg) is:",
    confidence_interval_99)
```

```
## The 99% confidence interval for blood pressure (mmHg) is: 136.7802 143.2198
```

It would be nice to plot these in R for demonstrating the differences. I have been failing to do this properly as none of my values in this question are adding up... will have to come back to this later...

```r
# NOTE: Either my confidence intervals are very wrong, or I am just REALLY BAD
# at plotting this correctly... ignore for now...
# Generate values for x-axis, in this case the blood pressure values
# We can take the mean (mu) and both add/subtract 4 standard deviations (sd)
# from the mean to demonstrate the full width of the distribution.
# Length just means 100 numbers
x_axis <- seq(mu - 4 * sd, mu + 4 * sd, length = 100)

# We can use the dnorm function for calculating the y-axis values. Dnorm gives you
# the probability density function (PDF) values for the normal distribution.
# Remember, the PDF describes what the liklihood of getting an exact value from some
# continuous random variable is.
y_axis <- dnorm(x_axis, mean = mu, sd = sd)

# Plot the values of the x and y axis
# use a plot of type "line" and line width of 2
# color the line black
# Label the x-axis blood pressure
# Lable y-axis probability density
plot(x_axis, y_axis, type = "l", lwd = 2, col = "black", xlab = "Blood Pressure (mmHg)",
     ylab = "Probability Density",
     main = "Patient Blood Pressure (n = 400) THIS IS WRONG GRRR")

# Add line for mean and standard deviation
# v = coordinate for vertical line, centered at mean value mu
# color it red, use line type 2, meaning dashed line
# line width of 2
# for the standard deviation line, color it green
abline(v = mu, col = "red", lty = 2, lwd = 2)
abline(v = c(mu - sd, mu + sd), col = "green", lty = 2, lwd = 2)

# Introduce 95th/99th percentile lines using qnorm function
percentile_95 <- c(qnorm(p = 0.975, mean = mu, sd = sd))
abline(v = percentile_95, col = "orange")


# Add legend
```

```r
legend("topright", legend = c("Normal Curve", "Mean", "Mean ± SD"),
       col = c("blue", "red", "green"), lty = c(1, 2, 2), lwd = 2)
```



3

We measure a treatment response on a set of 6 mice from a particular strain, and get the following data: 107, 101, 93, 94, 96, 114. Imagine the data are independent draws from some normal distribution.

I will define a vector in R that can be used later on for the rest of this question.

```r
# Define vector of data
mouse_data <- c(107, 101, 93, 94, 96, 114)
n <- length(mouse_data)
```

3a) Calculate a 95% confidence interval for the population mean.

First, we must calculate the mean and standard deviation of the dataset. This is easy to do in R!

```r
# Calculate mean of data
mean <- mean(mouse_data)
```

```r
# Calculate standard deviation of data
sd <- sd(mouse_data)

# Report the values of the mean and standard deviation
cat("Mean:", mean,
    "\n Standard Deviation:", sd, "\n")
```

```
## Mean: 100.8333
##  Standard Deviation: 8.280499
```

```r
# Now that we have these values, we can compute the 95% CI for the data using
# the assumption that the data are drawn from a normal distribution.
# The 95% confidence interval can be calculated with qt by finding the quantile
# where the test statistic has 2.5% of the area under the curve in each tail.
# Use qt with p = 0.975, df = n -1
quantile_95 <- qt(p = 0.975, df = n - 1)

# With the quantile calculated, we can calculate the specific values of our
# treatment response boundaries' and critical values
lower_bound_95 = mean - quantile_95 * sd/sqrt(n)
upper_bound_95 = mean + quantile_95 * sd/sqrt(n)

# Turn this into a vector
confidence_interval_95 <- c(lower_bound_95, upper_bound_95)

# Display the interval
cat("The 95% confidence interval for mean value of treatment response is:",
    confidence_interval_95, "\n")
```

```
## The 95% confidence interval for mean value of treatment response is: 92.14348 109.5232
```

```r
# We can check our calculations using the built in t-test function in R
t.test(mouse_data)$conf
```

```
## [1]  92.14348 109.52318
## attr(,"conf.level")
## [1] 0.95
```

3b) Calculate a 95% confidence interval for the population standard deviation.

To calculate a population standard deviation where we assume that our random variable is approximately normal, independent, and identically distributed around a mean μ and standard deviation σ. Identitically distributed effectively means that there are no overall trends in the data and all samples are taken from the same probability distribution.

We can do this calculation through the use of the Chi-square distribution.

The sampling distribution of S, the estimate of σ, is described by the following:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(df = n-1)$$

Where χ^2 is the Chi-squared distribution. The values computed using this distribution can inform which critical values should be used for the range of SD values.

We want to choose a lower and upper critical value so that the probability of finding a value between these boundaries is 95%. In other words, we are solving for:

$$\begin{aligned}
95\% &= Pr\left(L \leq \frac{(n-1)S^2}{\sigma^2} \leq U\right) \\
&= Pr\left(\frac{1}{U} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{L}\right) \\
&= Pr\left(\frac{(n-1)S^2}{U} \leq \sigma^2 \leq \frac{(n-1)S^2}{L}\right) \\
&= Pr\left(S\sqrt{\frac{n-1}{U}} \leq \sigma \leq S\sqrt{\frac{n-1}{L}}\right) \\
&= \left(S\sqrt{\frac{n-1}{U}},\ S\sqrt{\frac{n-1}{L}}\right)
\end{aligned}$$

as demonstrated in lecture.

We can compute this relatively easily in R!

```
# We now know our sample standard deviation, S, currently stored as sd above
# We need to first determine the critical values for our upper and lower bounds
# In R, we can use the qchisq function for the quantile of these values
# Calculate lower bound
l <- qchisq(0.025, n-1)

# Calculate upper bound
u <- qchisq(0.975, n-1)

# Calculate the confidence interval using the upper and lower bounds
ci_sd_95 <- round(c(sd * sqrt((n-1)/u), sd * sqrt((n-1)/l)), 2)

# Display the interval
cat("The CI for the SD of the treatment response is:", ci_sd_95)
```

```
## The CI for the SD of the treatment response is: 5.17 20.31
```

3c) Calculate 95% confidence interval for the population variance.

The population variance is simply the standard deviation squared! We saw this derived above. For clarity, the relevant lines are reproduced here:

$$\begin{aligned}
95\% &= Pr\left(L \leq \frac{(n-1)S^2}{\sigma^2} \leq U\right) \\
&= Pr\left(\frac{1}{U} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{L}\right) \\
&= Pr\left(\frac{(n-1)S^2}{U} \leq \sigma^2 \leq \frac{(n-1)S^2}{L}\right)
\end{aligned}$$

We can compute this in R as well.

```r
# We now know our sample standard deviation, S, currently stored as sd above
# We need to first determine the critical values for our upper and lower bounds
# In R, we can use the qchisq function for the quantile of these values
# Calculate lower bound
l <- qchisq(0.025, n-1)

# Calculate upper bound
u <- qchisq(0.975, n-1)

# Calculate the confidence interval using the upper and lower bounds
ci_var_95 <- c(sd^2 * (n-1)/u, sd^2 * (n-1)/l)

# Display the interval
cat("The CI for the SD of the treatment response is:", ci_var_95)
```

```
## The CI for the SD of the treatment response is: 26.71602 412.4501
```

4

Consider data on the treatment response of 12 mice from strain A and 9 mice from strain B.

I will create vectors for each of these before moving forward.

```r
# Define vectors for each strain
strain_a <- c(132, 72, 102, 115, 59, 103, 86, 159, 60, 94, 80, 97)
mean_a <- mean(strain_a)
sd_a <- sd(strain_a)
na <- length(strain_a)
var_a <- sd_a^2
strain_b <- c(101, 96, 93, 106, 81, 77, 106, 97, 74)
mean_b <- mean(strain_b)
sd_b <- sd(strain_b)
nb <- length(strain_b)
var_b <- sd_b^2
```

We can see that these datasets have different numbers of samples! This will be important later.

4a) Assume that the measurements from strain A are independent draws from a normal distribution, that the the measurements from strain B are independent draws from a normal distribution, and that the population standard deviation within each strain is the same. Calculate a 95% confidence interval for the difference between the strain means A - B. Derive the answer analytically, and also using the R function t.test().

We assume that both strains are using independent draws from a normal distribution with respective means μ_A and μ_B and standard deviations within each strain is the same.

First, the analytical solution.

```r
# We are assessing whether the expected values of random variables X and y
# representing strains A and B respectively, have the same means muA and muB
# As shown in lecture, we first calculate the difference between the two
# random variables.

# Subtract means from one another
mean_difference <- mean_a - mean_b

# Calculate the estimated standard deviation of our first calculation
# Since we assume they have same standard deviation sigma at the population level
# we can compute the following weighted sigmas for each population using
# a pooled calculation
spooled <- sqrt((var_a*(na-1) + var_b*(nb -1))/(na + nb - 2))

sd_difference <- spooled*sqrt((1/na) + (1/nb))

# Now that we have our values the differences in means, we find the 97.5 percentile
# of the t distribution with n+m-2 degrees of freedom. We can do this using the
# qt function from above.
percentile_97.5 <- qt(p = 0.975,
                      df = (na + nb - 2))

# With the critical t value calculated, we can finally calculate the interval!
ci_difference_95 <- c(mean_difference - percentile_97.5*sd_difference,
                      mean_difference + percentile_97.5*sd_difference)

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

```
## The 95% CI for the difference in means is: -17.43919 25.93919
```

We can then use R's built in `t.test` functionality with the variable `var.equal` set to `TRUE` because, in this part of the problem, we are assuming that each σ is the same between the two populations.

```r
# Use R's built in t.test object for x = strain_a, y = strain_b, and set
# var.equal=TRUE
t.test(x = strain_a, y = strain_b, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  strain_a and strain_b
## t = 0.41013, df = 19, p-value = 0.6863
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -17.43919  25.93919
## sample estimates:
## mean of x mean of y
##  96.58333  92.33333
```

To clarify and more directly answer the question, we can request just the specific property of the class we care about with the dollar sign "$" symbol.

```r
# Filter for just the confidence interval to answer the question directly
t.test(strain_a, strain_b, var.equal = TRUE)$conf
```

```
## [1] -17.43919  25.93919
## attr(,"conf.level")
## [1] 0.95
```

4b) Using the R function `t.test()`, derive a 95% confidence interval for the difference between the strain means A and B allowing for the possibility that the population standard deviations between the strains might be different.

This is effectively the same thing as using the `t.test()` function as above, but we can do it where we set `var.equal=FALSE` instead.

```r
# Perform the t test with R's built in functionality, filter for just the
# confidence interval to answer the question directly
t.test(x = strain_a, y = strain_b, var.equal = FALSE)$conf
```

```
## [1] -15.56442  24.06442
## attr(,"conf.level")
## [1] 0.95
```

4c) Which of the two approaches do you prefer? What do you conclude?

I would prefer to use the second approach, where we assume that the variability between the two distributions is equal, because it makes statistical significance more demanding to reach at the level of α set to 0.05. We see directly that when the variances are assumed to be equal, we must observe more "unusual" data to reject the null. While this risks the possibility of false negatives, we should strive to minimize false positives as much as possible. Ensuring that our hypothesis testing regimens are rigorous and demanding of our data is better for science at the level of an individual study as well as science as a whole. An argument could be made for the different strains having a different standard deviation between them since their biology could be significantly altered and change the underlying distributions of the outcomes being measured, but until the experiments explicitly testing for those changes in biology are performed, it is perhaps better to assume that very similar animals have similar distributions for the variables being measured.