Lecture Notes: Confidence Interval Estimation

October 10, 2022

1 Introduction

As we learnt in the last chapter, statistical inference involves making probabilistic statements about some aspect of a statistical model using sampled data. Following a brief discussion on common pitfalls in sampling data we discussed how maximum likelihood estimation can provide *point estimates* of parameters of a statistical model. We then focused on *point estimates* of mean and variance of a statistical model.

We learnt that the sample mean $\overline{X}_n = \sum X_i/n$ is an unbiased point estimator of the population mean ($E(X_i) = \mu$). It is called a point estimate as it provides one value $\overline{X}_n$ which serves as our best estimate of the population mean μ, based on the observed data. In this chapter we will learn an inference technique which augments the point estimate ($\overline{X}_n$) by providing an interval in which the population mean *might* be present. Before we get into it, let us remind ourselves about normally distributed random variables.

2 Normal distribution

A continuous random variable X defined in the range $(-\infty, \infty)$ is said to be normally distributed if it has the following probability density function (p.d.f.).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

The p.d.f. has two parameters μ and σ. These parameters are equal to the mean and standard deviation of the random variable X.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu \\ Var(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2 \end{aligned} \tag{2}$$

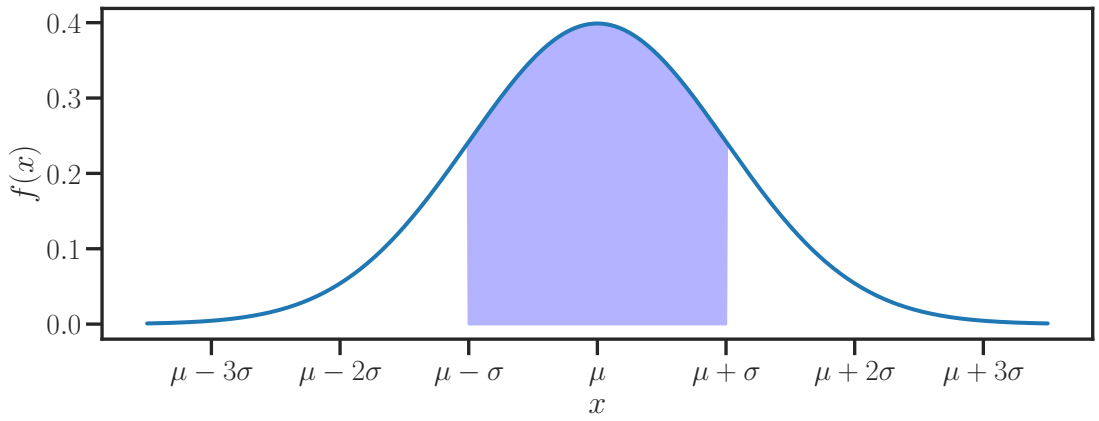


Figure 1: The figure shows the probability density function of a random variable X which is normally distributed. The shaded region corresponds to the area present under the curve within one standard deviation of the mean and is equal to $P(\mu - \sigma < X < \mu + \sigma)$.

In the above equations $E(X)$ and $Var(X)$ denote the mean (expectation) and variance of the random variable X. Let us compute the probability that X lies within one standard deviation of the mean (shaded region in figure 1).

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{3}$$

Here is a trick that is often employed to compute probabilities for normally distributed random variables. Consider the transformation $Z = (X - \mu)/\sigma$. The limits of the integral in equation 3 $X = \mu - \sigma$ and $X = \mu + \sigma$ then become $Z = -1$ and $Z = 1$. Furthermore $dz = dx/\sigma$. Equation 3 can then be rewritten as

$$P(\mu - \sigma < X < \mu + \sigma) = P(-1 < Z < 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.68 \tag{4}$$

The equation or the p.d.f. inside the integral 4 indicates that we are computing the probability associated with a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. Such a random variable Z has a special name and is referred to as the standard normal $Z \sim \mathcal{N}(0, 1)$. Let $F(z)$ be the cumulative distribution function (c.d.f.) of the standard normal. Then the probability in equation 4 can also be evaluated using the equation $P(-1 < Z < 1) = F(1) - F(-1)$.

As evident from the discussion above, $F(z)$ (c.d.f of the standard normal), can be used to compute the probability associated with any normally distributed random variable X with mean μ and standard deviation σ, between any two limits. For example the probability that $X \sim \mathcal{N}(\mu, \sigma)$ lies in the interval $[a, b]$ is equal to

$$P(a < X < b) = P((a - \mu)/\sigma) < Z < (b - \mu)/\sigma)) = F((b - \mu)/\sigma) - F((a - \mu)/\sigma) \tag{5}$$

Lower bound x_l	Upper bound x_u	Probability or Area under $\mathcal{N}(\mu, \sigma)$ between x_l & x_u
$\mu - \sigma$	$\mu + \sigma$	0.68
$\mu - 1.96\sigma$	$\mu + 1.96\sigma$	0.95
$\mu - 3\sigma$	$\mu + 3\sigma$	0.997

Table 1: Probability that $X \sim \mathcal{N}(\mu, \sigma)$ lies between the limits x_l and x_u.

You might wonder why goes through with this exercise since computing probabilities is quite straightforward for any random variable in Python. And you are right in that this method of computing probabilities was widely prevalent until a few decades ago when computers were not common. In those days the cumulative distribution function of the standard normal $F(z)$ was made available in the form of tables. Value of $F(z)$ evaluated on a fine grid on the real number line was found in these tables. You can find such a table in this link.

The table was put together by a few hard working statisticians and due to this effort computing probabilities associated with any normally distributed random variable was reduced to the simple task of looking up a table. These days we do not make use of such tables and rely on scientific packages like Scipy or programs like R. Yet understanding how the standard normal will works is useful for the techniques discussed below. In table 1 I list of probability values of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma)$, between the symmetric limits x_l and x_u. We will make use of these results later in this chapter.

3 Constructing confidence intervals

Engineers at the Morton Salt factory are interested in estimating the mean fill weight of their salt packages manufactured on a certain day. They suspect that a hardware error in the packaging equipment is causing the boxes to be incorrectly filled. They want all boxes to be as close to the labeled weight which is 200g. It is in the best interest of the company to assess the amount of salt in each packages as overfilling can hurt profits and under-filling can earn customer disapproval. A quality control engineer wants to estimate the weight of each package. But it is too time consuming to measure each one and therefore he walks to the warehouse and picks a random sample of 100 packages. The sample mean was measured to be 205 grams and the sample standard deviation was 10 grams. Below we will learn how the engineer can use this information to construct a confidence interval estimate for the population mean (mean weight of packages made)

Every box of salt in the packaging line will have a slightly different weight. Since we do not know how to accurately predict the weight of each package we can think of it as a random process. The weight of each package is then a random variable X with some unknown probability distribution $f(x)$ and an unknown mean μ and standard deviation σ. Every single package of salt that comes off the packaging line in this factory can be thought of as a random variable which has a probability distribution given by $f(x)$. Therefore we call $f(x)$ the p.d.f. of the *population* and μ as mean of the *population*.

The engineer has a pretty good estimate of the population mean μ in the sample mean $\overline{X}_n$ = 205 grams. The law of large numbers states that as sample size grows the sample mean $\overline{X}_n$ approaches μ. They also have an estimate of the standard deviation of the population σ, in the sample standard deviation s_n = 10 grams. *However, there is a possibility that the engineer who picked the samples by chance ended up picking packages that weighed much larger than 200g. How confident can the engineer be that packages have been incorrectly filled just by looking at these 100 random samples ?*

With just three pieces of information (sample size, sample mean and sample standard deviation) it is possible to express the confidence in quantitative terms by constructing what is called a confidence interval estimate for the population mean μ. Thanks to the central limit theorem, we know that if the sample size n is large enough, the probability distribution of $\overline{X}_n$ can be approximated by the normal distribution $\mathcal{N}(\mu, \sigma/\sqrt{n})$. Here μ, σ are the mean and standard deviation of the population from which the samples were obtained and n is the sample size. Both μ and σ are unknown and it is μ that we are interested in knowing about.

Let us assume that the sample standard deviation s is a good approximation for σ. The distribution of $\overline{X}_n$ can then be approximated as $\mathcal{N}(\mu, s/\sqrt{n})$. We know from our prior discussion on normally distributed random variables that there is a 95 % chance that $\overline{X}_n$ lies between the limits $\mu - 1.96s/\sqrt{n}$ and $\mu + 1.96s/\sqrt{n}$ (see table 1). In other words a sample of size 100 drawn from the population of salt packages, is highly likely (95 % chance) to have a sample mean $\overline{X}_n$, which lies within the aforementioned limits. We can express this as

$$P(\mu - 1.96\frac{s}{\sqrt{n}} \leq \overline{X}_n \leq \mu + 1.96\frac{s}{\sqrt{n}}) = 0.95 \tag{6}$$

Let us look at the limits between which $\overline{X}_n$ lies which high (95%) probability.

$$\mu - 1.96\frac{s}{\sqrt{n}} \leq \overline{X}_n \leq \mu + 1.96\frac{s}{\sqrt{n}} \tag{7}$$

Let us subtract $\mu + \overline{X}_n$ from the above inequality. This give us

$$-\overline{X}_n - 1.96\frac{s}{\sqrt{n}} \leq -\mu \leq -\overline{X}_n + 1.96\frac{s}{\sqrt{n}} \tag{8}$$

Multiplying (-1) we get

$$\overline{X}_n + 1.96\frac{s}{\sqrt{n}} \geq \mu \geq \overline{X}_n - 1.96\frac{s}{\sqrt{n}} \tag{9}$$

Note that multiplying by a negative number reverses the inequality. Rewriting the above equation we have

$$\overline{X}_n - 1.96\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X}_n + 1.96\frac{s}{\sqrt{n}} \tag{10}$$

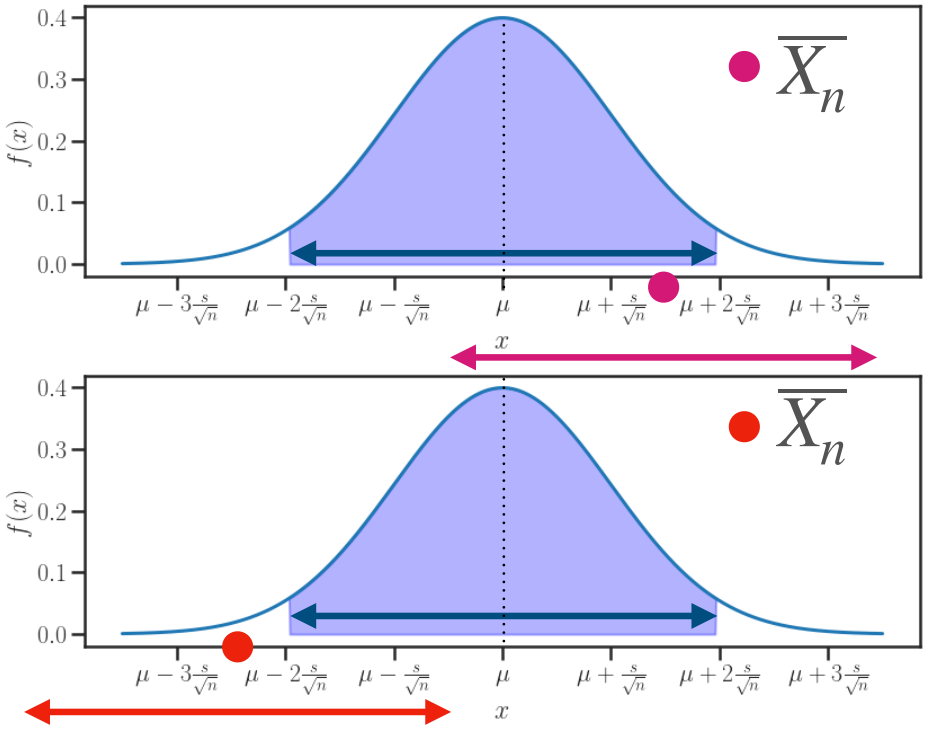


Figure 2: The figures at the top and bottom offer a visual explanation of confidence interval estimation. See text for more details

In equation 7, we started with our knowledge of normal distributions and central limit theorem to state that the sample mean $\overline{X}_n$ most likely (95 % chance) lies between the limits $[\mu - 1.96s/\sqrt{n}, \mu + 1.96s/\sqrt{n}]$. According to equation 10 the previous statement is equivalent to stating that the population mean (μ) most likely (95 % chance) lies between the limits $[\overline{X}_n - 1.96s/\sqrt{n}, \overline{X}_n + 1.96s/\sqrt{n}]$. *This is a useful result, since we have gone from just having a point estimate for μ, to providing bounds within which μ might be found.*

Figure 2 offers a visual explanation for the above argument. Look at the figure at the top. The light blue curve in the figure is the p.d.f. of the sample mean $\overline{X}_n$. The pdf is centered around the population mean μ as dictated by the central limit theorem. The area of the shaded region which is symmetric about the mean μ equals 0.95. Therefore the limits of the double headed blue arrow are $\mu - 1.96s/\sqrt{n}$ and $\mu + 1.96s/\sqrt{n}$. When a sample of size n is drawn from the population, the sample mean has a 95% chance of being within the interval given by the double headed blue arrow. The pink dot, is one such sample drawn from the population. Look at the double headed pink arrow centered around the pink dot shown below the x-axis. This arrow is equal in length to the blue one. If the pink dot (sample mean $\overline{X}_n$) lies with the double headed blue arrow, the population mean μ will certainly lie within the double headed pink arrow.

At the bottom of figure 2 is a case where the sample mean (red dot) lies outside the shaded region. Such outcomes are less likely and occur with a 5% chance. The red dot lying outside the shaded region implies that the confidence interval estimate computed from this value of $\overline{X}_n$ will not contain the population mean μ. As shown in the figure the

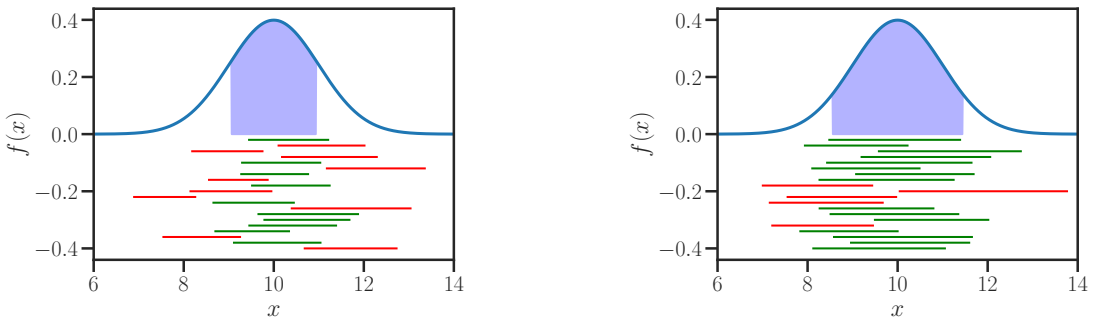
double headed red arrow centered around the sample mean does not contain the population mean μ.

The interval $[\overline{X}_n - 1.96s/\sqrt{n}, \overline{X}_n + 1.96s/\sqrt{n}]$ *is the confidence interval estimate for* μ *at the* 95% *confidence level*. Going back to the problem the engineer can estimate that the mean fill weight of the boxes lies in the interval $[205 - 1.96 \times 10/\sqrt{100}, 205 + 1.96 \times 10/\sqrt{100}] = [203.04, 206.96]$ at 95% confidence level.

3.0.1 What does a 95 % confidence interval mean ?

As figure 2 shows, some confidence interval estimates will contain the population mean μ (top), and some will not contain μ (bottom). There is no such thing as μ being probably present in an interval; it is either present or absent. However, in the long run, 95% of all confidence interval estimates computed from samples of size n from this population, will contain the population mean μ. In other words if the engineer had the practice of estimating confidence intervals every single day, 95% of his predictions will contain the true population mean μ.

3.0.2 Why pick 95% as the confidence level ?



(a) Confidence level = 0.65

(b) Confidence level = 0.85

Figure 3: In both figures the blue curve is the p.d.f. of the sample mean $\overline{X}_n$. The population from which the samples are drawn has mean $\mu = 10$. The horizontal red and green lines are confidence intervals constructed from samples of size 100 obtained from the population. The confidence levels for the plots are 65% (left), and 85% respectively. The horizontal lines are colored in green if they contain the population mean μ and are colored red if they do not.

The choice to set 95% as the confidence level is completely arbitrary. The confidence level can be increased to say 99% or more if there is a need to construct estimates which have a very low chance of not containing the population mean μ. The trade-off is that increasing the confidence level, widens the interval and therefore reduces the precision of our probabilistic prediction. On the other hand having a low confidence interval (e.g. 60%) helps us arrive at a tighter interval estimate, however most of our interval estimates will not contain the population mean μ.

Confidence level	Lower bound x_l	Upper bound x_u	Interval width(gms)
68%	$\overline{X}_n - s/\sqrt{n} = 204$	$\overline{X}_n + s/\sqrt{n} = 206$	2
95%	$\overline{X}_n - 1.96s/\sqrt{n} = 203.04$	$\overline{X}_n + 1.96s/\sqrt{n} = 206.96$	3.92
99.7%	$\overline{X}_n - 3s/\sqrt{n} = 202$	$\overline{X}_n + 3s/\sqrt{n} = 208$	6

Table 2: Confidence interval estimates for the problem presented above.

The plots in figure 3 are presented to validate the preceding argument. They show confidence interval estimates constructed for a population with mean $\mu = 10$. The green and red bars are the interval estimates constructed from 20 different samples of size 100. The estimates are colored green if they contain the sample mean and red if they do not. As the confidence level increases from 65% to 85%, the number of errors (red bars) in the estimates reduce, however the estimates become wider. If we choose to be 100% confident in our estimate (zero error), the corresponding confidence interval includes the entire range of the number line $(-\infty, \infty)$, and this estimate has no utility. In table 2 you will find a few confidence interval estimates for the problem we discussed earlier at a three different confidence levels.

3.1 One sided vs two sided confidence intervals

The interval estimates we constructed above (see table 2) provide a lower bound and an upper bound within which the population mean μ might be found. For this reason they are called as *two-sided confidence interval estimates*. Sometimes we might be interested in knowing just one of the two bounds. In the problem presented above, the engineers might just be interested in making sure the mean fill weight of the boxes is above 200 gms. In such a case we construct a one-sided confidence interval estimate since we are only interested in estimating a lower bound for the population mean μ.

The procedure to construct a one-sided confidence interval estimate for the lower bound or the upper bound is very similar to how we constructed two-sided confidence intervals. Let us pick 95% as the confidence level. Following our earlier discussion we know that the probability distribution of the sample mean $\overline{X}_n$ can be approximated by $\mathcal{N}(\mu, s/\sqrt{n})$.

We know that since $\overline{X}_n$ is normally distributed it can takes values between $(-\infty, \infty)$. Since we are only interested in the lower bound of μ, we look towards the lower end of the distribution and find the value x^*, such that 95% of all samples of size n drawn from the population have a sample mean $\overline{X}_n \leq x^*$, i.e $P(\overline{X}_n \leq x^*) = 0.95$. One way to compute x^* is to map $\overline{X}_n$ to the standard normal $Z = (\overline{X}_n - \mu)/s/\sqrt{n}$ and then find the value of z^* such that

$$\int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.95 \tag{11}$$

Here is a small code snippet that use the `scipy.stats` module to compute z^*.

```
from scipy.stats import norm
from scipy.optimize import fsolve

def t_eqn_to_solve_lb(ta, c = 0.95):

    # Return value of eqn F(z) - c = 0
    # F is the c.d.f. of the standard normal
    return norm.cdf(ta) - c


def get_z_limits_lbound ( conf = 0.95 ):

    # Use fsolve to numerically solve F(z) - c = 0
    sol = fsolve(t_eqn_to_solve_lb, x0 = 0.1, args=(conf))
    return sol

print(get_z_limits_lbound(0.95))
```

[1.64485363]

The value of $z^* \simeq 1.645$.

$$\begin{aligned} z^* &= 1.645 \\ \frac{x^* - \mu}{s/\sqrt{n}} &= 1.645 \\ x^* &= \mu + \frac{1.645s}{\sqrt{n}} \end{aligned} \tag{12}$$

As shown in figure 4 of all samples drawn from the population the lowest 95% will have a sample mean $\overline{X}_n$ that lies in the interval $[-\infty, \mu + 1.645s/\sqrt{n}]$. In other words is a 95 % chance that the sample mean is smaller than $\mu + 1.645s/\sqrt{n}$.

$$\overline{X}_n \leq \mu + 1.645\frac{s}{\sqrt{n}} \tag{13}$$

Let us subtract $\mu + \overline{X}_n$ from the above inequality. This give us

$$-\mu \leq -\overline{X}_n + 1.645\frac{s}{\sqrt{n}} \tag{14}$$

Multiplying (-1) we get

$$\mu \geq \overline{X}_n - 1.645\frac{s}{\sqrt{n}} \tag{15}$$

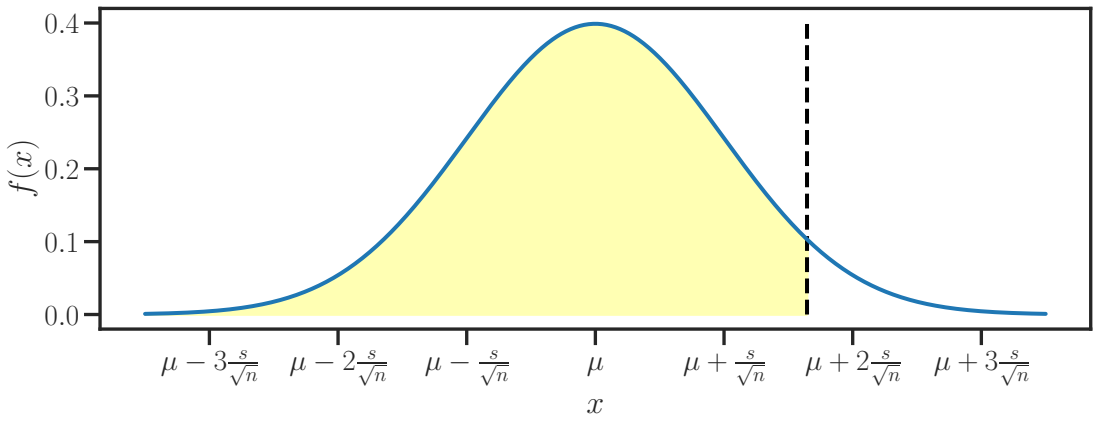


Figure 4: The graph shows the p.d.f. of the sample mean which a normally distributed random variable $X \sim \mathcal{N}(\mu, s/\sqrt{n})$. 95% of all samples drawn from this distribution will lie in the shaded region. Since we are interested in the lower bound of values obtained from this distribution, the shaded region lies towards the lower end of the distribution. The black dashed line corresponds to $\mu + 1.645s/\sqrt{n}$.

Note that multiplying by a negative number reverses the inequality. Rewriting the above equation we have

$$\overline{X}_n - 1.645\frac{s}{\sqrt{n}} \leq \mu \tag{16}$$

This implies that the population mean μ is most likely (95% chance) greater than $\overline{X}_n - 1.645\frac{s}{\sqrt{n}}$. This is our 95% confidence level estimate for the lower bound of the population μ. The engineers can have 95% confidence that the salt packages weigh at least $\overline{X}_n - 1.645\frac{s}{\sqrt{n}} = 205 - 1.645 = 203.355g$. One-sided confidence interval estimates, at a few different values of confidence level are shown in table 3.

Confidence level	Lower confidence bound
90%	$\overline{X}_n - 1.28s/\sqrt{n} = 203.72g$
95%	$\overline{X}_n - 1.645s/\sqrt{n} = 203.455g$
99%	$\overline{X}_n - 2.33s/\sqrt{n} = 202.67$

Table 3: Lower confidence bounds for the problem presented above.

Following the same idea discussed above we can construct upper confidence bounds as well. Table 4 lists the expressions you can use to estimate upper (one-sided) confidence bounds.

3.2 Estimating confidence interval estimates for small samples

In all cases presented above the sample size was large and this allows us to use the central limit theorem to approximate the sample mean $\overline{X}_n$ as a normally distributed random

Confidence level	Upper confidence bound
90%	$\overline{X}_n + 1.28s/\sqrt{n}$
95%	$\overline{X}_n + 1.645s/\sqrt{n}$
99%	$\overline{X}_n + 2.33s/\sqrt{n}$

Table 4: Expressions to compute upper confidence bounds.

variable $\overline{X}_n \sim \mathcal{N}(\mu, s/\sqrt{n})$. What do we do in situations where large sample sizes are unavailable ? This is a possibility when the sampling process is either time or resource consuming. In such situations our point estimate s of the population standard deviation σ is poor. The validity of the central limit theorem is questionable and therefore we do not have a good approximation for the probability distribution of the sample mean $\overline{X}_n$. Without knowing the probability distribution of $\overline{X}_n$ it is impossible to construct confidence intervals.

This problem was first carefully studied by William Gossett, a chemist and statistician working at the Guinness brewing company in Ireland. He was using statistical methods to quantify the quality of raw materials used in beer productions. He was constrained to work with small sample sizes and was aware of the limitations of the central limit theorem.

Gossett realized that the distribution of the sample mean $\overline{X}_n$ can be predicted even when sample sizes are small, *if the population from which the samples are drawn is normally distributed*. He was able to prove two important results.

- Consider a small sample with sample mean $\overline{X}_n$. Let the samples be drawn from a population which is normally distributed. In this case if the population standard deviation σ is known then the probability distribution of $\overline{X}_n$ is simply $\mathcal{N}(\mu, \sigma/\sqrt{n})$.
- Again consider a a small sample with sample mean $\overline{X}_n$. Let the samples be drawn from a population which is normally distributed. Let the population standard deviation σ be an unknown. The sample standard deviation is given by $s = \sqrt{\sum \frac{(X_i - \overline{X}_n)^2}{n-1}}$. Let us define a statistic t as

$$t = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \tag{17}$$

  Gossett proved that the probability distribution function (p.d.f.) of the t statistic is given by

$$f(t) = \frac{\Gamma(\nu + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \tag{18}$$

In the above equation Γ stands for the gamma function and ν is called the number of degrees of freedom of the distribution and it equals $n - 1$, where n is the sample size. Gossett published his work under the pseudonym *Student* and so the p.d.f. of the t statistic is called Student's t distribution. In figure 5 you'll find plots of the t -distribution p.d.fs for

a few different values of the parameter ν. For low values of ν the curves are broader with fat tails, since the sample standard deviation s, exhibits a lot more variability when n is small. As $\nu = n-1$ increases, or as the sample size increases, the t distribution approaches the standard normal $\mathcal{N}(0,1)$. All the p.d.f.s are symmetric and centered around $t = 0$ or equivalently around the population mean $\overline{X}_n = \mu$. Now let us learn how to use the t distribution to construct confidence interval estimates for small samples.

A metallurgist is studying a new welding process [1]. He has manufactured five welded joints and measures the yield strength of each. The five values are 56.3, 65.4, 58.7, 70.1 and 63.9. Assume that these values are a random sample from an approximately normal population. Construct a 95% *confidence interval for the mean strength of welds made by the process ?*

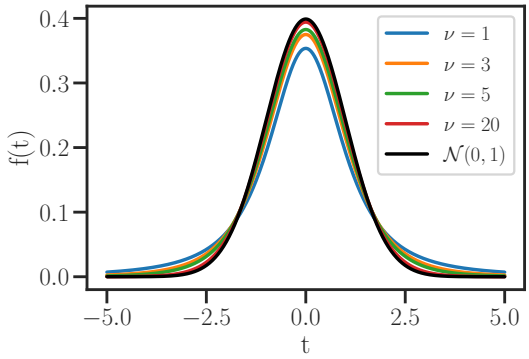


Figure 5: The graphs contain plots of the Student t-distribution function for a few select values of parameter ν. Also shown in black is the standard normal distribution function $\mathcal{N}(0,1)$. As ν (sample size) increases, the t-distribution approaches the normal distribution.

Here we have a small sample of size $n = 5$. The sample mean $\overline{X}_n = 62.88$ and the sample standard deviation $s = 5.48$. Since the sample size is small we cannot rely on the central limit theorem to give us the probability distribution of $\overline{X}_n$. Let us estimate the statistic $t = (\overline{X}_n - \mu)/s/\sqrt{n}$. The sample size $n = 5$ and $\nu = n - 1 = 4$ and therefore thanks to Gossett we know the p.d.f. of t is given by

$$f(t) = \frac{\Gamma(5/2)}{\sqrt{\pi\nu}\Gamma(2)}(1 + \frac{t^2}{4})^{-\frac{5}{2}} \tag{19}$$

Henceforth the procedure to estimate the confidence interval, is very similar to how we did it for the case of large sample sizes. Since we are constructing a two-sided confidence interval first let us find a value t^* such that

$$\int_{-t^*}^{t^*} f(t)dt = 0.95 \tag{20}$$

Here is a Python code snippet that can be used to find the value of t^*.

```
1 from scipy.stats import t,norm
2 from scipy.optimize import fsolve
3
```

[1] Examples discussed in this chapter have been taken from Statistics for Engineers and Scientists by William Navidi.

```
4  def t_eqn_to_solve(ta, nu = 2, c = 0.95):
5
6      # Return value of eqn F(t) - F(-t) - c = 0
7      # F is the c.d.f. of the t statistic
8      return t.cdf(ta,nu) - t.cdf(-ta,nu) - c
9
10
11 def get_t_limits ( nu = 2, conf = 0.95 ):
12
13     # Use fsolve to numerically solve F(t) - F(-t) - c = 0
14     sol = fsolve(t_eqn_to_solve, x0 = 0.1, args=(nu, conf))
15     return sol
16
17 print(get_t_limits(4,0.95))
```

[2.77644511]

The value of $t^* \simeq 2.776$. 95% of all samples of size 5 drawn from the population, will have a sample mean $\overline{X}_n$, such that the corresponding t value will lie between [-2.776,2.776]. We can express this as

$$P(-2.776 \leq t \leq 2.776) = 0.95 \tag{21}$$

Let us look at the range of probable t values.

$$\begin{aligned} -2.776 \leq t &\leq 2.776 \\ -2.776 \leq \frac{\overline{X}_n - \mu}{s/\sqrt{n}} &\leq 2.776 \\ -2.776\frac{s}{\sqrt{n}} \leq \overline{X}_n - \mu &\leq 2.776\frac{s}{\sqrt{n}} \end{aligned} \tag{22}$$

95% of all samples having a t value that lies in [-2.776,2.776] is equivalent to stating that 95% of all samples of size 5 will have a sample mean $\overline{X}_n$, that lies in the interval $[\mu - 2.776\frac{s}{\sqrt{n}}, \mu + 2.776\frac{s}{\sqrt{n}}]$. Adding μ we get

$$\mu - 2.776\frac{s}{\sqrt{n}} \leq \overline{X}_n \leq \mu + 2.776\frac{s}{\sqrt{n}} \tag{23}$$

Let us subtract $\mu + \overline{X}_n$ from the above inequality. This give us

$$-\overline{X}_n - 2.776\frac{s}{\sqrt{n}} \leq -\mu \leq -\overline{X}_n + 2.776\frac{s}{\sqrt{n}} \tag{24}$$

Multiplying (-1) we get

$$\overline{X}_n + 2.776\frac{s}{\sqrt{n}} \geq \mu \geq \overline{X}_n - 2.776\frac{s}{\sqrt{n}} \tag{25}$$

Rewriting the above equation we have

$$\overline{X}_n - 2.776\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X}_n + 2.776\frac{s}{\sqrt{n}} \tag{26}$$

The above inequality (eqn. 26) gives the bounds for the 95% confidence interval estimate for the population mean μ. Going back to the problem, the mean yield strength of the population at the 95% confidence level is $[62.88 - 2.776 \times 5.48/\sqrt{5}, 62.88 + 2.776 \times 5.48/\sqrt{5}] = [56.072, 69.688]$.

3.2.1 A few things to remember when using the t distribution

- For the t distribution to be used, the population from which the samples are drawn should be normally distributed. This might seem to be too restrictive a condition and you may wonder how widely applicable the t distribution is. Let X be a random variable that is a linear combination of several independent random variables, which are not necessarily identically distributed. The random variables that X is composed off, can each have a different p.d.f., yet it was proven by Lyupanov that X is normally distributed. This is why the normal distribution is very common. For example, it has been observed that height of a certain species say squirrels is typically normally distributed. Since the random variable height is composed of several independent contributions like nutrition, genetic factors, environment etc, it ends up being normally distributed.

- Samples drawn from a population that is normally distributed typically do not contain outliers. For a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma)$, there is less than a 0.3% chance, that it takes a value outside the limits $[\mu - 3\sigma, \mu + 3\sigma]$. Let us assume that we are interested in constructing a confidence interval for the mean diameter of a population of pipes. The diameters of 6 sample pipes are measured to be 2, 2.1, 2.2 ,1.9 ,2.1 6.0. While most measurements are centered around 2.0, one of them (6.0) is an outlier. Therefore it is highly likely that the diameter of pipes in the population, is not normally distributed and therefore the t distribution cannot be used to construct a confidence interval.

- In some cases, the standard deviation σ of the population might be known. In such situations if the quantity of interest is normally distributed, then the sample mean is also normally distributed, even when the sample size is small. This is because $\overline{X}_n$ is just a linear combination of several normally distributed random variables, and therefore it has been proved that $\overline{X}_n$ is also normally distributed $\overline{X}_n \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$. In such situations, confidence intervals are constructed using the appropriate normal distribution and not from the Student t distribution. Only when the standard devia-

tion of the population is unknown, and all we have is the sample standard deviation s, we use the Student t-distribution.