Chapter 7: Statistical Intervals Based on a Single Sample

Curtis Miller

2018-07-02

Introduction

While we appreciate a parameter estimate we know that with any estimate there is uncertainty. Rather than report a single number, statisticians prefer to report a range of plausible values for the parameter being estimated. The shorter the range, the more we know about the location of the parameter.

In this chapter we will be looking at more common statistical intervals, such as confidence intervals. We will see how to construct them and how to properly interpret them. (Statisticians care a lot about the correct interpretation!)

Section 1: Basic Properties of Confidence Intervals

A $100(1-\alpha)\%$ **confidence interval (CI)** is a **random interval** (an interval with random endpoints) intended to describe the location of a parameter θ. Suppose the endpoints of the random interval are $l(x_1, \ldots, x_n)$ and $u(x_1, \ldots, x_n)$ (recall the distinction between x_i and X_i; here, the former is an observed number, perhaps from a sample, while X_i is a random variable). The CI for θ is an interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ such that

$$\mathbb{P}\left(l(X_1, \ldots, X_n) \leq \theta \leq u(X_1, \ldots, X_n)\right) = 1 - \alpha$$

In short, in the long run, $100(1-\alpha)\%$ of intervals constructed this way capture the true value of θ.[1] Common confidence intervals include 90%, 95%, and 99%[2].

Suppose that σ is known and we have a dataset of i.i.d. data, with observed values $x_1, \ldots, x_n$. A confidence interval for the population mean μ is

This interval is exact when the data follows a Normal distribution[3] and approximately correct (due to the CLT) for large n when σ exists,

[1] This is *not* the same as saying that the *probability* the interval captured θ is $1-\alpha$. The distinction is subtle but important. When we construct a confidence interval from a particular dataset, the endpoints are not random, and *that particular interval may or may not* include the true value of θ. We have to use this frequentist notion of a long-run capture rate in order to make sense of the interval. There are intervals out there where we can refer to the probability of whether a particular interval captured the true θ, such as the Bayesian **credible interval**, but this uses a completely different theory and interpretation of probability, in addition to being more computationally difficult.

[2] Alternatively, common α includes 0.1, 0.05, and 0.01.

[3] We got this interval in the Chapter 5 notes.

for any underlying distribution. This interval takes the commonly-seen[4] form

For this interval, the **margin of error (moe)** is

Consider for a second the variables involved in the margin of error, and consider changing their values. Which variables (all others being equal) lead to the margin of error being larger when they increase? Which would lead to a decrease in the margin of error?

Consider the denominator of the moe. What is the relationship between the amount of data and the size of the moe?

Call the moe m. When planning our study we may want to specify the value of m. We do not want to change α[5], and σ is viewed as a property of nature and thus impossible to change. Thus we can only change n.

We can solve the equation for n and thus get a formula for the sample size needed to attain a margin of error m[6]:

[4] This is not law; we will see intervals not of this form.

[5] The relationship between α and m can be thought of as a trade-off between precision and accuracy. Here, *precision* refers to the size of the margin of error; it describes how well we know the location of the parameter of interest. We like being precise. We can gain precision by sacrificing *accuracy*, which is how likely the CI achieves its goal of containing the parameter of interest. While we want to be precise, we also want to be accurate, and wider intervals are naturally more accurate, all else being equal (or *ceteris paribus*, as the economists like to say). The only way to gain precision without sacrificing accuracy is increasing the sample size, n.

[6] The textbook has a similar formula but it involves the **width** of the CI, which is $w = 2m$. I prefer to use the margin of error here.

Example 1

An automated assembly line producing ball bearings should produce bearings with a diameter of 5mm. Quality control personnel run the line and get a sample of ten bearings. The bearings are known to have a standard deviation of $\sigma = 0.1$ mm[7]. The measured ball bearing diameters are listed below:

```
bearings <- c(10.396, 10.497, 10.655, 10.578, 10.543,
              10.575, 10.563, 10.549, 10.546, 10.489)
mean(bearings)
```

```
## [1] 10.5391
```

1. Construct a 95% CI for the mean diameter of the ball bearings.

2. Management is not satisfied with the margin of error, and want an estimate accurate up to 0.01 mm. Find a sample size n that attains this (while using a 95% CI).

[7] This assumption is clearly unrealistic; not only that, the mean μ is usually known before σ is, as you should expect from your study of probability and the nature of σ; thus its unlikely to see a study where σ is known but not μ. We will see in the next section what happens when we drop this assumption, but if n is large, you could replace σ with the sample standard deviation s and still get a quality CI, thanks to the law of large numbers and a result known as Slutzky's Theorem [Slutsky, 1925].

```
suppressPackageStartupMessages(library(BSDA))
z.test(bearings, sigma.x = 0.1)
```

```
##
##  One-sample z-Test
##
## data:  bearings
## z = 333.28, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  10.47712 10.60108
## sample estimates:
## mean of x
##   10.5391
```

```
ceiling((qnorm(1 - .05/2) * 0.1 / 0.01)^2)  # Needed n
```

```
## [1] 385
```

In many cases we can get formulas for confidence intervals that are either exact (if the assumptions hold) or approximately accurate for large n. This is not always the case, though, and we may need to use numerical techniques, such as **bootstrapping**, to get confidence intervals. This involves resampling the data and computing an estimate of the parameter of interest, $\hat{\theta}$, many times to get an estimate of the sampling distribution of $\hat{\theta}$. The percentiles of the simulated data can then be used to form the confidence interval.

Example 2

1. Use bootstrapping to estimate a 95% CI for the mean ball bearing diameter mentioned in Example 1.

```
(xbar <- mean(bearings))  # Estimate
```

```
## [1] 10.5391
```

```
xstars <- replicate(1000, {  # Simulations
  sim_bearings <- sample(bearings, size = 10, replace = TRUE)
  mean(sim_bearings)
})
head(xstars)
```

```
## [1] 10.5606 10.5395 10.5509 10.5280 10.5219
## [6] 10.5170
```

```
(xbarstar <- mean(xstars))  # Mean of simulated means
```

```
## [1] 10.53926
```

```
# Percentiles of simulated means
(xbar_perc <- quantile(xstars - xbarstar, c(0.025, 0.975)))
```

```
##       2.5%      97.5%
## -0.0396615  0.0379435
```

```
(xbar + xbar_perc)  # Bootstrap-estimated CI
```

```
##     2.5%    97.5%
## 10.49944 10.57704
```

2. Repeat the above procedure for the sample median. Which interval is more precise?

```
# Below I committed a programming sin: copy/paste programming!
# I should have written a function to generalize the
# procedure. But I have other goals, such as showing the
# intermediate steps.
(xtilde <- median(bearings))  # Estimate
```

```
## [1] 10.5475
```

```
xstars2 <- replicate(1000, {  # Simulations
  sim_bearings <- sample(bearings, size = 10, replace = TRUE)
  median(sim_bearings)
})
head(xstars)
```

```
## [1] 10.5606 10.5395 10.5509 10.5280 10.5219
## [6] 10.5170
```

```
(xtildestar <- mean(xstars2))  # Mean of simulated medians
```

```
## [1] 10.5469
```

```
# Percentiles of simulated medians
(xtilde_perc <- quantile(xstars2 - xtildestar, c(0.025, 0.975)))
```

```
##      2.5%     97.5%
## -0.053901  0.028099
```

```
(xtilde + xtilde_perc)  # Bootstrap-estimated CI
```

```
##    2.5%   97.5%
## 10.4936 10.5756
```

```r
# Compare widths
(w1 <- diff(xbar_perc))  # Ignore the column name; not informative here
```

```
##    97.5%
## 0.077605
```

```r
(w2 <- diff(xtilde_perc))  # Wider
```

```
## 97.5%
## 0.082
```

```r
(w2 / w1 - 1) * 100  # The percentage by which the second interval is larger
```

```
##    97.5%
## 5.663295
```

Section 2: Large-Sample Confidence Intervals for a Population Mean and Proportion

The assumption that we know σ is clearly unrealistic. If n is large, though[8], we can replace σ[9] with the sample standard deviation, s. This is because of the following:

Proposition 1. *For a collection of i.i.d.r.v. $X_1, \ldots, X_n$ with sample mean $\bar{X}$ and sample standard deviation S, if $\mathbb{E}[X_1] = \mu$ and $\text{Var}(X_1) < \infty$, for n large, the approximate distribution of $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is the standard Normal distribution.*

Thus we have the (approximate) $100(1-\alpha)\%$ CI:[10]

How would we go about sample size planning in this case? Our formulas seem to require future information. The easiest approach is to guess σ, erring on the side of large values as large σ yield larger n and thus smaller margin of errors.[11]

Example 3

At the behest of management a new sample of ball bearings was collected, this time with $n = 61$ (people decided that 385 ball bearings were too many; the study should not cost that much money). The new sample mean is $\bar{x} = 10.488$ mm, and the sample standard deviation is $s = 0.105$ mm. Compute a 95% confidence interval for the mean diameter μ. Based on this CI, is it plausible the assembly line does not produce ball bearings of the desired diameter of 10 mm?

[8] As a rule of thumb, we can consider $n > 40$ as "large" in this context.

[9] In this context statisticians view σ as a **nuisance parameter**. We are not interested in the value of σ, but in order to make a statement about μ we are forced to estimate it.

[10] The quantity $s/\sqrt{n}$ is called the **standard error** of the mean, since it estimates the mean's standard deviation.

[11] This ethos of this approach is known as being "conservative", since we are trying to err on the side of more precision than desired. In this case, we err on the side of collecting more data than needed rather than collect too little and get a margin of error that is larger than desired.

```
xbar <- 10.488
s <- 0.105
n <- 61
(m <- qnorm(0.975) * s / sqrt(n))  # moe

## [1] 0.02634951

c(xbar - m, xbar + m)

## [1] 10.46165 10.51435
```

Confidence intervals have a close cousin, called **confidence bounds**[12]. The number $l(x_1, \ldots, x_n)$ is a $100(1-\alpha)$**% confidence lower bound** for a parameter θ if

$$\mathbb{P}\left(l(X_1, \ldots, X_n) \leq \theta\right) = 1 - \alpha$$

Similarly, $u(x_1, \ldots, x_n)$ is a $100(1-\alpha)$**% confidence upper bound** for a parameter θ if

$$\mathbb{P}\left(\theta \leq u(X_1, \ldots, X_n)\right) = 1 - \alpha$$

We have the following large-sample confidence bounds for the population mean μ

Example 4

The stock with ticker symbol CGM had an average daily return of 0.07% over the last 200 days, with a standard deviation of 0.8%. Compute a 99% confidence lower bound for the mean return of the stock.

[12] Confidence bounds can be viewed as confidence intervals with one of the end points being infinite.

```
0.07 - 0.8 * qnorm(.99)/sqrt(200)
```

```
## [1] -0.06159811
```

Up until now we have been working with continuous data and our objective was to describe the location of the mean μ of the data. Suppose instead that we are working with binary/Bernoulli data and want to estimate the population proportion p of "successes". We can find a confidence interval for p[13] by working with

After isolating p in the inequality so that it's bounded by two computable numbers requiring only a sample of data, we get the following confidence interval:

We can turn the CI into a confidence bound by replacing $\alpha/2$ with α and $\pm$ with $+$ or $-$, depending on whether we want an upper bound or lower bound.

Prior to our study, if we want to choose a sample size n to achieve a moe m, our sample size should be

Here, $\tilde{p}$ is a *guess* at what the population proportion will be. If we are uncomfortable with making a guess, use $\tilde{p} = 0.5$; this will maximize m and guarantee that the observed moe will not exceed m (this is the most conservative approach). If we have a belief about the location of p we could economize during data collection somewhat by choosing $\tilde{p}$ to be near our belief, bearing in mind that the close $\tilde{p}$ is to 0.5, the larger our sample size (and smaller our observed moe) will be.

Example 5

Jack Johnson and John Jackson are running for mayor of New New York. The Johnson campaign conducts a survey of voters to deter-

[13] The problem of finding a confidence interval for p demonstrates how many different procedures can be used to get different results intended to solve the same problem. Wikipedia [2018] lists eight different intervals CIs for p. The traditional CI used was

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

but this interval exhibits pathological behavior for strange combinations of n and p. The interval recommended in this class is known as the Wilson score interval, which biases the parameter estimate slightly to 0.5. An interval not mentioned in Wikipedia [2018] is the CI obtained when adding two "imaginary" successes and two "imaginary" failures to the sample; this interval seems to work well.

mine who they support in the upcoming election.

1. The Johnson campaign will be constructing a 95% CI and does not want the moe to exceed 0.03 (or 3%). What sample size does the campaign need to achieve this?

2. In aa sample of 1068 New New York voters, 560 reported they planned to vote for Jack Johnson. Construct a 95% CI for the proportion of voters supporting Johnson. Based on the CI, who is winning?

```
suppressPackageStartupMessages(library(Hmisc))
ceiling((qnorm(.975) * 0.5 / 0.03)^2)  # Sample size

## [1] 1068

binconf(560, 1068, alpha = 0.05, method = "wilson")  # CI

##   PointEst     Lower     Upper
##  0.5243446 0.4943595 0.5541551
```

Section 3: Intervals Based on a Normal Population Distribution

From this point on in the chapter, we will assume that our data is an i.i.d. random sample from a *Normal* distribution with unknown mean and standard deviation. The intervals mentioned in Section 2 work for any underlying distribution so long as n is large enough. Here, we want intervals for when n is not considered large. The procedures mentioned in this section often work fine when n is large and the data doesn't follow a Normal distribution, though.

We start with the following theorem:

Theorem 1. *Suppose $\bar{X}$ is the sample mean of n i.i.d. Normal random variables with mean μ and S is the sample standard deviation. The random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t distribution with $\nu = n - 1$ degrees of freedom (denoted $T \sim t(n-1)$).

The $t(\nu)$ distribution[14] is a probability distribution with the following properties:

[14] This distribution is often called Student's t distribution in honor of the pseudonym of William Gosset. Gosset was employed by Guinness (the brewer), and at the time Guinness was engaged in a program to make beer brewing scientific. Eventually the experiments Guinness's burgeoning R & D department wanted to conduct required statistical methods that did not yet exist, so Gosset, then one of their brewers, began studying statistics and mathematics to develop methods for addressing Guinness's problems. Gosset's work was innovative and Guinness allowed him to publish his results in journals, but in order to not attract the attention of rival brewers, Gosset published under the pseudonym "Student". [Box, 1987]

The t distribution depends on a parameter known as the **degrees of freedom (df)**. This name comes from the fact that among the n deviations $X_1 - \bar{X}, \ldots, X_n - \bar{X}$, the condition $\sum_{i=1}^{n}(X_1 - \bar{X}) = 0$ means only $n-1$ of these deviations are freely determined.

The t **critical value** $t_{\alpha,\nu}$ satisfies

Table A.5 gives critical values for the t distribution for various α and ν.

Confidence intervals based on the t distribution resemble those from the previous section, but with z_α replaced with $t_{\alpha,n-1}$.

We can get confidence bounds rather than confidence intervals by replacing ± with either $+$ or $-$ and $t_{\alpha/2,n-1}$ with $t_{\alpha,n-1}$.

Since we assume the data follows a Normal distribution, we should check that this assumption is reasonable for our dataset. Techniques for checking the normality assumption range from probability plots to box plots to statistical tests. Use whatever method you prefer.

Example 6

Assume that the diameter of the ball bearings from Example 3 follow a Normal distribution. Compute the requested CI but using the t distribution. Compare to the CI found in Example 3.

```
(m2 <- qt(0.975, df = n - 1) * s / sqrt(n))  # moe
```

```
## [1] 0.02689175
```

```
c(xbar - m2, xbar + m2)
```

```
## [1] 10.46111 10.51489
```

There are other satistical intervals than confidence intervals. A **prediction interval (PI)** is an interval intended to describe the range of values that will likely include a future observation. If we denote our future observation with X_{n+1} the interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ is a $100(1 - \alpha)\%$ PI if

For Normally distributed data our PI is given below:

Again, we can get formulas for prediction upper bounds or prediction lower bounds with the usual substitutions.

Example 7

Over the past 121 days, the daily percentage change of the price of the stock with ticker symbol CGM had the following sample mean and standard deviation:

```
mean(cgm)
```

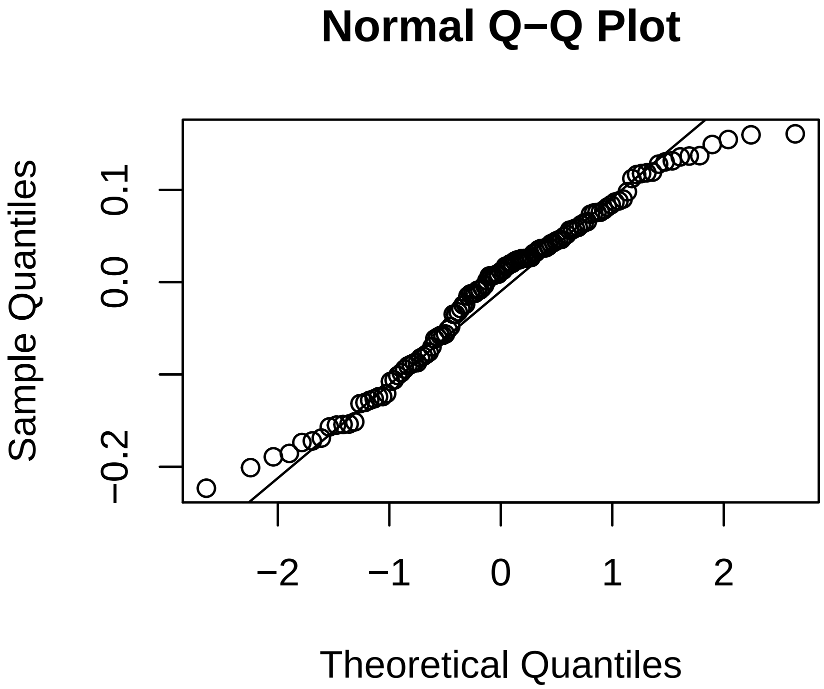
```
## [1] -0.005115001
```

```
sd(cgm)
```

```
## [1] 0.0922781
```

A look at these daily returns' probability plot suggests that we can reasonably assume that the price fluctuations follow a Normal distribution[15]:

```
qqnorm(cgm)
qqline(cgm)
```



Construct a 99% prediction lower bound for price movements.

[15] Actual asset price fluctuations are usually *not* Normally distributed. Instead, asset price fluctuations exhibit "heavy tails"; that is, extreme price movements are far more likely than the Normal distribution would suggest. Nevertheless, many models in finance for asset prices assume that price fluctuations follow a Normal distribution. See Mandelbrot and Hudson [2007] to learn more.

```
mean(cgm) - sd(cgm) * qt(.99, df = length(cgm) - 1 * sqrt(1 + 1/length(cgm)))
```

```
## [1] -0.2226907
```

Confidence intervals are meant to capture the mean and prediction intervals are meant to capture future values. **Tolerance intervals** are intervals such that at least k% of the population should be between the bounds of the interval; this statement is made with confidence level $100(1-\alpha)\%$[16].

The visualization of what is done by a tolerance interval is given below:

Tolerance intervals take the form

We still have the obvious translation to tolerance bounds. Tolerance critical values are given in Table A.6 in the textbook.

Example 8

In light of previous studies, management has instructed the assembly line producing 10mm ball bearings to retool. After the retooling a sample of 50 ball bearings is produced by the line. Management will be satisfied if 99% of ball bearings produced by the line have a diameter that is within 0.1mm of the specified diameter of 10mm. Construct a 99% tolerance interval for the diameter of the ball bearings that is correct with 95% confidence, using the following data.

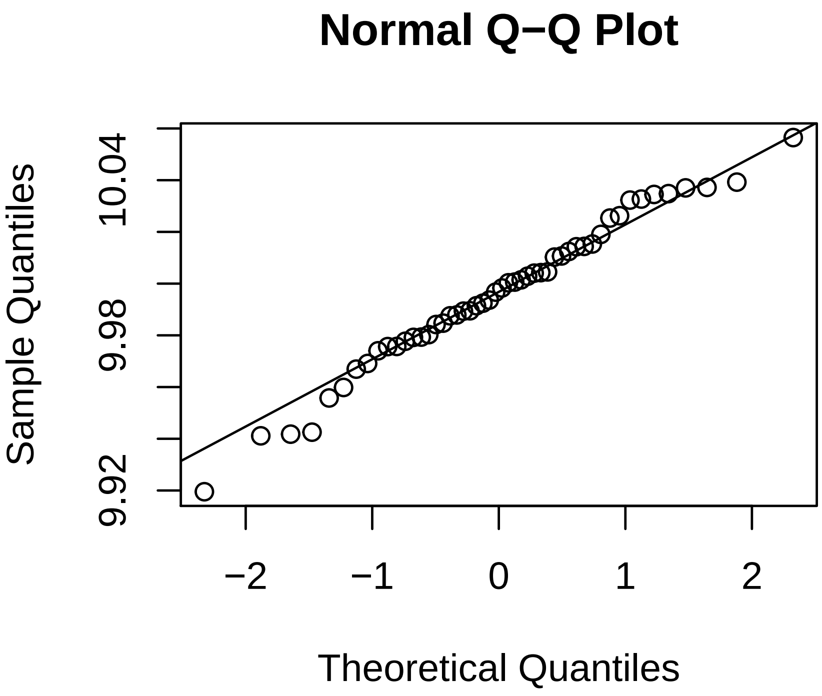
```
bearings2
```

```
##  [1] 10.001461 10.034805 10.014253  9.955770
##  [5] 10.012418  9.975701 10.000594 10.000315
##  [9] 10.010690 10.004044 10.015320 10.014393
```

[16] For example, we may have an interval such that, with 95% confidence, 99% of the population is within the bounds of the interval

```
## [13]  9.959830  9.987546  9.941776 10.026255
## [17] 10.025361 10.002873 10.019028 10.056476
## [21] 10.037196 10.004245 10.037012  9.974021
## [25] 10.039246  9.993619  9.996703  9.980263
## [29]  9.987879  9.942542  9.991442  9.941127
## [33]  9.975634  9.984178  9.989484 10.032692
## [37]  9.979320  9.977702 10.010261 10.034477
## [41] 10.004526  9.998308  9.992430  9.979205
## [45]  9.966931  9.919507  9.969121  9.989352
## [49] 10.032301  9.984713
```

```
qqnorm(bearings2)
qqline(bearings2)
```



```
mean(bearings2)
```

```
## [1] 9.996087
```

```
sd(bearings2)
```

```
## [1] 0.02894869
```

```r
# For constructing tolerance intervals
suppressPackageStartupMessages(library(tolerance))
normtol.int(bearings2, alpha = .05, P = 0.99, side = 2)
```

```
##   alpha    P    x.bar 2-sided.lower
## 1  0.05 0.99 9.996087      9.905473
##   2-sided.upper
## 1       10.0867
```

What do you do when you don't have Normally distributed data and n is not large? This depends on what you are attempting to do. Some procedures, such as the t procedures for constructing confidence intervals, are **robust** to non-normality in some contexts; that is, failure of holding to the assumption does not seem to change the end result very much. But prediction intervals and tolerance intervals are *not* robust to the normality assumption and you may need to an interval constructed for a more appropriate distribution. Bootstrapping and other non-parametric procedures (not discussed in this course) could also provide a solution. Perhaps consider reading the book Hahn and Meeker [2011] to learn about other intervals that may be useful for your problem.

Section 4: Confidence Intervals for the Variance and Standard Deviation of a Normal Population

We may be interested in constructing a confidence interval for the population variance σ^2 or standard deviation σ. We will be keeping the assumptions made in Section 3; in fact, those assumptions are more crucial. Not only are the procedures I will suggest *not* robust to the Normality assumption, if our data isn't Normally distributed, we may not even consider σ a good measure of spread in the data (especially if our underlying distribution is not symmetric).

Theorem 2. *Suppose $\bar{X}$ is the sample mean of n i.i.d. Normal random variables with mean μ and S^2 is the sample variance The random variable*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

follows a $\chi^2(n-1)$ distribution.

Let $\chi^2_{\alpha,\nu}$ satisfy

We can derive the CI for σ^2 by working with

The resulting CI is given below:[17]

We can get a CI for σ by taking the square root of the lower and upper bounds. We can get one-sided intervals by using either the upper or lower bound exclusively and replacing $\alpha/2$ with α.

Example 9

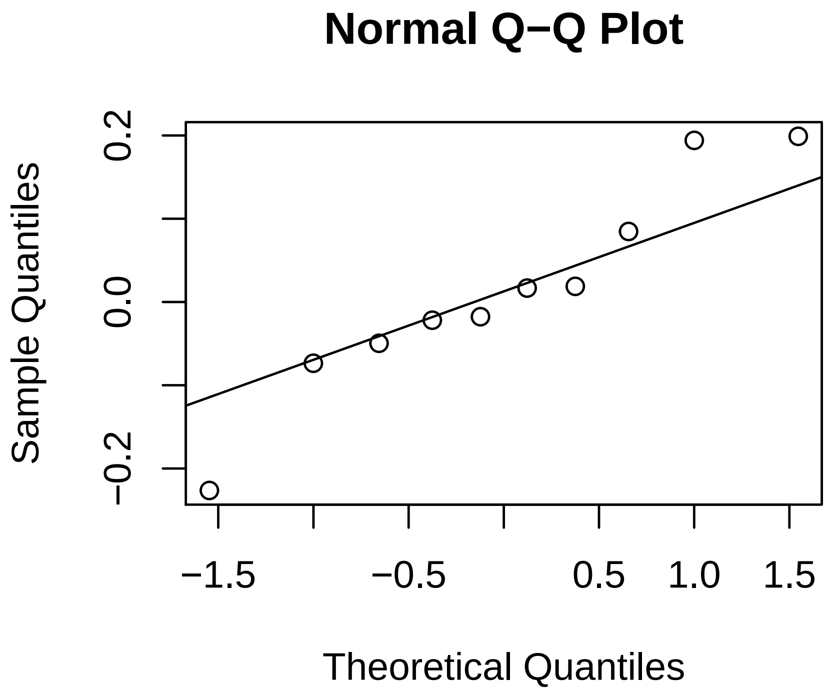
We have the following returns from the previous ten days of the stock with ticker symbol CGM:

```
cgm2 <- c(-0.2264,  0.0188, -0.0496,  0.1990,  0.1941,
          -0.0219, -0.0177,  0.0847,  0.0167, -0.0736)
```

Based on the plot below the returns seem to follow a Normal distribution:

```
qqnorm(cgm2)
qqline(cgm2)
```

[17] Notice this is *not* an equal-tail interval!



The standard deviation and variance of the stock's daily returns are given below:

```
var(cgm2)
```

```
## [1] 0.01594101
```

```
(vol <- sd(cgm2))
```

```
## [1] 0.1262577
```

Construct a 90% CI for the true σ[18] of the stock's returns.

[18] In finance, σ is frequently referred to as the *volatility* of the asset's price.

```
n <- length(cgm2)
(l <- (n - 1) * vol^2 / qchisq(.05, df = n - 1,
                               lower.tail = FALSE))  # Variance lower bound

## [1] 0.008479775

(u <- (n - 1) * vol^2 / qchisq(1 - .05, df = n - 1,
                               lower.tail = FALSE))  # Upper bound

## [1] 0.04314715

c(sqrt(l), sqrt(u))  # Bounds for the standard deviation

## [1] 0.0920857 0.2077189
```

References

Joan Fisher Box. Guinness, gosset, fisher, and small samples. *Statistical Science*, 2(1):45–52, 1987. ISSN 08834237. URL http://www.jstor.org/stable/2245613.

Gerald J Hahn and William Q Meeker. *Statistical intervals: a guide for practitioners*, volume 92. John Wiley & Sons, 2011.

Benoit Mandelbrot and Richard L Hudson. *The Misbehavior of Markets: A fractal view of financial turbulence*. Basic books, 2007.

E. Slutsky. Über stochastische Asymptoten und Grenzwerte., 1925.

Wikipedia. Binomial proportion confidence interval — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Binomial%20proportion%20confidence%20interval&oldid=832472232, 2018. [Online; accessed 06-April-2018].