STAT 515 Lec 12

Confidence interval for the mean when variance unknown

Karl Gregory

Confidence interval for μ with σ unknown

Until now, we have constructed confidence intervals for the mean μ assuming that the variance σ^2 of the population is known. In practice, we will know neither μ nor σ^2. We must estimate μ *and* σ^2 from the data.

Recall that if $X_1, \dots, X_n$ are a random sample of X values such that we may assume $\bar{X}_n$ has the Normal$(\mu, \sigma^2/n)$ distribution, then we may write the probability statement

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

which we can rearrange to get

$$P\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

which implies that a $(1-\alpha)100\%$ confidence interval for μ may be constructed as

$$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Our best guess of σ from the data is the square root of the sample quantity

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

We find, however, that if we replace σ with s, then we must replace $z_{\alpha/2}$ by a larger quantity; that is, we must make the confidence interval wider to compensate for using an estimate of σ rather than the true value of σ.

Figure 1: William Sealy Gosset (1876 – 1937)

Construction of the t-distribution

In order to construct a confidence interval for μ when σ is estimated by S_n, we first note that the quantity

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

does *not* have a Normal distribution. This quantity has a distribution called a t-distribution, which was first considered by the fellow in Figure 1.

Sampling distribution result: t-distribution result

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$ and let

$$\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Then

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \quad \text{has the } t\text{-distribution with degrees of freedom equal to } n-1.$$

If a random variable T has the t-distribution with degrees of freedom ν, we write $T \sim t_\nu$.

A t-distribution arises when a standard Normal random variable Z is divided by the square root of chi-squared random variable W divided by its degrees of freedom, where Z and W are independent from each other. That is, if Z and W are independent random variables such that $Z \sim \text{Normal}(0,1)$ and $W \sim \chi^2_\nu$, the quantity

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has what is called the t-distribution with ν degrees of freedom. So a t-distribution inherits its degrees of freedom from the chi-squared random variable in its construction.

Now, we may write

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{(S_n/\sigma)} = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) \frac{1}{\sqrt{S_n^2/\sigma^2}} = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) \frac{1}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}}.$$

We know that when the population is Normally distributed,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0,1) \quad \text{and} \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We also know (from a theorem called Cochran's Theorem) that these quantities are independent from each other when the population is Normal. We may therefore claim that

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \quad \text{has the same distribution as} \quad \frac{Z}{\sqrt{W/(n-1)}},$$

where Z and W are independent random variables such that Z has the $\text{Normal}(0,1)$ distribution and W has the chi-squared distribution with degrees of freedom $n-1$.

Properties of the t-distributions

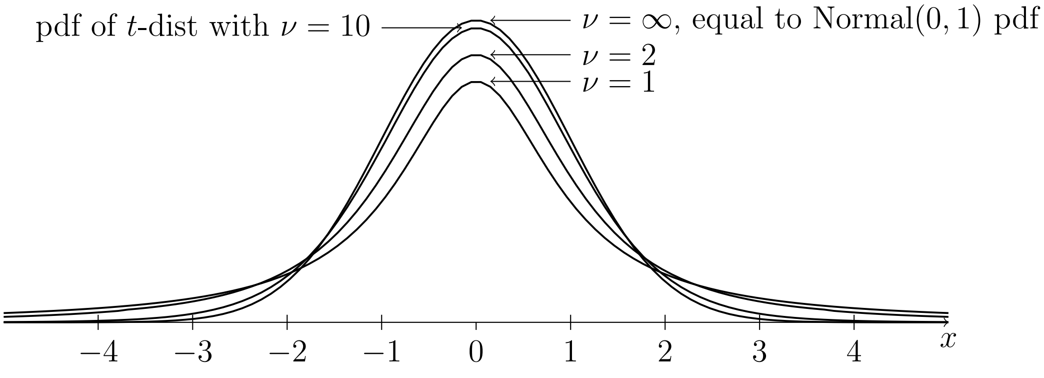
There is a different t-distribution for every degrees of freedom $\nu = 1, 2, 3, 4, \dots$. Technically, t-distributions exists for non-integer values of ν (all real numbers $\nu > 0$), but these t-distributions are not of interest to us. The t-distributions are symmetric, bell-shaped distributions centered at the origin, much like the $\text{Normal}(0,1)$ distribution, but they differ in an important way: The cardinal property of the t-distributions is that they have *heavier tails* than the $\text{Normal}(0,1)$ distribution for each $\nu < \infty$. For smaller values of ν the tails are heavier and for larger values of ν the tails become more like those of the $\text{Normal}(0,1)$ distribution.

The probability density function of the t_ν-distribution is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\nu\pi\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < x < \infty,$$

where $\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$ for $z > 0$, but we will not need to use this function in class.

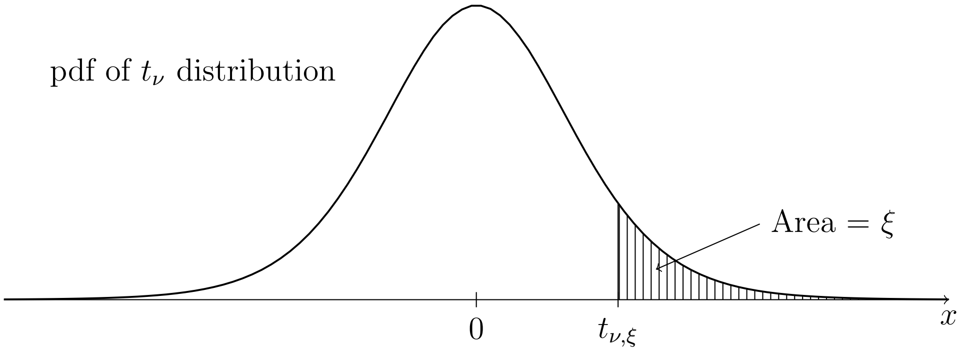
If we let $\nu \to \infty$, the probability density function of the t_ν-distribution approaches that of the $\text{Normal}(0,1)$ distribution, as the plot below depicts:



As we have done for the Normal$(0, 1)$ distribution and for the chi-squared distributions, we would like to define a symbol for each quantile of each t-distribution: For any number $0 < \xi < 1$, let $t_{\nu,\xi}$ be the value such that

$$P(T > t_{\nu,\xi}) = \xi,$$

where T is a random variable having the t-distribution with degrees of freedom ν. The quantity $t_{\nu,\xi}$ thus admits the depiction



The quantiles $t_{\nu,\xi}$ for several values of ν and some oft-encountered values of ξ are tabulated on page 817 of the textbook.

We are now prepared to give an expression for a $(1-\alpha)100\%$ confidence interval for μ when the population is Normal with unknown variance.

Confidence intervals based on the t-distribution

We work towards an expression for our confidence interval in the same way as before. Given the discussion in the previous two sections, we may write

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha.$$

Rearranging the expression in the probability statement until the mean μ remains alone in the middle, we get

$$P\left(\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right) = 1 - \alpha.$$

This implies that a $(1-\alpha)100\%$ confidence interval for μ may be constructed as

$$\bar{X}_n \pm t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}.$$

We now state this formally:

Result: Confidence interval for mean of Normal population with σ unknown

For a random sample $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, with σ unknown,

$$\bar{X}_n \pm t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$$

is a $(1-\alpha) \times 100\%$ confidence interval for μ.

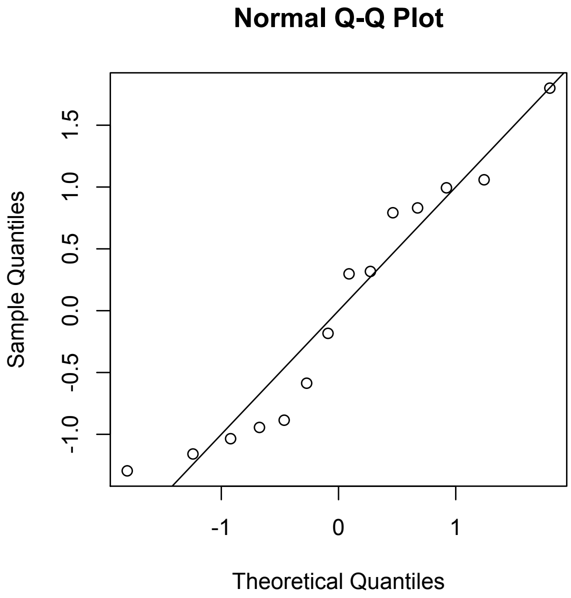
Exercise. Using the data set `Loblolly` in R, which one can access by entering `data(Loblolly)` into the console, build a 95% confidence interval for the average height μ of Loblolly pines which are ten years old.

Answer: Type

```
x <- Loblolly$height[Loblolly$age==10]
```

This stores the desired values in the vector `x`.

The sample size is $n = 14$, which we can get by entering `length(x)` into the console. Since the sample size is less than 30, we must make sure that the data are Normally distributed before we can use the confidence interval based on the t-distribution. We make a Normal QQ plot with the commands `qqnorm(scale(x))` and `abline(0,1)`. This produces the plot



There is some wiggliness to the points in the plot, but it looks pretty safe to assume that the population from which the sample was drawn is Normally distributed.

We can compute $\bar{X}_n$ by typing `mean(x)`, which gives $\bar{X}_n = 27.44214$, and we can compute s by typing `sd(x)`, which gives $s = 1.537887$. As the sample size is $n = 14$, the relevant t-distribution is the t-distribution with degrees of freedom 13. The value of α corresponding to a 95% level of confidence is $\alpha = .05$.

We get from the table on page 817 that

$$t_{13,.025} = 2.160.$$

A 95% confidence interval for μ is thus given by

$$27.44214 \pm 2.160 \frac{1.537887}{\sqrt{14}} = (26.55, 28.33).$$

Large-sample confidence interval for μ with σ unknown

So far we have assumed that the population distribution was Normal(μ, σ^2), with σ^2 unknown. We now consider the case in which the population distribution is not Normal. It turns out the confidence interval

$$\bar{X}_n \pm t_{n-1} \frac{S_n}{\sqrt{n}}$$

does not have a justification when the population distribution is not Normal. For large sample sizes, however, we find that the central limit theorem works in our favor. A result called Slutzky's theorem allows us to make the following claim about how the quantity $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ behaves when σ is replaced by S_n, as n becomes large:

Sampling distribution result: Corollary of central limit theorem and Slutzky's theorem

Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and $\mathbb{E}|X_1|^4 < \infty$. Then

$$\frac{\bar{X} - \mu}{S_n/\sqrt{n}} \text{ has a distribution more and more like the } \mathrm{Normal}(0,1)$$

distribution for larger and larger sample sizes n.

We can use the above result to argue that when our sample comes from a non-Normal population, the interval

$$\bar{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}},$$

in which $z_{\alpha/2}$ is used instead of $t_{n-1,\alpha/2}$ in spite of the fact that σ has been replaced by S_n, will still contain the population mean μ with probability close to $1-\alpha$, provided the sample size is “large”.

We now state this formally:

Result: Large-sample CI for μ, population not-Normal

Let $X_1, \dots, X_n$ be a random sample from a population with mean μ and $\mathbb{E}|X_1|^4 < \infty$. Then

$$\bar{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}}$$

is an approximate $(1-\alpha) \times 100\%$ confidence interval for μ when n is large ($n \geq 30$, say).

The condition $\mathbb{E}|X_1|^4 < \infty$ in the result is a technical one that we needn't worry about in this course; it basically says that the distribution cannot be too heavy-tailed, which ensures that accuracy of S_n in estimating σ improves as the sample size is increased.

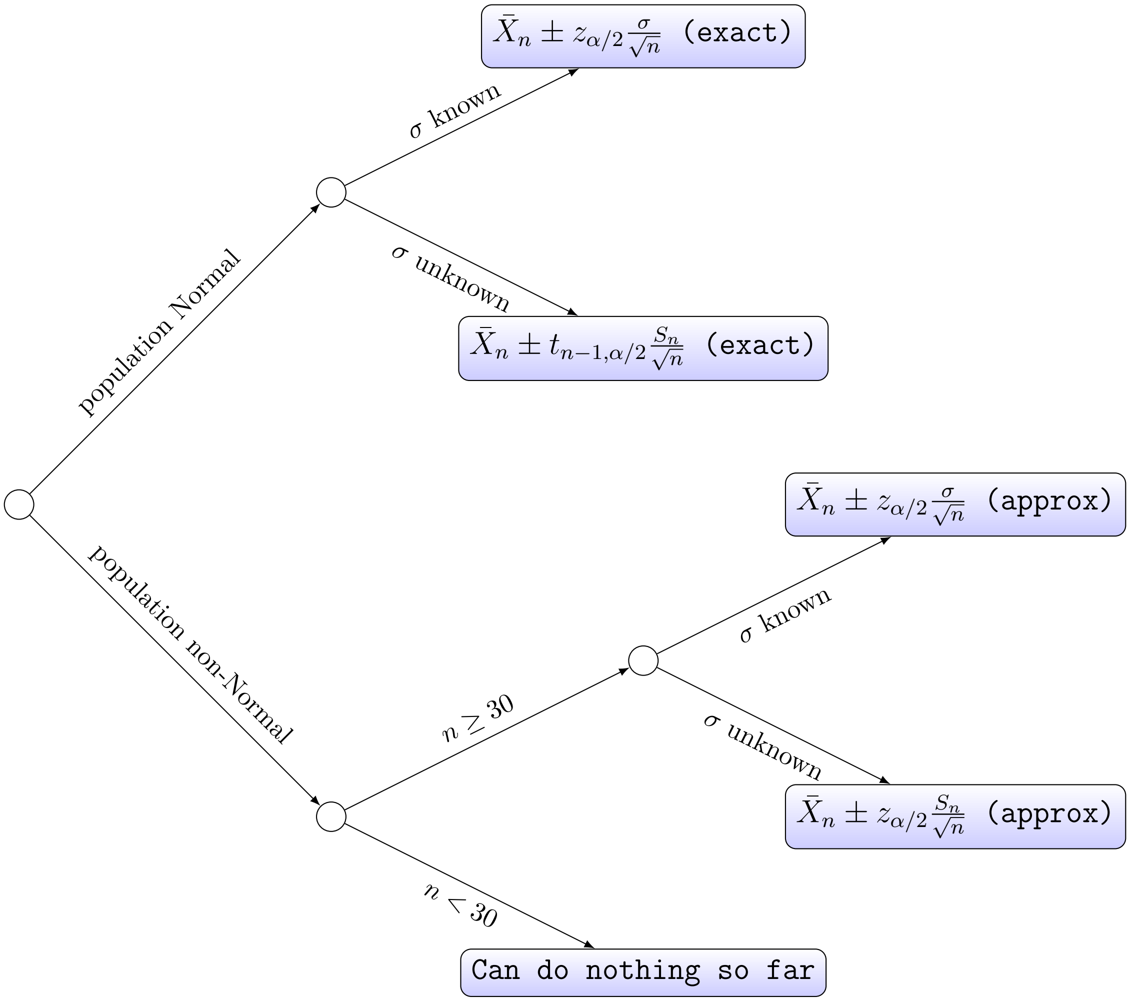
Another way to justify the confidence interval $\bar{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$ when n is large is by recalling that the t-distributions become more and more like the standard Normal distribution for larger and larger values of the degrees of freedom parameter. That is, for larger and larger sample sizes n, the quantity $t_{n-1,\alpha/2}$ becomes closer and closer to the quantity $z_{\alpha/2}$ for any choice of α. This means that for large sample sizes it is not very important whether we use the t-distribution or the standard Normal distribution to build confidence intervals!

Specifically, we see that for large n we have

$$\bar{X}_n \pm t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \approx \bar{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}}.$$

The following section summarizes when we should use which interval.

When to use which interval?


$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (exact)
σ known
σ unknown
$\bar{X}_n \pm t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}$ (exact)
population Normal
population non-Normal
$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (approx)
σ known
$n \geq 30$
σ unknown
$\bar{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}}$ (approx)
$n < 30$
Can do nothing so far

Note that in order to use the t-distribution, we *must* assume that the data come from a Normal distribution. If the data do not come from a Normal distribution, we require a large sample size, $n \geq 30$ as a rule of thumb. In the non-Normal but large-n case, we rely on the phenomenon described by the Central Limit Theorem—that the standardized sample mean becomes more and more Normal as the sample size grows.