Nonparametric estimators

Snehal M. Shekatkar

Department of Scientific Computing, Modeling, and Simulation,
Savitribai Phule Pune University, Pune, India 411007

The estimators defined for quantities that are "distribution-free" are called nonparametric estimators. Consider the following three quantities which are distribution-free in the sense that they can be computed for any probability distribution (the formulae below assume that all the variables are continuous):

$$\mu = \int xp(x)dx \tag{1}$$

$$\sigma^2 = \int (x-\mu)^2 p(x)dx \tag{2}$$

$$r = \frac{1}{\sigma_x \sigma_y} \int\int (x-\mu_x)(y-\mu_y)p(x,y)dxdy \tag{3}$$

Note that these quantities belong to the population as a whole and not to a particular sample drawn from the population. We need to construct estimators to estimate the values of these quantities from a given sample $X_1, X_2, \cdots, X_n$. Here are the most commonly used nonparametric estimators for these three quantities in the same order:

Sample mean

$$\widehat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4}$$

Sample variance

$$\widehat{\sigma^2} = S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{5}$$

Sample correlation coefficient

$$\widehat{r} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{S_x S_y} \tag{6}$$

1 Sample mean

Consider a sample $X_1, X_2, \cdots, X_n$ of size n where X_i are IID random variables with finite mean μ and finite variance σ^2. Then the central limit theorem implies that the sample mean $\overline{X}_n$ will converge to a normal random variable in distribution. Thus, $\overline{X}_n$ is asymptotically normal, and its sampling distribution normal to a good degree of accuracy even for moderate values of n. Given that we know the shape of the sampling distribution (at least asymptotically), we can now ask two questions:

1. What is the mean of the sampling distribution of $\widehat{\mu} = \overline{X}_n$?

2. What is the standard error i.e. what is the standard deviation of the sampling distribution of $\overline{X}_n$?

These questions can be answered in a straightforward fashion by using the properties of expectation value operator $\mathbb{E}$ and variance operator $\mathbb{V}$. Consider:

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\sum_i \left(\frac{X_i}{n}\right)\right) = \sum_i \mathbb{E}\left(\frac{X_i}{n}\right) = \sum_i \left(\frac{\mu}{n}\right) = \mu \tag{7}$$

$$\mathbb{V}(\overline{X}_n) = \mathbb{V}\left(\sum_i \left(\frac{X_i}{n}\right)\right) = \sum_i \mathbb{V}\left(\frac{X_i}{n}\right) = \sum_i \left(\frac{\sigma^2}{n^2}\right) = \frac{\sigma^2}{n} \tag{8}$$

Hence the standard error is:

$$se = \sqrt{\mathbb{V}(\overline{X}_n)} = \frac{\sigma}{\sqrt{n}} \tag{9}$$

Thus, we see that the sampling distribution of the sample mean $\overline{X}_n$ is normal with mean same as the mean of the random variables X_i and standard deviation that is reduced by a factor of $\sqrt{n}$. Notice that we haven't assumed that the sample $X_1, X_2, \cdots, X_n$ is drawn from a normal distribution but only that these IID random variables have finite mean and finite variance! This also means that whenever these conditions are not satisfied, our estimator may not be valid. This can happen, for example, when X_i are drawn from the *Cauchy distribution* or *Power-law distribution*. For the Cauchy distribution, the population mean does not exist, whereas for Power-law distribution, variance is infinite which implies that the standard error cannot be reduced by taking a larger sample.

Another important observation is that the standard error goes as $1/\sqrt{n}$ where n is the sample size. This means that to reduce the variability in the estimate of $\overline{X}_n$ by a factor of N, we need to increase the sample size by a factor of N^2, not by N.

Because $\overline{X}_n$ is asymptotically normal, the random variable $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to the *standard normal variable* Z:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow Z \quad \text{where} \quad Z \sim \mathcal{N}(0,1) \tag{10}$$

For this reason, $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ is also called the Z-score of $\overline{X}_n$.

1.1 Bias and consistency

Now let us find out the bias of this estimator:

$$\text{bias}(\overline{X}_n) = \mathbb{E}(\overline{X}_n) - \mu = \mu - \mu = 0 \tag{11}$$

Thus, $\overline{X}_n$ is unbiased.

To understand the consistency of the estimator, let Z denote the standard normal random variable, and let Φ be its cumulative distribution function:

$$\Phi(z) = \mathbb{P}(Z \leq z) \tag{12}$$

Therefore,

$$\lim_{z\to\infty} \Phi(z) = 0$$

Consider,

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) &= \mathbb{P}\left(\frac{|\overline{X}_n - \mu|}{\sigma/\sqrt{n}} > \frac{\sqrt{n}\varepsilon}{\sigma}\right) = \mathbb{P}\left(|Z| > \frac{\sqrt{n}\varepsilon}{\sigma}\right) \\
&= 2\mathbb{P}\left(Z > \frac{\sqrt{n}\varepsilon}{\sigma}\right) = 2\left(1 - \mathbb{P}\left(Z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right)\right) \\
&= 2\left(1 - \Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)\right) \\
\therefore \lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) &= 2\left(1 - \lim_{n\to\infty} \Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)\right) = 2(1-1) = 0
\end{aligned}$$

Thus, $\overline{X}_n$ also converges in probability to the true mean μ and hence is also consistent.

1.2 Confidence interval for $\overline{X}_n$

Although in the limit $\overline{X}_n$ converges in probability to the true mean μ, its value for a finite sample is most likely different. In fact, as we have already seen, $\overline{X}_n$ is a normal random variable with mean μ and variance σ^2/n (provided that the conditions of the Central Limit theorem are satisfied). Because of this, only quoting a single number $\overline{X}_n$ is not much useful in most cases, and it is customary to also state a range of values within which we are confident that the true μ lies. This range is an example of a confidence set in one dimension, and is called *confidence interval for the mean* or just an "*error bar*". In principle, $\overline{X}_n$ can take any value in the range $(-\infty, \infty)$, and so if we want to quote an interval in which true μ lies with 100% confidence, then we must quote this whole interval. Of course this is a useless statement since we already know that μ lies somewhere on the real line.

But now consider an alternative perspective. Because $\overline{X}_n$ has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, we know that with high probability $\overline{X}_n$ takes values close to μ. From this perspective, we can ask what is the interval whose midpoint is μ and in which $\overline{X}_n$ lies with say 95% confidence (again, for 100% confidence, we get the useless $(-\infty, \infty)$ interval). First, let us ask this questions for the standard normal random variable Z. Let $\phi(z)$ be the probability density of Z and let $z_{\alpha/2}$ represent the value such that Z takes value in the interval $(-z_{\alpha/2}, z_{\alpha/2})$ with probability $1-\alpha$:

$$\int_{-z_{\alpha/2}}^{z_{\alpha/2}} \phi(z)dz = 1-\alpha \tag{13}$$

Here, as we have seen, $1-\alpha$ is called **level of confidence** whereas α is called **level of significance**. You can try to verify numerically that for $1-\alpha = 0.95$, $z_{\alpha/2} \approx 1.96$. Returning to the original question, since we know that $\frac{\overline{X}_n - \mu}{\sigma} \rightsquigarrow Z$ as $n \to \infty$, we see that $\overline{X}_n$ takes the value in the interval

$$\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{14}$$

with probability $1-\alpha$. Notice that μ is the midpoint of this interval, not $\overline{X}_n$. But now consider the interval

$$\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{15}$$

whose midpoint is $\overline{X}_n$. Since $\overline{X}_n$ is a random variable, this interval is also random and we can imagine constructing many such intervals for many different samples, and any such interval is called a $1-\alpha$ confidence interval for μ. **Because there is $1-\alpha$ chance that $\overline{X}_n$ deviates by amount less than $z_{\alpha/2}$ from μ, the fraction $1-\alpha$ of all such intervals contain the true value μ.** These are true $1-\alpha$ confidence intervals for μ. The remaining fraction α of intervals will not contain the true value of μ and are false $1-\alpha$ confidence intervals for μ. Obviously, when we have only one sample available, we can construct only one $1-\alpha$ confidence interval and we can't be sure whether it is true or false. But we can be reasonably sure that our constructed confidence interval contains the true mean μ.

Notice an extremely important point here: μ is a fixed unknown number and these intervals are random. Therefore, we shouldn't interpret a $1-\alpha$ confidence interval to be an interval that contains μ with probability $1-\alpha$. This is because a given interval either contains the true μ or it doesn't, and hence we can't make probability statements in this fashion. Unfortunately, even some professional data scientists interpret confidence intervals in this fashion while reporting their results.

Another important point that we should note is that here we could construct $1-\alpha$ confidence interval only because we assumed that although we don't know population mean, we do know population standard deviation σ which allows us to compute the true standard error. In practice, this is a strong assumption and hence we must estimate the standard error itself by first estimating sigma and then dividing by $\sqrt{n}$. Estimation of σ is covered in the next section.

2 Sample variance

Estimation of variance from the IID data $X_1, X_2, \cdots, X_n$ provides a good example of existence of two different estimators for the same quantity of interest. Although we have defined the sample variance as,

$$\widehat{\sigma_1^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{16}$$

There is an equally popular estimator in use:

$$\widehat{\sigma_2^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{17}$$

When n is large, it doesn't matter which of these is used. However, there is an important conceptual difference between the two estimators. To appreciate it, let us calculate the bias of these estimators by first calculating their expectation values.

$$(n-1)\mathbb{E}(\widehat{\sigma_1^2}) = \mathbb{E}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right) = \sum_{i=1}^{n}\left(\mathbb{E}(X_i^2) - 2\mathbb{E}(X_i\overline{X}_n) + \mathbb{E}(\overline{X}_n^2)\right) \tag{18}$$

Let us separately evaluate the individual expectations $\mathbb{E}(X_i^2)$, $\mathbb{E}(X_i\overline{X}_n)$, and $\mathbb{E}(\overline{X}_n^2)$.

Since

$$\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}^2(X_i)$$

we have:

$$\mathbb{E}(X_i^2) = \mathbb{V}(X_i) + \mathbb{E}^2(X_i) = \sigma^2 + \mu^2 \tag{19}$$

Now consider,

$$\mathbb{E}(X_i\overline{X}_n) = \frac{1}{n}\mathbb{E}\left(\sum_{j=1}^{n}X_iX_j\right) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_iX_j)$$

At this point it is important to realize that $\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$ *if and only if* X_i and X_j are independent. In our case, we have assumed that $X_1, X_2, \cdots, X_n$ are independent, but when $i = j$, this assumption doesn't hold, and this one term must be treated separately. Thus we have:

$$\mathbb{E}(X_i\overline{X}_n) = \frac{1}{n}\mathbb{E}(X_i^2) + \frac{1}{n}\sum_{i\neq j}\mathbb{E}(X_i)\mathbb{E}(X_j)$$

Note that there are only total $(n-1)$ terms in the summation in R.H.S. Also, the first term is already known from 19, and $\mathbb{E}(X_i) = \mu$.

$$\therefore \quad \mathbb{E}(X_i\overline{X}_n) = \frac{(\sigma^2+\mu^2)}{n} + \frac{n-1}{n}\mu^2 = \frac{\sigma^2}{n} + \mu^2 \tag{20}$$

Finally, consider:

$$\mathbb{E}(\overline{X}_n^2) = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right)$$

Again we have to be careful here. Whenever $i = j$, the random variables in the product are not independent. Since there are total n^2 terms in the double summation, and that independence doesn't hold for exactly n out of them, there are total $(n^2 - n)$ terms for which independence holds. Each of those terms can be evaluated as $\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\mathbb{E}(X_j) = \mu^2$. Hence again using 19, we get:

$$\mathbb{E}(\overline{X}_n^2) = \frac{1}{n^2} n(\sigma^2 + \mu^2) + \frac{1}{n^2}(n^2 - n)\mu^2$$

$$\therefore \quad \mathbb{E}(\overline{X}_n^2) = \frac{\sigma^2}{n} + \mu^2 \tag{21}$$

Using 19, 20 and 21 in 18, we get:

$$(n-1)\mathbb{E}(\widehat{\sigma_1^2}) = \sum_{i=1}^{n} \left[(\sigma^2 + \mu^2) - 2\left(\frac{\sigma^2}{n} + \mu^2\right) + \left(\frac{\sigma^2}{n} + \mu^2\right) \right]$$

$$(n-1)\mathbb{E}(\widehat{\sigma_1^2}) = \sum_{i=1}^{n} \left[\left(1 - \frac{1}{n}\right)\sigma^2 + (1 - 2 + 1)\mu^2 \right] = (n-1)\sigma^2$$

$$\therefore \mathbb{E}(\widehat{\sigma_1^2}) = \sigma^2 \tag{22}$$

Hence the estimator $\widehat{\sigma_1^2}$ that uses $(n-1)$ in the denominator is unbiased. But the same calculation shows that if we replace $(n-1)$ by n, then for the estimator $\widehat{\sigma_2^2}$,

$$\mathbb{E}(\widehat{\sigma_1^2}) = \frac{n-1}{n}\sigma^2,$$

and thus this estimator is biased! (However, we should note that the bias vanishes as $n \to \infty$). This is the reason that we prefer normalization using $(n-1)$ instead of n for an estimator of σ^2.

3 Degrees of freedom

Given a IID sample $X_1, X_2, \cdots, X_n$, we have n independent numbers available to us. Thus, when we calculate $\overline{X}_n$ we say that we have n degrees of freedom available to us. But now consider the process of calculating the sample variance. In this case, we must first calculate the quantities $X_i - \overline{X}_n$. However these n numbers are not actually independent. We know that their sum must be zero. Hence, once we calculate say the first $n-1$ of those, the last one is automatically fixed by this *zero-sum* requirement. When we use these quantities to calculate the sample variance, we have only $n-1$ independent numbers instead of n. Thus, we can say that for variance we have only $n-1$ degrees of freedom.

This argument can be extended to more degrees of freedom. For example, it is possible that in a given calculation, we have to find three quantities: sample mean, sample variance, one more which involves sample mean and variance. In this case, the third quantity involves only $n-2$ independent numbers since the sample mean and sample variance are fixed. Henceforth we will often make use of this phrase and will indicate the correct number whenever required.

4 Student's t-distribution

We have seen that the random variable $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ converges to the standard normal random variable Z in distribution as $n \to \infty$. Based on this, we also found how to construct a $1-\alpha$ confidence interval for μ. However, when we don't know the true value of the population standard deviation sigma, we need to replace its value by its sample estimate $\widehat{\sigma}_n$. What do we make of this? Of course again in the limit $n \to \infty$, we are absolutely fine since $\widehat{\sigma}_n$ converges to true σ in probability, and so it is a consistent estimator (We haven't really proved this, but it can be done).

But now assume that the population variance is not known and n is not really that large. That is, it is large enough to assume that the normal approximation for $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ holds. But now if we consider the following quantity by replacing σ by $\widehat{\sigma}_n$,

$$t = (\overline{X}_n - \mu)/(\widehat{\sigma}_n/\sqrt{n})$$

then normality does not hold for it.

Then would it be okay to keep using our estimated $\widehat{\sigma}_n$ in place of σ to construct a $1-\alpha$ confidence interval for μ? The answer depends on what the sampling distribution of this quantity

$$t = \frac{\overline{X}_n - \mu}{\widehat{\sigma}_n/\sqrt{n}}$$

.

Turns out that even when the original distribution of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ is standard normal, once the population σ is replaced by the sample $\widehat{\sigma}_n$, the sampling distribution changes significantly, especially when n is small. This was first discovered by William Gosset while working at Guiness Brewery where he was not allowed to publish under a real name. Thus, he published this discovery under a pen name *Student* and hence in the honour of Gosset, the sampling distribution of $t = (\overline{X}_n - \mu)/(\widehat{\sigma}/\sqrt{n})$ is now called as *Student's t-distribution.*

This is a somewhat simplified definition of the t-distribution though. The reason is that the analytical form of the distribution actually assumes that the original sample $X_1, X_2, \cdots, X_n$ are actually normally distributed IID random variables. Recall that we have been making a more general assumption that the sample consists of IID random variables with finite mean and variance but didn't assume anything about the form of their distribution. This was in the spirit of nonparametric inference. Unfortunately, when n is small, the distribution of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ actually depends a lot on the underlying distribution of the sample. This is because although $\overline{X}_n$ asymptotically does have normal distribution independent of the distribution of X_i (given finite mean and variance constraints), the rate of convergence to normal does depend on the underlying distribution.

When the IID random variables $X_1, X_2, \cdots, X_n$ actually have a Gaussian distribution, we don't have to wait till n becomes large for the Central Limit Theorem to be applicable because $\overline{X}_n$ has a normal distribution even when $n = 2$. Thus, the only thing we need to worry about is what happens when we replace σ by $\widehat{\sigma}_n$. The probability density of the distribution of t is then given by:

$$\psi(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}} \tag{23}$$

where k is called the degrees of freedom of the distribution and $\Gamma(z)$ is the Gamma function defined as:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

for $z > 0$. Thus, Student's t-distribution is not a single distribution but rather a family of distributions, one for each $k > 0$. I must admit that this whole definition looks intimidating (at least to me)! However, observe that when z is an integer, Gamma function simply reduces to factorials: $\Gamma(z) = (z-1)!$ and even when z is not an integer (e.g. when k is odd), Python lets us evaluate Gamma function in an effortless fashion! The important point to note is that after replacing σ by $\widehat{\sigma}_n$, the sampling distribution of t is a t-distribution with degrees of freedom $k = n - 1$.

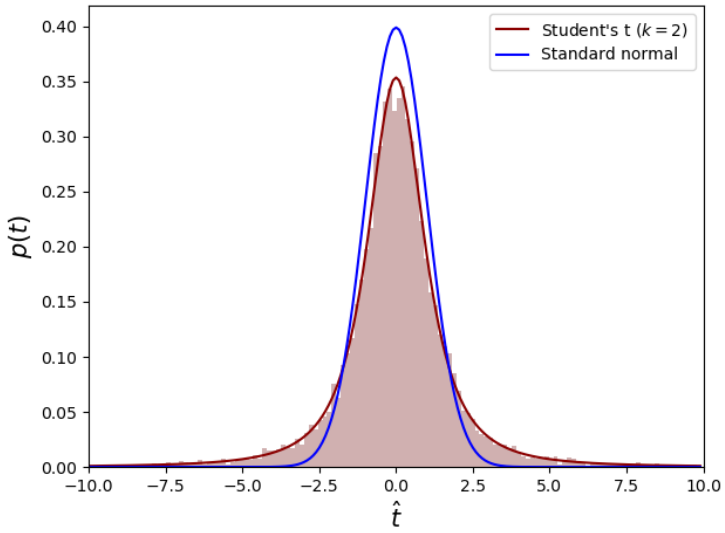

Figure 1: Histogram of t-values generated by drawing 10000 random samples each of size $n = 3$ from normal distributions with mean $\mu = 1.5$ and standard deviation $\sigma = 1.5$. The overlapped curves show that the t-distribution is a much better approximation to the histogram (especially in the tails) than the standard normal distribution.

Fig. 1 is a plot generated by this code for $n = 3$. We see that indeed the distribution of t deviates significantly from the standard normal. In particular, t-distribution has heavy tails than standard normal. Thus, if we construct $1 - \alpha$ confidence interval for μ, we assuming the sampling distribution to be normal, we will get lesser coverage than $1 - \alpha$. In other words, the fraction of confidence intervals that will contain the true value of μ will be less than $1 - \alpha$. But now that we know the correct sampling distribution, we can construct a $1 - \alpha$ confidence interval in a correct fashion. To do this, we simply need to find out the value $t_{\alpha/2}$ such that:

$$\int_{-t_{\alpha/2}}^{t_{\alpha/2}} \psi(t)dt = 1 - \alpha$$

where $\psi(t)$ is the probability density function of t-distribution given by 23. Just like the case of the standard normal, we can try to find out $t_{\alpha/2}$ for a given α. However, in most cases researchers choose the level of significance α to have one of the few values like 0.01, 0.05 or 0.1, and most statistics textbook provide tables of $t_{\alpha/2}$ for these α. Notice that $t_{\alpha/2}$ also depends on the degrees of freedom k and hence such tables need to provide these values for several values of k. One such table is shown in Table. 1 for only three values of degrees of freedom k. Of course tables in books list values for many more k.

k	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$
1	636.62	63.657	12.706
2	31.598	9.925	4.303
3	12.924	5.841	3.182

Table 1: Table listing the $t_{\alpha/2}$ values for 3 different levels of significance α and for 3 different degrees of freedom k.

Thus, if we want to construct say 95% confidence interval for μ when say $n = 3$, then we read the value of $t_{\alpha/2}$ from such table for $\alpha = 0.05$ and $k = n - 1 = 2$. The value is 4.303. As you may be expecting, you don't need such tables when you already have Python to serve you. The library 'scipy' can be used to directly find $t_{\alpha/2}$ for a given α and k as the following code shows:

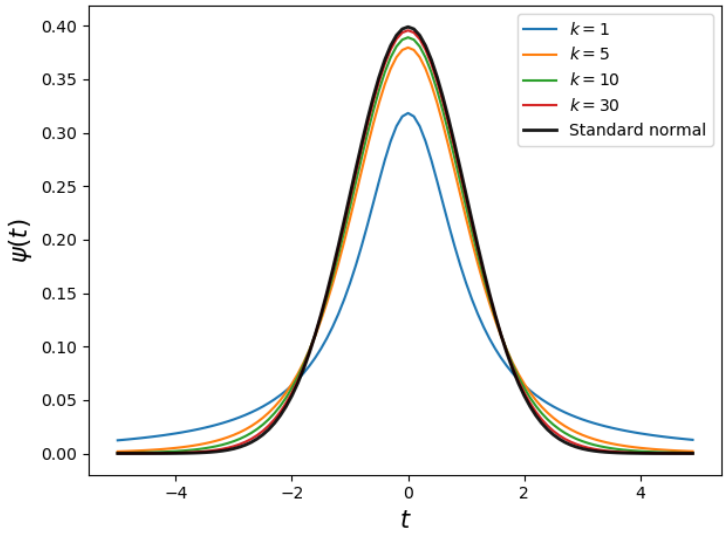

Figure 2: Plot of t-distribution for degrees of freedom $k = 1, 5, 10, 30$. The standard normal distribution is also shown for comparison.

```
from scipy.stats import t

k = 3 # degrees of freedom

for alpha in [0.001, 0.01, 0.05]:
    # Print level of significance and the corresponding critical t
    # Each critical value is rounded to 3 decimal places for printing
    print(alpha, ":", round(t.ppf(q=1-alpha/2, df=k), 3))
```

Thus, in this case our 95% confidence interval is:

$$\left(\overline{X}_n - 4.303\frac{\widehat{\sigma}}{\sqrt{n}}, \overline{X}_n + 4.303\frac{\widehat{\sigma}}{\sqrt{n}}\right)$$

Notice that if we want a confidence interval with higher confidence (i.e. larger coverage), say 99%, then the corresponding value of t would be much larger (in this case it can be seen to be 9.925) and so we get a much broader interval.

4.1 Relation between t-distribution and the standard normal distribution

I have said before that when n is large, $\widehat{\sigma}_n$ provides a good approximation to σ and using z-based confidence interval is fine even if we replace σ by $\widehat{\sigma}_n$. This implies that as n increases (i.e. k increases), t-distribution should converge to the standard normal distribution. In Fig. 2 I have plotted t-distribution curves for $k = 1, 5, 10, 30$ and also plotted the standard normal distribution for comparison. We see that t-distribution quickly converges to the standard normal. In fact for $k = 30$, the two distributions almost overlap. Another way to look at it would be to say that the standard normal is the t-distribution with an infinite number of degrees of freedom. Hence, for $n > 30$ practically it wouldn't matter whether we use $t_{\alpha/2}$ or $z_{\alpha/2}$ while constructing confidence intervals.