Forum 4
Inference: Estimating Mean(s)
Data Skills for Scientists

2025-09-24

Inference

The 3 topics in the second module, 'Inference' are:

- Estimating Mean(s);
- Testing Mean(s); and
- Testing Median(s) and Categories.

In this forum we will cover the first topic, Estimating Mean(s).

Key topics for Estimating Mean(s)

- Confidence intervals (known population standard deviation)
- Sample size calculations for population mean using the margin of error (known population standard deviation)
- Confidence intervals (unknown population standard deviation)
- Introduction to inference and hypothesis testing

Confidence intervals (known population standard deviation)

The problem

Assume:

- we have a population that is Normally distributed, or the sample size n > 30.
- we know the population standard deviation σ, but we do not know the population mean μ.
- we have a simple random sample of size n.

How can we estimate the value of μ?

Example (Tablets)

Consider the case of estimating the population mean amount of active ingredient in manufactured tablets. You know that the population is normally distributed and the population standard deviation is 0.5mg. You have taken a simple random sample of 10 tablets and obtained the following mg of active ingredient:

29.07 29.32 29.95 30.37 29.96 28.91 28.53 30.55 29.61 30.09

The sample mean is 29.636 mg, and so we estimate the population mean active ingredient in the tablets to be 29.636 mg.

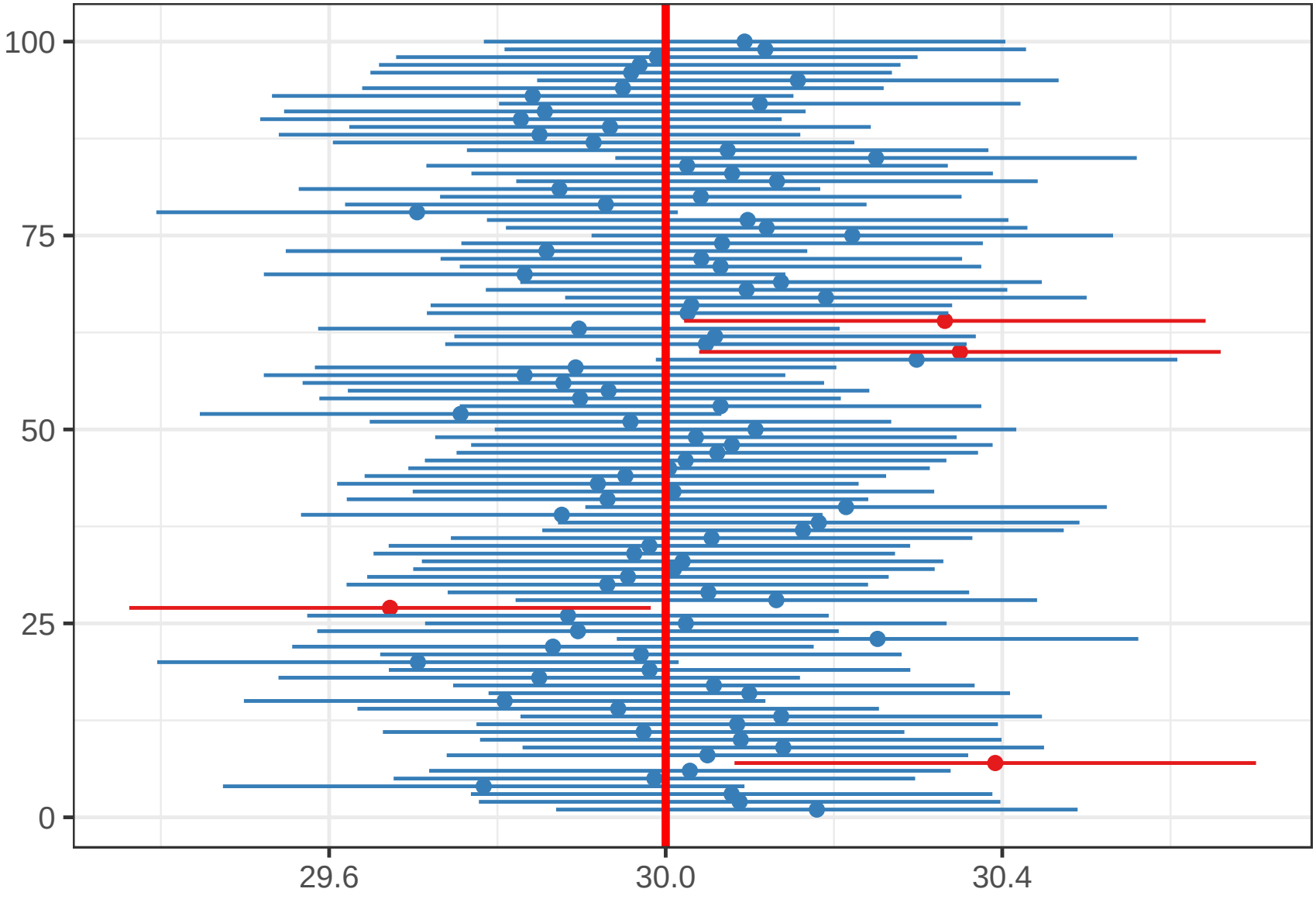
Confidence intervals

If you would like a range for the population mean rather than a point estimate, then you use a confidence interval.

Confidence intervals

A confidence interval will give a range of values for the population mean that we are confident about to a level $C\%$, usually 95%.

What do we mean by 95% confident?



Applet

Confidence interval Applet:

https:
//digitalfirst.bfwpub.com/stats_applet/stats_applet_4_ci.html

Confidence interval

All confidence intervals we construct will have a similar form:

- estimate $\pm$ critical value $\times$ standard error.

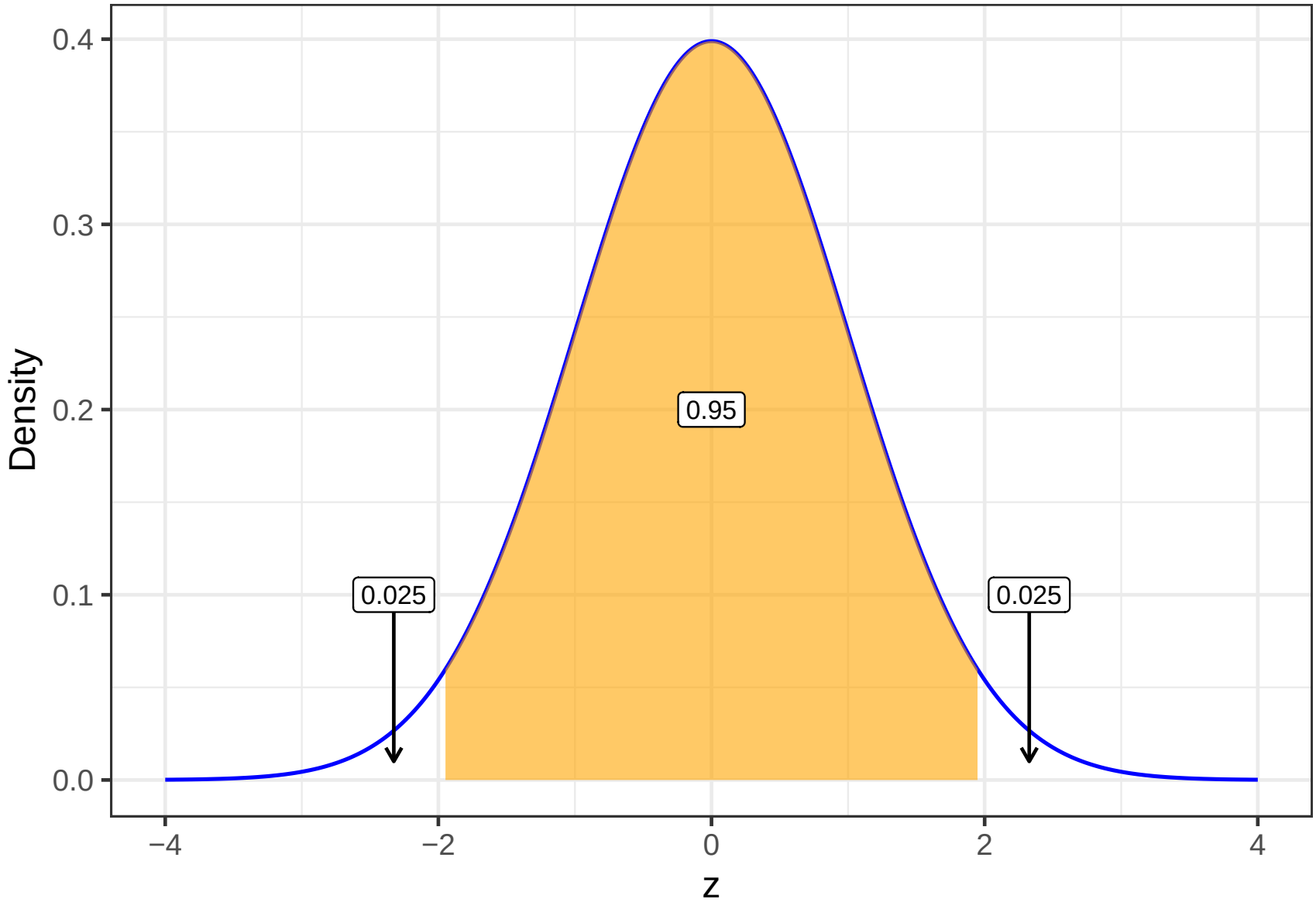
Confidence interval for population mean with known population standard deviation

In the case of a population mean for a Normal distribution with a known population standard deviation, we have

- **estimate:** the sample mean.
- **critical value:** we will get this from a Normal distribution and denote it as z^*.
- **standard error:** this is the standard deviation of the sample mean $\sigma/\sqrt{n}$.

How to calculate the critical value z^*

Consider confidence level 95%



How to calculate the critical value z^*

For a specified confidence level, e.g. 95%, we can compute the critical value z^* using R.

```
confidence_level<-0.95
alpha<-1-confidence_level
qnorm(1-alpha/2,mean=0,sd=1)
```

```
[1] 1.959964
```

Example

Calculate the 95% confidence interval for the tablet example.

Remember we have

- Sample mean, $\bar{x} = 29.636$ mg.
- Population standard deviation, $\sigma = 0.5$ mg.
- Sample size, $n = 10$.

Example (continued)

We can use R to compute the confidence interval.

```
sample_mean <-  29.636 # sample mean
population_sd <- 0.5 # population standard deviation
n <- 10 # sample size
confidence_level <- 0.95
alpha <- 1-confidence_level
zcrit <- qnorm(1-alpha/2,mean=0,sd=1)
lower <- sample_mean - zcrit*population_sd/sqrt(n)
upper <- sample_mean + zcrit*population_sd/sqrt(n)
```

Interpretation of confidence interval

We are <CI level> confident that the true <parameter> of <population> lies between <lower> and <upper> <units>.

What is the interpretation of the 95% CI in this case?

Summary - Confidence interval with known population standard deviation

- The problem - For a population that is normally distributed, or with sample size n > 30, assume that we ***know*** the population standard deviation σ, but we do not know the population mean μ. How can we estimate the value of μ?
- If we want a single point estimate of the population mean μ, we can take a simple random sample from the population and then use the sample mean $\bar{x}$ to estimate the population mean.
- A confidence interval will give a range of values for the population mean that we are confident about to a level $C\%$, usually 95%.
- The formula for calculating the C% confidence interval for the population mean (known population standard deviation) is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

Sample size calculations for population mean using the margin of error (known population standard deviation)

Interval width

Consider the *symmetric* interval centred at 5, with interval width (either side) of 1.

The interval would be

$$5 \pm 1 = (5 - 1, 5 + 1) = (4, 6).$$

Any value in this interval is *no more than* 1 away from 5, *i.e.* has a **margin of error** of 1.

Interval width

1. What is the margin of error for the symmetric interval $(5, 15)$?
2. What is the margin of error for the interval 6 ± 3?
3. What is the margin of error for the interval $\mu \pm m$?

Recall

The formula for finding the confidence interval for a population mean with known population standard deviation is

$$\bar{x} \pm \left(z^* \times \frac{\sigma}{\sqrt{n}} \right)$$

This interval is centered at $\bar{x}$ and has margin of error $m = z^* \times \frac{\sigma}{\sqrt{n}}$.

Note that as n gets large, m gets smaller and we are more 'certain' about the true value of μ.

Margin of error

The **margin of error** is a numerical measure of the spread of a sampling distribution.

We denote the margin of error m.

Before starting an experiment, we may state that we wish to have some *maximum* margin of error.

Sample size calculations

We can have a *large* confidence level and a *small* margin of error by choosing an appropriate sample size n.

As an example we focus on a confidence interval assuming the population standard deviation is known, although the method is generalisable to other tests.

Calculating the margin of error

Given z^*, σ and m (a desired margin of error), we can solve for n.

If we would like,

$$m \leq \frac{z^*\sigma}{\sqrt{n}},$$

then we require,

$$n \geq \left(\frac{z^*\sigma}{m}\right)^2.$$

Example

Let μ be the true population mean hours of television watched by university students. Assume we know that the population standard deviation σ is 17.5 hours.

We would like to calculate the sample size n such that we have 95% confidence level and margin of error *less than* 5 hours.

For this problem: $\sigma = 17.5$, $z^* = 1.96$ and we need $m \leq 5$.

Hence, we require,

$$n \geq \left(\frac{1.96 \times 17.5}{5} \right)^2 = 47.05.$$

which rounded up gives us $n = 48$.

Confidence intervals (unknown population standard deviation)

The problem

Assume:

- we have a population that is Normally distributed, or the sample size n > 30.
- we ***do not know*** the population standard deviation σ ***and*** we do not know the population mean μ.
- we have a simple random sample.

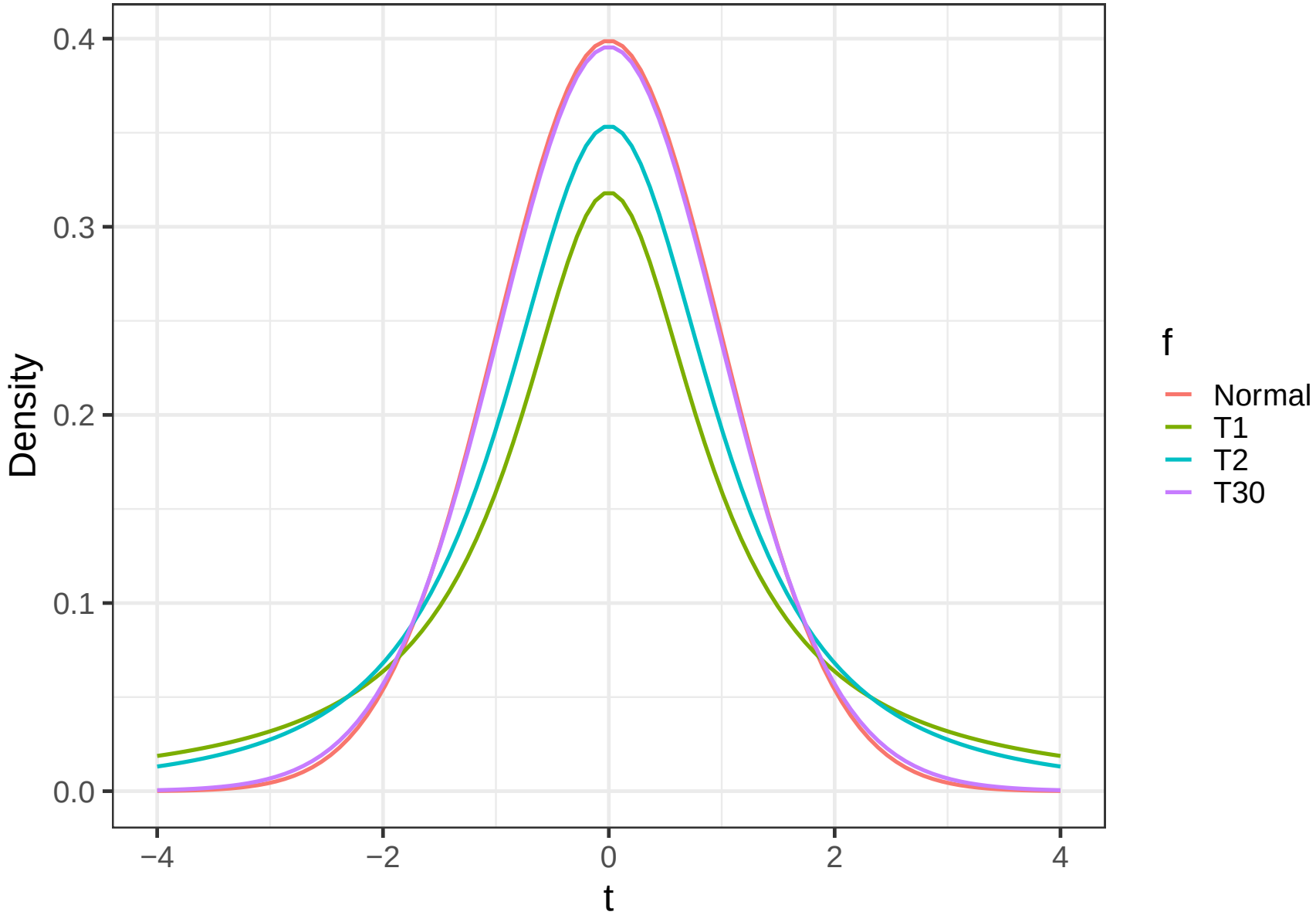
How can we estimate the value of μ?

This is the situation we will typically face in practice.

t-distribution

- We use the ***sample standard deviation*** in place of the population standard deviation.
- To account for the additional uncertainty (as we now estimate the sample mean and sample standard deviation), we use a t-distribution.
- There is a different t-distribution for each sample size, we identify each distribution by its ***degrees of freedom (df)***.
- For a sample size of n we use a t-distribution with degrees of freedom n-1.

t-distribution examples



Confidence interval for population mean with known population standard deviation

Recall that all confidence intervals we construct will have a similar form:

- estimate $\pm$ critical value $\times$ standard error.

In the case of a population mean for a Normal distribution with an ***unknown*** population standard deviation, we have

- **estimate:** the sample mean.
- **critical value:** we will get this from a t distribution and denote it as t^*.
- **estimated standard error:** this is the estimated standard deviation of the sample mean $s/\sqrt{n}$, where s is the sample standard deviation.

Confidence interval for population mean with known population standard deviation

The formula for calculating the C% confidence interval for the population mean with an unknown population standard deviation is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}.$$

How to calculate the critical value t^*

For a specified confidence level, e.g. 95%, we can compute the critical value t^* using R.

```
confidence_level<-0.95
alpha<-1-confidence_level
n <- 10 # number of samples
qt(1-alpha/2,df=n-1)
```

```
[1] 2.262157
```

Example (Tablets)

Consider the case of estimating the population mean amount of active ingredient in manufactured tablets. ***Now assume that you know the population is normally distributed but you do not know the population standard deviation***. You have taken a random sample of 10 tablets and obtained the following mg of active ingredient:

29.07 29.32 29.95 30.37 29.96 28.91 28.53 30.55 29.61 30.09

The sample mean is 29.636 mg, and so we estimate the population mean active ingredient in the tablets to be 29.636 mg.

The sample standard deviation is 0.663 mg, and so we estimate the population standard deviation of active ingredient in the tablets to be 0.663 mg.

Example

Calculate the 95% confidence interval for the tablet example.

Remember we have

- Sample mean, $\bar{x} = 29.636$ mg.
- Sample standard deviation, $s = 0.663$ mg.
- Sample size, $n = 10$.

Example

The formula for calculating the C% confidence interval for the population mean (unknown population standard deviation) is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}.$$

We can use R to compute the confidence interval.

```
sample_mean <-  mean(drug)
sample_sd <-sd(drug)
n <- 10 # number of samples
confidence_level <- 0.95
alpha <- 1-confidence_level
tcrit <- qt(1-alpha/2,df=n-1)
lower <- sample_mean - tcrit*sample_sd/sqrt(n)
upper <- sample_mean + tcrit*sample_sd/sqrt(n)
```

Example

Alternatively, we can use the t.test function to compute confidence intervals.

```
drug <- c(29.07,29.32,29.95,30.37,29.96,
28.91,28.53,30.55,29.61,30.09)
t.test(drug)
```

```
    One Sample t-test

data:  drug
t = 141.37, df = 9, p-value = 2.253e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 29.16178 30.11022
sample estimates:
mean of x
   29.636
```

Interpretation of confidence interval

We are <CI level> confident that the true <parameter> of <population> lies between <lower> and <upper> <units>.

What is the interpretation of the 95% CI in this case?

Summary - Confidence interval with known population standard deviation

- The problem - For a population that is normally distributed, or with sample size n > 30, assume that we ***do not know*** the population standard deviation σ ***and*** we do not know the population mean μ. How can we estimate the value of μ?
- Point estimate. If we want a single point estimate of the population mean μ, we can take a simple random sample from the population and then use the sample mean $\bar{x}$ to estimate the population mean.
- Confidence intervals. A confidence interval will give a range of values for the population mean that we are confident about to a level $C\%$, usually 95%.
  - We can use the t.test function to compute the confidence interval. This function uses the t-distribution that accounts for the additional uncertainty.

Introduction to inference

Inference

The process of drawing conclusions about a population on the basis of data collected on a sample from that population.

Hemple's Paradox

Suppose we aim to show

All crows are black

Every time we see an example of a crow that is black, evidence mounts.

Hemple's Paradox

Suppose we aim to show

All crows are black

Every time we see an example of a crow that is black, evidence mounts.

But if we see a red pencil - does this provide evidence that all crows are black?

Hemple's paradox

All crows are black

is logically equivalent to

Not black, not a crow

The pencil is not black and not a crow - so what is truly meant by *evidence* is not entirely clear...

Statistical Questions

Are koalas better than possums at living in Adelaide?

What does this mean?

- Does it refer to the urban suburbs, the Adelaide Hills, or parklands?
- Does it refer to areas with native vegetation or those dominated by human development?
- What aspects of living does the question refer to? Finding food? Avoiding traffic and predators? Coping with heatwaves?

Can we answer this question?

Statistical Questions

> *Do koalas in the Adelaide Hills have higher survival and reproduction rates than common brushtail possums in the same region, based on data collected by local ecologists between 2015 and 2025?*

Statistical questions must relate to a well defined population.

Hypothesis Testing

Suppose we have the hypothesis that

All crows are black

Every time we see an example of a crow that is black, evidence mounts. If we only ever observe black crows, can we prove our hypothesis?

Hypothesis testing framework

Four steps of tests of significance

Tests of significance - four steps:

- State the null and alternative **hypotheses**.
- Calculate the value of the **test statistic**.
- Find the **p-value** for the observed data.
- State a **conclusion** (typically against a **significance level**).

Example

	Serum Cholesterol				Mean
Behaviour Type A					
233	291	312	250	246	
197	268	224	239	239	
254	276	234	181	248	
252	202	218	212	325	245.05
Behaviour Type B					
220	185	263	246	224	
212	188	250	148	169	
226	175	242	252	153	
183	137	202	194	213	204.10

Is there a genuine difference in the serum cholesterol levels of the two behavioural types?

Example

1) **Hypothesis**

H_0: People with type A behaviour and people with type B behaviour have the same mean serum cholesterol.

H_1: People with type A behaviour and people with type B behaviour have different mean serum cholesterol.

2) **Test Statistic** We will learn how to do this later.

3) **P-Value** A standard test yields the p-value $p = 0.0011$. So if H_0 were true we would expect to see a difference in means as large or larger than 40.95 only 11 times in 10,000.

4) **Conclusion** Either H_0 is false, or we have been very unlucky in selecting our sample.

No Evidence

When the p value is large, the data are not inconsistent with H_0 – if H_0 were true, it would not be unlikely to see data like this.

This means there is no evidence against H_0, it does not provide any evidence that H_0 is true.

Strong Evidence

When the p value is very small, the data are very inconsistent with H_0. It is unlikely we would see data like this if H_0 were true – unlikely, not impossible.

If $p = 0.01$, it means that 1 in 100 times we would see data as extreme as this when H_0 is true.

Either we have seen an outcome that is unlikely, or H_0 is false.

Errors

There are two types of error we can commit

Type I Rejecting H_0 when it is true

Type II Failing to rejecting H_0 when it is false

Power The power of a study is probability of correctly rejecting H_0 when it is false.

Importance

Suppose in the cholesterol example, every type A person had a serum cholesterol level of exactly 245, and every type B a level of exactly 244.9

Then an hypothesis test would show very strong evidence the mean serum cholesterol differs between the two behaviours.

But a difference of 0.1 in this context is negligible.

The p-value is only an indicator of whether the difference is real, it does not imply that the difference is important.

Further reading/tools

- Textbook:
  - 6.1 Estimating with confidence
  - 6.2 Tests of significance
  - 6.3 Use and abuse of tests
  - 6.4 Power and inference as a decision (optional)
- Normal distribution Applet: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_7_norm.html
- Confidence interval Applet: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_4_ci.html