FIT2086 Studio 4
Central Limit Theorem and Confidence Intervals

Daniel F. Schmidt

August 17, 2017

Contents

1 Introduction

This Studio session will introduce you to using confidence intervals to quantify the accuracy of your estitates. During your Studio session, your demonstrator will go through the answers to Sections 2 through 4 with you, both on the board and on the projector as appropriate. Any questions you do not complete during the session should be completed out of class before the next Studio. Complete solutions will be released on the Friday after your Studio.

2 Confidence Interval of the Mean for Normal Populations

Under the assumption that the population follows a normal distribution $N(\mu, \sigma^2)$, the maximum likelihood estimate for the unknown mean μ is the sample mean, i.e.,

$$\hat{\mu}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

If we assume that the population variance σ^2 is known, we can construct a $100(1-\alpha)\%$ confidence interval using

$$\left(\hat{\mu}_{\text{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \ \hat{\mu}_{\text{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{1}$$

where $z_{\alpha/2}$ is the value that satisfies

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \tag{2}$$

if Z came from a standard normal distribution. This says we need to find a symmetric interval of the number line that has probability $1-\alpha$. The amount of probability not in this interval is therefore α; due to the symmetry of the normal distribution, this tells us we have

$$\alpha/2 = \mathbb{P}(Z < -z_{\alpha/2}) = \mathbb{P}(Z > z_{\alpha/2})$$

so that the probability in both of the "tails" of the distribution is the same (see the figure in Slide 50 of Lecture 4). Rewriting equation (2) as the equivalent formula

$$\mathbb{P}(Z < z_{\alpha/2}) - \mathbb{P}(Z < -z_{\alpha/2}) = 1 - \alpha, \tag{3}$$

and adding $\alpha/2$ to both sides of equation (3) tells us that we need to find the value $z_{\alpha/2}$ that satisfies

$$\mathbb{P}(Z < z_{\alpha/2}) = 1 - \alpha/2, \tag{4}$$

which is just equal to the $100(1-\alpha/2)$-th percentile of the standard normal distribution and can easily be found using the quantile function `qnorm()`. For the very common case of a 95% confidence interval, $\alpha = 0.05$ and $z_{0.05/2} = \texttt{qnorm}(1 - 0.05/2) \approx 1.96$. The quantity

$$\frac{\sigma}{\sqrt{n}}$$

is called the standard error of the estimator. In this case, the standard error is the square-root of the variance of the sample mean, which is σ^2/n. It is the average amount the estimate of the mean would vary from sample to sample, if we repeatedly resampled from our population. It is generally interpreted as a measure of accuracy of an estimate; the smaller the standard error, the more accurate your estimate is.

1. If the variance of the population goes from $\sigma^2 = 2$ to $\sigma^2 = 4$, what happens to the standard error, and what happens to the width of the inverval?

   A: Both the standard error and the confidence interval will increase by a factor of $\sqrt{4}/\sqrt{2} = \sqrt{2} \approx 1.4142$. In general, both will increase proportional to the square-root of the population variance σ^2, i.e., proportional to the population standard deviation σ.

2. If the sample size goes from $n = 10$ to $n = 100$, what happens to the standard error, and what happens to the width of the interval? If you wanted to halve the width of the interval (i.e., double your accuracy) what increase in sample size do you need?

   A: Both the standard error and the confidence interval will increase by a factor of $\sqrt{10}/\sqrt{100} = 1/\sqrt{10} \approx 0.3162$, i.e., they will get smaller. In general, both will increase proportionally to the inverse of the square-root of the sample size. This means to halve the width of the interval you will need to increase the sample size by $2^2 = 4$ (i.e., $\sqrt{4} = 2$).

3. What does a "95% confidence interval" mean?

   A: A 95% confidence interval is an interval that, if constructed for all possible samples that we could draw from our population, will contain (cover) the true population parameter for 95% of those samples.

4. Suppose we wanted a 80% confidence interval; what is the corresponding value of $z_{\alpha/2}$? *(hint: use the* `qnorm()` *function)*

   A: For an 80% confidence interval, $\alpha = 1 - 0.8 = 0.2$; from equation (4) we then need to find the value $z_{0.2/2}$ such that

   $$\mathbb{P}(Z < z_{0.1}) = 1 - 0.1$$

   if Z is distributed as per a standard normal distribution, which is the 90% percentile. We can get this from the R code

   $$\texttt{z} = \texttt{qnorm}(1 - 0.2/2)$$

   which is approximately 1.281. We can confirm our interval covers $100(1 - 0.2)\% = 80\%$ of the standard normal distribution by the code

   $$\texttt{pnorm}(\texttt{z}) - \texttt{pnorm}(-\texttt{z})$$

   which returns 0.8. As $1.281 < 1.96$, the 80% confidence interval will be narrower than the corresponding 95% confidence interval.

5. What does a 100% confidence interval look like?

   A: A 100% confidence interval covers the entire parameter space and would be $(-\infty, \infty)$. This is the only way of ensuring our interval covers the population value of the parameter for 100% of possible samples we could draw from our population. It is obviously of no real use as it tells us nothing about the possible values of the population parameter!

In practice we do not know the variance of the population, and must also estimated it from the sample along with the mean. Let us assume we estimate it using the unbiased estimate

$$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\mathrm{ML}})^2. \tag{5}$$

If we just replace σ by the estimated standard deviation $\hat{\sigma}_u$ our confidence intervals will not be quite right – they will no longer cover the unknown population μ for 95% of samples. This is because we are not taking into account the uncertainty we have in our estimating the population variance. If we do this, we instead use the interval

$$\left(\hat{\mu}_{\mathrm{ML}} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}},\ \ \hat{\mu}_{\mathrm{ML}} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right) \tag{6}$$

where $t_{\alpha/2,n-1}$ is $100(1-\alpha/2)$-th percentile of the standard Student-t distribution with $(n-1)$ degrees-of-freedom. If we compare (1) and (6) we see that the definition of the intervals for known and unknown population variance differ in two ways: (i) the unknown variance case replaces the population standard deviation σ with an estimate, and (ii) the multipliers $z_{\alpha/2}$ and $t_{\alpha/2,n-1}$ will be different. Both, however, tell us that the width of the 95% confidence interval will be a multiple of the (estimated) standard deviation – it is just the size of the multiplier that is different.

6. If we wanted a 95% confidence interval when $n = 5$, what is the corresponding value of $t_{\alpha/2,n-1}$? How does this compare to the multiplier $z_{\alpha/2}$ in the case of known variance (i.e., $z_{0.05/2} = 1.96$)? *(hint: use the* `qt()` *function and refer to Slide 58 in Lecture 4 for help)*

   A: In this case, $t_{0.05/2,4} = \texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 4) \approx 2.776$. This multiplier is larger than 1.96, as we know need to take into account the fact that the population variance σ^2 is replaced by $\hat{\sigma}_u$, which is an estimate from the data.

7. What happens to $t_{0.05/2,n-1}$ as n increases? Try $n = \{10, 50, 100, 1000\}$? How do these values compare to $z_{0.05/2} = 1.96$?

   A: We have:

$$\begin{aligned}
\texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 9) &\approx 2.262\\
\texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 49) &\approx 2.009\\
\texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 99) &\approx 1.984\\
\texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 999) &\approx 1.963
\end{aligned}$$

   So as $n \to \infty$, the multiplier essentially comes the same as the normal.

Now, let us briefly revisit the height data we examined last week. The file `test.csv` contained a sample of heights from a population that we used to estimate the parameters of a normal distribution.

8. The file `CIunknownvar.R` contains a function, `calcCI(y, alpha)` that estimates the mean, variance and provides a $100(1-\texttt{alpha})\%$ confidence interval for the estimate of the mean, from data sample `y`. As the variance is unknown, the CI is based on the the interval (6). Inspect this function to see how it works, and then use it to calculate a confidence interval on the mean from the heights data in `train.csv`. Write a statement describing your results (estimated mean, and the confidence interval) *(hint: see Slide 53 in Lecture 4 for assistance).*

   A: See `studio4_solns.R`; our statement of results is:

"The estimated mean height of people in our sample (sample size $n = 10$) is $1.659m$. We are 95% confident that the population mean height for this group is between $1.592m$ and $1.727m$."

9. Finally, load the `test.csv` data. This contains a large number of realisations from the population. Calculate the mean heights of people in this very large population, and compare it to the 95% confidence interval you calculated previously.

   A: The population mean of the heights (from `test.csv`) is $1.649m$. Our confidence interval was $(1.592, 1.727)m$, so in this case it contained the true population mean.

3 Confidence Interval of Difference of Means

During 2007 through to 2009, the world experience financial upheaval that was termed the global financial crisis. At the end of September, 2007 the Lehman Brothers investment bank collapsed, which was suggested to be the primary trigger for the worst period of economic slump. Using data from the time period September 7th, 2007 through to August 28th, 2009, we can can analyse whether there was a slump in the economy by examining the behaviour of the Standard & Poor's financial index.

1. Load the data `S&P500.csv` into R, and plot the recorded S&P index.
2. Treat the months from September 7th, 2007 through to September 26th, 2008 as one group ("pre-Lehman Brothers collapse"), and treat the data from October 3rd, 2008 through to August 28th, 2009 as a second group ("post-Lehman Brothers collapse"). Estimate the mean, standard deviation and 95% CI of the Indices for each of these two groups using the `calcCI()` function provided.

   A: See `studio4_solns.R`.

3. Calculate a difference in means between the two groups, and then find the approximate 95% confidence interval for this difference. What does the confidence interval suggest about the difference in S&P index pre- and post- the collapse the Lehman Brothers investment bank? Write a statement to summarise these findings. *(hint: use Slides 66–69 for help answering this question)*

   A: `studio4_solns.R`; our statement is:

   "The estimated difference in mean S&P Index between the 58 weeks prior to the Lehman Brothers investment bank collapse, and in the 50 weeks after the Lehman Brothers investment bank collapse was 494.7 units, i.e., the average S&P index was 494.7 units higher before the Lehman Brothers bank collapsed. We are 95% confident that the population mean difference in S&P Index between these two groups is between 486.6 units and 502.9 units. As both ends of the confidence interval are positive, and far away from zero, the data suggests that the collapse of the investment bank had an adverse effect on the United States economy, as measured by the S&P Index."

4 Maximum Likelihood Estimation of Bernoulli (Part 2)

Last week we examined maximum likelihood estimation of the probability parameter θ in the Bernoulli distribution. This week we will look at the distribution and statistics of the ML estimator. Remember

that the probability distribution for Bernoulli distribution with probability θ is:

$$\mathbb{P}(y \mid \theta) = \theta^y (1-\theta)^{1-y} \tag{7}$$

where $y \in \{0, 1\}$ is a binary variable. If Y comes from a Bernoulli distribution with probability θ, we can write $Y \sim \text{Be}(\theta)$. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a vector of n binary random variables (0 or 1). We can model this data using the n independent and identical Bernoulli distributions (7). Also recall that the maximum likelihood estimator for θ is

$$\hat{\theta}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{8}$$

We are interested in exploring the behaviour of this estimator.

1. First, derive the:

   (a) Bias;

   (b) variance;

   (c) mean squard error;

   for the estimator (8). *(hint: use the results on Slide 25–26 in Lecture 4 for help)*

   A: From Lectures 3 (and 4), we know that if an estimator is equal to the sample mean, we can easily find its bias and variance using only the mean and variance of the population. The ML estimator for the probability of success (8) of a Bernoulli distribution is of this form, so we can apply those formulae. From the properties of the Bernoulli distribution, we know that $\mathbb{E}[Y_i] = \theta$ and $\mathbb{V}[Y_i] = \theta(1-\theta)$, where θ is the probability of seeing a success. The expected value of $\hat{\theta}_{\text{ML}}$ is then

   $$\mathbb{E}\left[\hat{\theta}_{\text{ML}}\right] = \mathbb{E}[Y_i] = \theta$$

   so that the bias $b_\theta(\hat{\theta}_{\text{ML}}) = 0$, i.e, it is an unbiased estimator of θ. The variance of the estimator $\hat{\theta}_{\text{ML}}$ is then given by:

   $$\text{Var}_\theta(\hat{\theta}_{\text{ML}}) = \frac{\mathbb{V}[Y_i]}{n} = \frac{\theta(1-\theta)}{n},$$

   and the mean-squared error is then

   $$\text{MSE}_\theta(\hat{\theta}_{\text{ML}}) = b_\theta^2(\hat{\theta}_{\text{ML}}) + \text{Var}_\theta(\hat{\theta}_{\text{ML}}) = \frac{\theta(1-\theta)}{n}.$$

2. Is the maximum likelihood estimator consistent for this problem?

   A: From Lecture 3, we know that an estimator is consistent if both its bias and variance go to zero as the sample size $n \to \infty$. The bias of $\hat{\theta}_{\text{ML}}$ is zero, so the first condition is met. The variance is $\theta(1-\theta)/n$ which goes to zero as $n \to \infty$, so the estimator is consistent.

3. Next, using the central limit theorem, derive the asymptotic (large n) distribution of $\hat{\theta}_{\text{ML}}$. *(hint: look at Slides 27–30 in Lecture 4 for help).*

A: Again, we use the fact that the ML estimator of θ is equivalent to the sample mean. Then, from Lecture 4, we know that if an estimator is equal to the sample mean, it's large n distribution is

$$\hat{\theta}_{\text{ML}} \xrightarrow{d} N\left(\mathbb{E}\left[Y_i\right],\ \frac{\mathbb{V}\left[Y_i\right]}{n}\right) = N\left(\theta,\ \frac{\theta(1-\theta)}{n}\right)$$

4. Now, derive an approximate 95% confidence interval for the probability parameter θ. *(hint: use the procedure in Slides 71-75 in Lecture 4 for help)*

   A: Once again, we use the fact that the ML estimator $\hat{\theta}_{\text{ML}}$ is equivalent to the sample mean. We also note that the population variance $\mathbb{V}\left[Y_i\right] = \theta(1-\theta)$, which is a function of θ (denoted $v(\theta)$ in the Lecture notes). Then, using the procedure for approximate confidence intervals of the sample mean we can find an approximate 95% confidence interval:

$$\left(\hat{\theta}_{\text{ML}} - 1.96\sqrt{\frac{\hat{\theta}_{\text{ML}}(1-\hat{\theta}_{\text{ML}})}{n}},\ \ \hat{\theta}_{\text{ML}} + 1.96\sqrt{\frac{\hat{\theta}_{\text{ML}}(1-\hat{\theta}_{\text{ML}})}{n}}\right) \tag{9}$$

   where we have substituted the estimate $\hat{\theta}_{\text{ML}}$ into the variance of a single observation $\mathbb{V}\left[Y_i\right] = \theta(1-\theta)$.

5. Imagine we are having a game of "guess the coin toss" with our friend. She tosses the coin $n = 12$ times and we observe the following sequence of heads (coded as 1s) and tails (coded as 0s):

$$\mathbf{y} = (0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1)$$

   Estimate the value of the success probability θ using maximum likelihood, and also find the 95% approximate confidence interval using the procedure you came up with previously. Using this information, do you think your friend is using a fair coin ($\theta = 1/2$) or not?

   A: The maximum likelihood estimate of $\hat{\theta}_{\text{ML}}$ is $4/12 = 1/3$. Using this estimate in (9) yields the approximate 95% confidence interval

$$\left(1/3 - 1.96\sqrt{\frac{(1/3)(1-1/3)}{12}},\ \ 1/3 + 1.96\sqrt{\frac{(1/3)(1-1/3)}{12}}\right) \approx (0.066, 0.600)\,.$$

   In our sample of $n = 12$ coin tosses the observed probability of a head is 1/3. We are 95% confident that the true population probability of a head lies between 0.066 (heavily biased towards a tail) and 0.6 (slightly biased towards a head). The interval does not rule out the possibility of the population probability of success being 1/2 (i.e., an unbiased coin).

5 Simulation: Coverage of Confidence Intervals

We have learned that if we use the "known variance" confidence interval given by equation (1) in the situation that we actually do not know the variance, by "plugging-in" the estimated variance $\hat{\sigma}_u^2$, we are no longer guaranteed to get 95% coverage; that is, the interval will no longer contain the true,

population parameter for 95% of possible samples. To see how far off 95% the actual coverage is, we can use a computer simulation. The provided R file `CIsim.R` contains a function `testCIknownSigma2()`. The code is shown below on the last page of this document.

Let us walk through this code. The function takes a population mean `pop_mu`, population variance `pop_sigma2`, a sample size `n` and a number of times to run the simulation `niter`. For each of these iterations, the code generates a sample of `n` datapoints from a $N(\texttt{pop\_mu}, \texttt{pop\_sigma2})$ distribution. It then calculates the maximum likelihood estimate of μ, i.e., the sample mean, and computes a 95% confidence interval for the sample, by treating the population variance `pop_sigma2` as known. It then calculates whether the confidence interval contains the true population mean `pop_mu`, and finally returns the proportion of the simulation iterations for which the confidence interval covered the true value.

A: See `studio4_solns.R` for the solutions to this question.

1. Run the `testCIknownSigma2()` function for several different values of population mean, variance and sample size. Try using `niter = 10000` iterations. Do the results change as you vary the parameters, and do you expect them to change?
2. Now create a copy of the function, which you will need to modify in the following ways:
   (a) For each iteration, estimate the population variance using the unbiased estimator of variance $\hat{\sigma}_u^2$, given by equation (5)
   (b) Change the calculation of the 95% confidence interval for the "known-variance" case by replacing σ with the estimate $\hat{\sigma}_u$. This is the simple approach to obtaining (approximate) confidence intervals for $\hat{\mu}_{\text{ML}}$ when the population variance is unknown.
   (c) For each iteration, also calculate the exact 95% confidence interval for $\hat{\mu}_{\text{ML}}$ using equation (6). Extend your code to also count the number of times this interval covers the true population mean `pop_mu`. Then, also return the coverage achieved by this exact interval.
   (d) Write a loop to run this code for all the samples sizes from $\texttt{n} = 3$ up to $\texttt{n} = 100$, and record the coverage obtained by the two methods for each of these sample sizes. Plot the coverage against the sample size. How do the curves compare? What happens as `n` increases?

We can also use this program, with some minor modification, to test the approximate confidence intervals for the Poisson distribution derived in Lecture 4 (see Slide 75). Given a sample of data $\mathbf{y} = (y_1, \ldots, y_n)$, the maximum likelihood estimate of the Poisson rate parameter λ is

$$\hat{\lambda}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

which is equal to the sample mean. The approximate 95% confidence interval for $\hat{\lambda}_{\text{ML}}$ is given by

$$\left(\hat{\lambda}_{\text{ML}} - 1.96\sqrt{\hat{\lambda}_{\text{ML}}/n},\ \ \hat{\lambda}_{\text{ML}} + \sqrt{\hat{\lambda}_{\text{ML}}/n}\right). \tag{10}$$

To test how good this approximation is, do the following:

3. Create a copy of your function from the previous question, and modify it so that:
   (a) It now takes a population rate `pop_lambda` in place of a mean and variance;

(b) Each iteration, it generates a sample of size `n` random numbers from a Poisson distribution with rate `pop_lambda`

(c) For each sample, it calculates the ML estimate of λ and a confidence interval using (10), and records whether the interval contains (covers) the population rate `pop_lambda`

(d) It returns the proportion of iterations that lead to a confidence interval that covered the population rate `pop_lambda`.

4. Run your function for all combinations of `pop_lambda` $= \{1, 5, 10, 50\}$ and `n` $= \{5, 10, 25, 50, 100\}$. Create a table showing these results. How can you interpret the results? What are the observable trends as `pop_lambda` and `n` vary?

```
# Function to test coverage of confidence intervals
testCIknownSigma2 <- function(pop_mu, pop_sigma2, n, niter)
{
  retval = list()
  retval$coverage = 0

  # Do niter simulations
  for (i in 1:niter)
  {
    # Generate data from the population
    y = rnorm(n, pop_mu, sqrt(pop_sigma2))

    # Compute the 95% confidence interval
    mu_hat = mean(y)
    CI = mu_hat + c(-1.96*sqrt(pop_sigma2)/sqrt(n),
                    1.96*sqrt(pop_sigma2)/sqrt(n))

    # Does it cover the population parameter?
    if (pop_mu >= CI[1] && pop_mu <= CI[2])
    {
      retval$coverage = retval$coverage + 1
    }
  }

  # Estimate coverage as proportion of times
  # the interval covered the population parameter
  retval$coverage = retval$coverage/niter

  return(retval)
}
```