

You already know this from previous lecture:

use Z when
 σ is known

Consider a sample from a Normal distribution with known variance σ^2 .

Then sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$

where μ =population mean

σ^2 =population variance

n =sample size

The test statistic is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

This is new to you:

this is realistic situation

use t
when σ
is
unknown

Consider a sample from a Normal distribution with unknown variance σ^2

Since σ^2 is unknown, we estimate σ^2 by its unbiased estimate s^2 .

The test statistic is $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ and it follows a t-distribution with $(n - 1)$ degrees of freedom.

For very large samples, t_{n-1} and $N(0,1)$ are almost identical and we can use Z as the test statistic.

For small samples, t_{n-1} and $N(0,1)$ distributions are vastly different and we use $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ as the test statistic.

(show class t-table)

Introducing the t-distribution

① If the population standard deviation σ is known and the underlying population is normal or sample size is large,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

② If population standard deviation σ is unknown, we estimate σ by the sample standard deviation s . (we do not use Z in this case)

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

has a t-distribution with degrees of freedom = $n - 1$

③ $100(1 - \alpha)\%$ CI for μ when σ is unknown

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Steps to calculate a one-sample t-CI.

Assumptions:

- Important
- ① Normal population
 - ② Simple random sample
 - ③ σ unknown

Step ① For confidence level $1-\alpha$, look up t-table to find $t_{\alpha/2}$ with degrees of freedom = $n - 1$

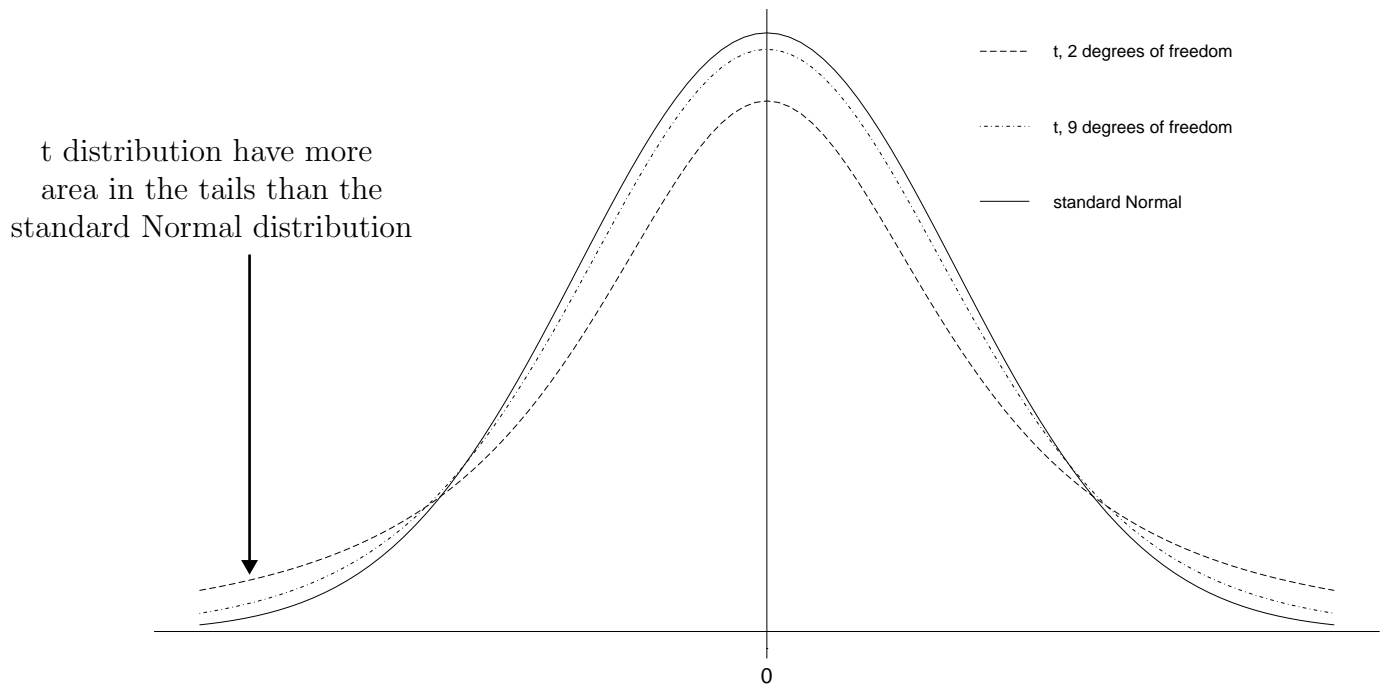
Step ② The CI for μ is

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

underlined part is called margin of error

When n is very large, s is a very good estimate of σ , and the corresponding t distribution are very close to the normal distribution.

The t distribution becomes wider for smaller sample sizes.



Example:

Same cadmium in mushroom problem, except now, assume σ is unknown.

Cadmium, a heavy metal, is toxic in animals. Mushrooms, however are able to absorb and accumulate cadmium at high concentrations. The Czech and Slovak governments have set a safety limit for cadmium in dry vegetables at 0.5 parts per million (ppm). Below are cadmium levels of a random sample of the edible mushroom *Boletus pinicola*:

0.24 0.59 0.62 0.16 0.77 1.33 0.92 0.19 0.33 0.25 0.59 0.32

Find and interpret a 95% confidence interval for the mean cadmium level of all *Boletus pinicola* mushrooms. Assume the population standard deviation of cadmium levels is unknown.

Solution

Think: check assumptions

- (a) We assume population is normal
- (b) Assume simple random sample
- (c) Note that σ is unknown (so use t)

$$\left. \begin{array}{l} \bar{x} = 0.52583 \\ s = 0.352122353 \end{array} \right\} \text{from earlier question}$$

$$n = 12$$

$$\begin{array}{l} t_{\alpha/2, n-1} = 2.201 \\ \downarrow \text{df}=11 \\ = \frac{0.05}{2} = 0.025 \end{array}$$

make sure you know to get this from t-table

95% CI for μ is

$$\begin{aligned} & \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \\ & 0.52583 \pm (2.201) \left(\frac{0.352122353}{\sqrt{12}} \right) \\ & = (0.3021, 0.7496) \end{aligned}$$

Example

One of the quantities calculated by some modern vehicles is fuel efficiency (or gas mileage), usually expressed in miles per gallon (mpg). For one vehicle equipped this way, the mpg were recored each time the gas tank was filled, and the computer was then reset. Here are the mpg values for a random sample of 20 of these records:

41.5 50.7 36.6 37.3 34.2 45.0 48.0 43.2 47.7 42.2
43.2 44.6 48.4 46.4 46.8 39.2 37.3 43.5 44.3 43.3

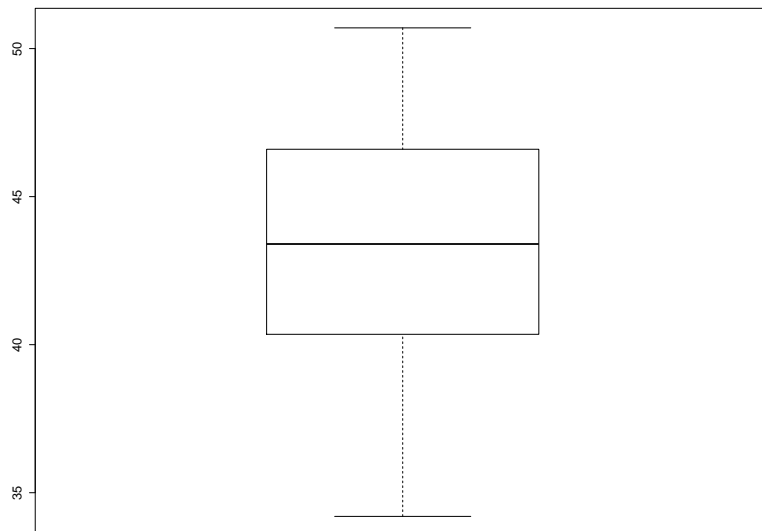
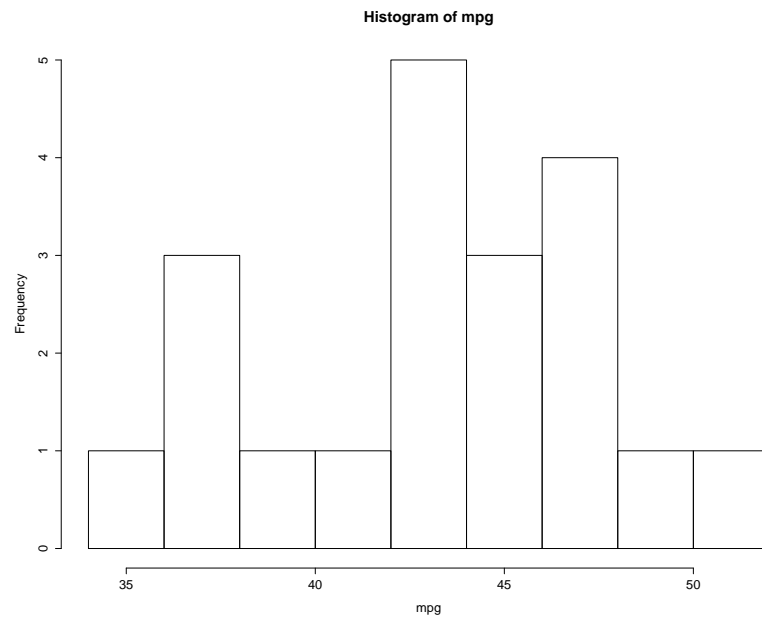
(a) Think: before we work on this problem, do you think we should be using one-sample z CI, or one-sample t CI?

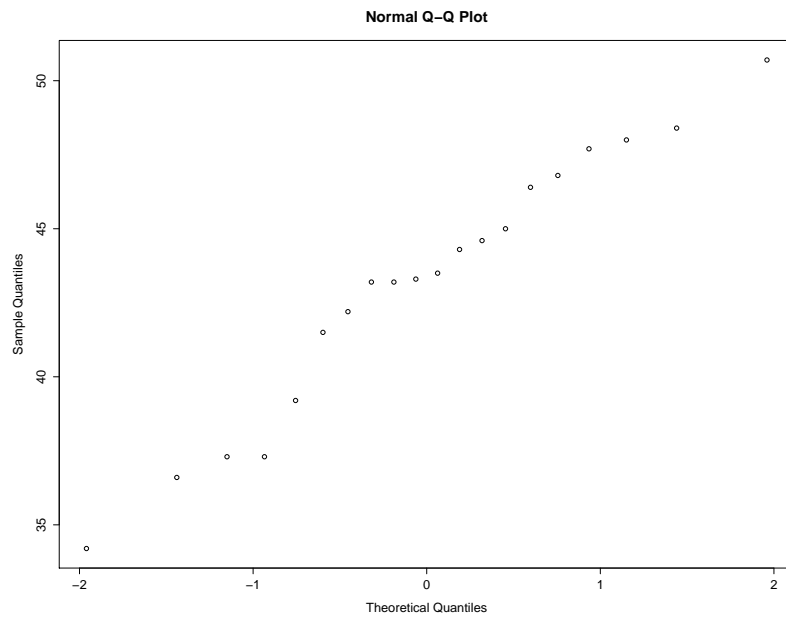
(b) By looking at the following displays, is it appropriate to analyze these data using methods based on Normal distribution? Why or why not?

R output:

```
>summary(mpg)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
34.20	40.92	43.40	43.17	46.50	50.70





(c) Find the mean, standard deviation, standard error and margin of error for 95% confidence.

(d) Report the 95% confidence interval for μ , the mean mpg for this vehicle based on these data.


(c)

$$\text{mean} = 43.17$$

$$\begin{aligned}\text{standard dev} &= \sqrt{19.49168} \\ &= 4.415\end{aligned}$$

$$\begin{aligned}\text{standard error} &= \frac{s}{\sqrt{n}} \\ &= \frac{4.415}{\sqrt{20}} \\ &= 0.987\end{aligned}$$

$$\begin{aligned}\text{margin of error} &= t \cdot \frac{s}{\sqrt{n}} = 2.093 \times 0.987 \\ &= 2.066\end{aligned}$$

 $t_{0.025,19}$

(d) 95% CI for μ :

$$\begin{aligned}\bar{x} \pm t \cdot \frac{s}{\sqrt{n}} \\ 43.17 \pm 2.066 = (41.10, 45.24)\end{aligned}$$

Table below is very useful in exams to help you decide whether to use z or t

* consider a small sample of an unknown population, we need to assume the population is normally distributed in order to calculate the confidence interval of its mean (or conduct a hypothesis test on its mean)

* use the t-distribution for a small sample drawn from a normal population with unknown variance. use the normal-distribution for all other cases (where the population variance is known and the population is normally distributed).

Distribution of population	population variance	sample size	use t or z?
Normal	known	Any size	z
Normal	unknown (estimate σ^2 by s^2)	Large	z
		Small	t
Unknown. For large samples,	known	Large	z
we can assume a normal distribution by central limit theorem	unknown. Estimate σ^2 by s^2	Large	z