University of BRISTOL

WEEK 4

Confidence Intervals

Felipe Campelo

bristol.ac.uk

In this lecture…

In this lecture, we'll go over the main concepts related to quantifying our uncertainty when estimating parameters. More specifically, we will:

- Explore the concept of a **confidence interval** and its interpretation in terms of sampling variability.
- Learn how to construct and interpret CIs for a population mean, under different assumptions (normal distribution with σ^2 known and with σ^2 unknown, large-sample approximate intervals).
- See how sample size, variability, and confidence level affect the width / precision of the interval.

Motivation

Last week we discussed concepts related to point estimation of parameters – that is, given a sample, how to get the best possible *estimate* of the true value of a populational parameter. However, given that point estimates are subject to sampling variability, we know that they will almost always be a bit "*off-target*".

We also saw how we can use *standard errors* to quantify the uncertainty associated with our estimates. However, although standard errors provide a sense of how much an estimate would vary *across repeated samples*, it is an incomplete quantification of uncertainty.

What is commonly needed to support decision-making is a range of plausible values for the true population parameter, which captures the real value of the parameter with *a known level of confidence.*

Definition

Confidence intervals (CIs) quantify the degree of uncertainty associated with the *estimation of population parameters* such as the mean, variance, proportion, difference of means etc..

A CI can be defined as *an interval that is expected to contain the true value of a given population parameter with a confidence level of* $100(1-\alpha)\%$;

A useful way to think about confidence intervals in terms of confidence *in the method*: The **method** used to derive the interval has a hit rate of $100(1-\alpha)\%$ - i.e., **the interval generated has a $100(1-\alpha)\%$ chance of having captured the true value of the population parameter**.

Intuition

True parameter value (*fixed; unknown*)

Interval (width is determined by factors such as sample size, desired confidence level and variance of the data-generating process.

Data-generating process (*with its own properties, parameter values, variance etc*).

Formalisation

A **confidence interval** estimate for a given parameter θ is an interval of the form

$$P(L \leq \theta \leq U) = 1 - \alpha$$

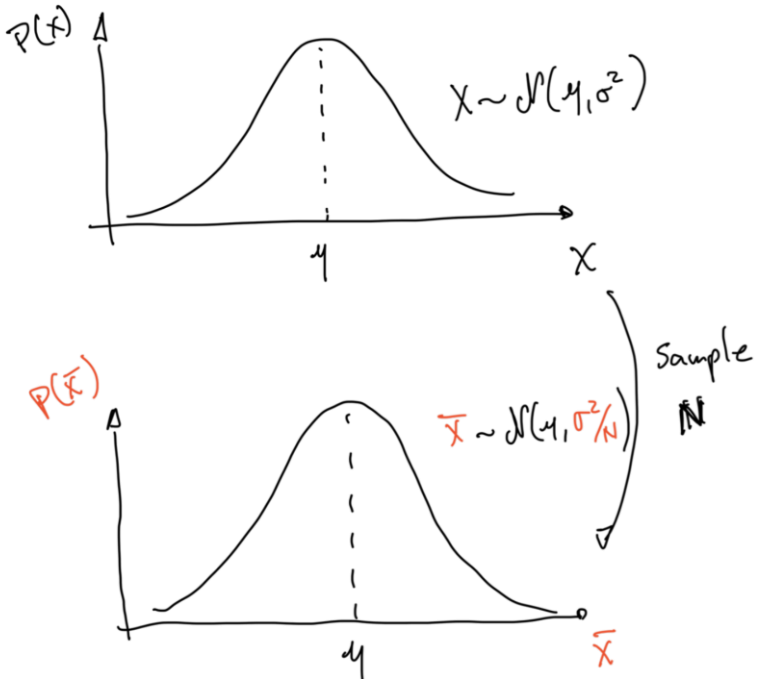
where the end-points *L* and U are computed from the sample data.

Given a sample, there is a probability that the CI calculated from it will contain the true value of θ.

The bounds L and U are called the lower- and upper-confidence limits.

The quantity $1 - \alpha$ is called the confidence level (or confidence coefficient) of the interval.

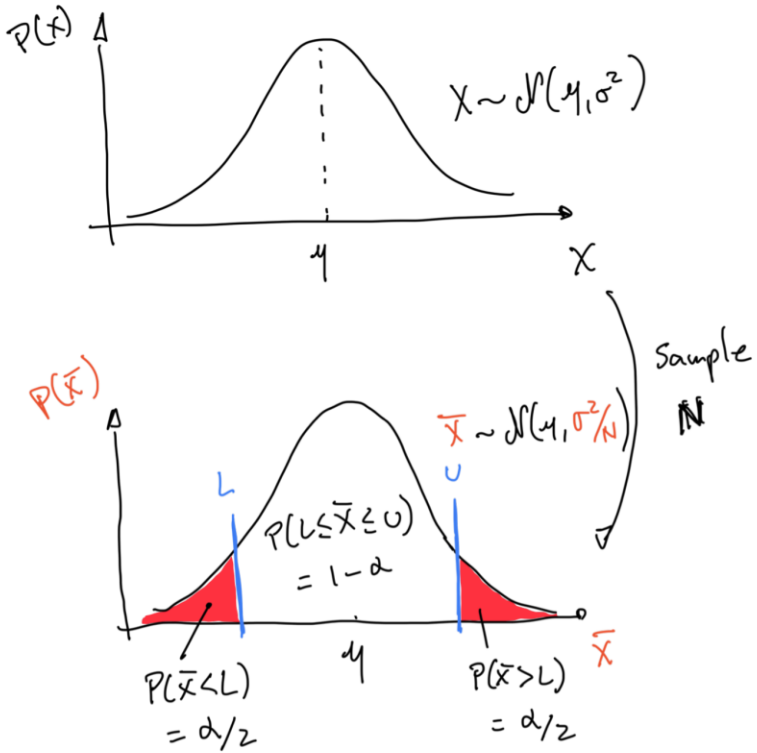
CI on the mean of a normal variable, known σ^2



Suppose that you have a data-generating process that's distributed according to a normal distribution with *known variance* σ^2 and *unknown mean* μ, and we get a sample of N observations, $X_1, \ldots, X_N$.

We know that the sample mean $\bar{X}$ will also be normally distributed, with mean μ and variance σ^2/N.

CI on the mean of a normal variable, known σ^2



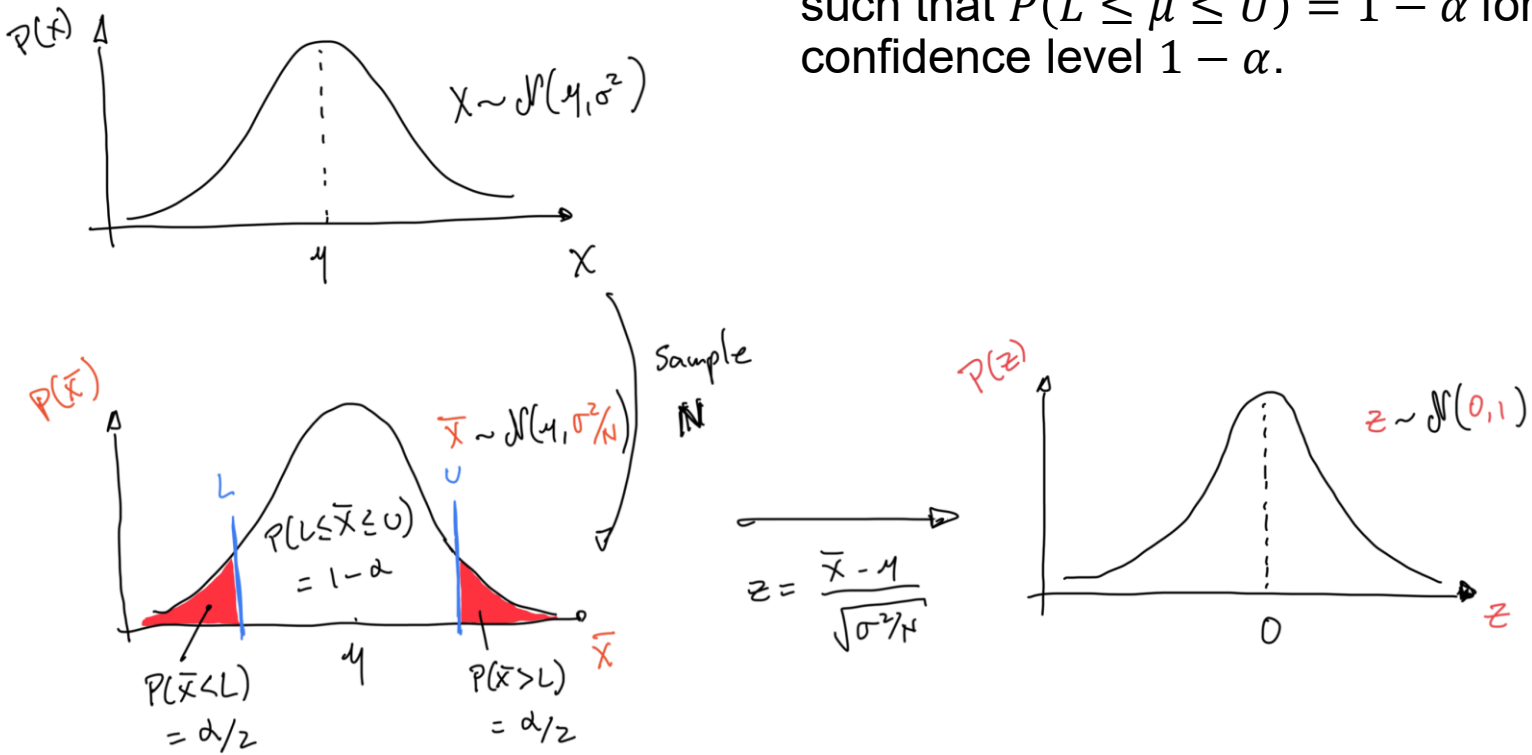
Suppose that you have a data-generating process that's distributed according to a normal distribution with *known variance* σ^2 and *unknown mean* μ, and we get a sample of N observations, $X_1, \ldots, X_N$.

We know that the sample mean $\bar{X}$ will also be normally distributed, with mean μ and variance σ^2/N.

We want to find out what are the values of L and U such that $P(L \leq \mu \leq U) = 1 - \alpha$ for any desired confidence level $1 - \alpha$.

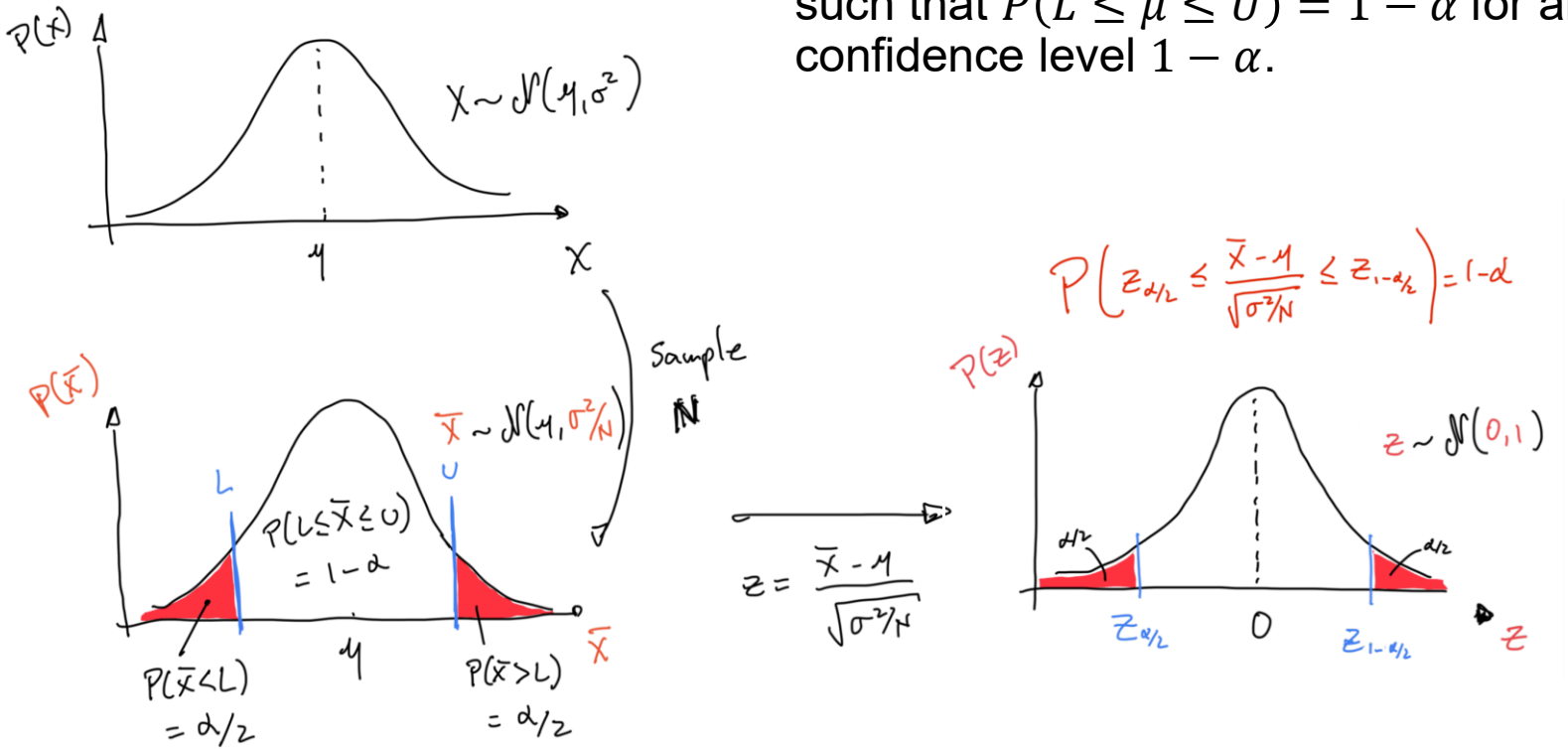
CI on the mean of a normal variable, known σ^2

We want to find out what are the values of L and U such that $P(L \leq \mu \leq U) = 1 - \alpha$ for any desired confidence level $1 - \alpha$.

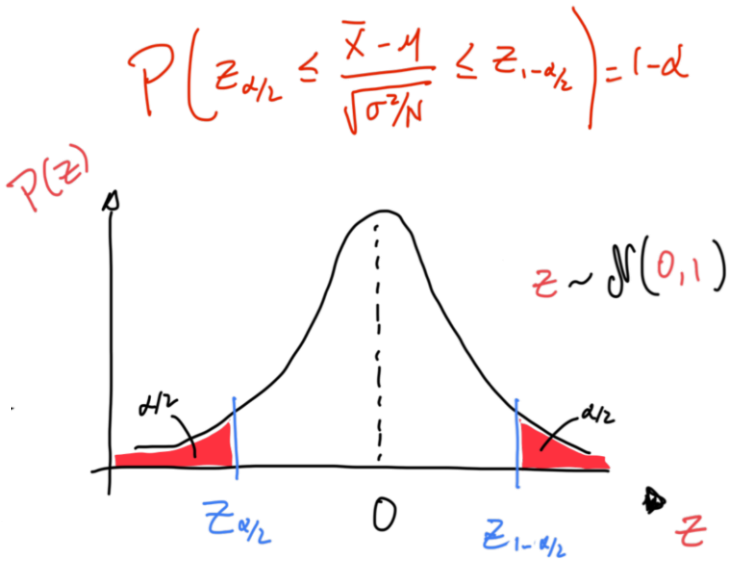


CI on the mean of a normal variable, known σ^2

We want to find out what are the values of L and U such that $P(L \leq \mu \leq U) = 1 - \alpha$ for any desired confidence level $1 - \alpha$.



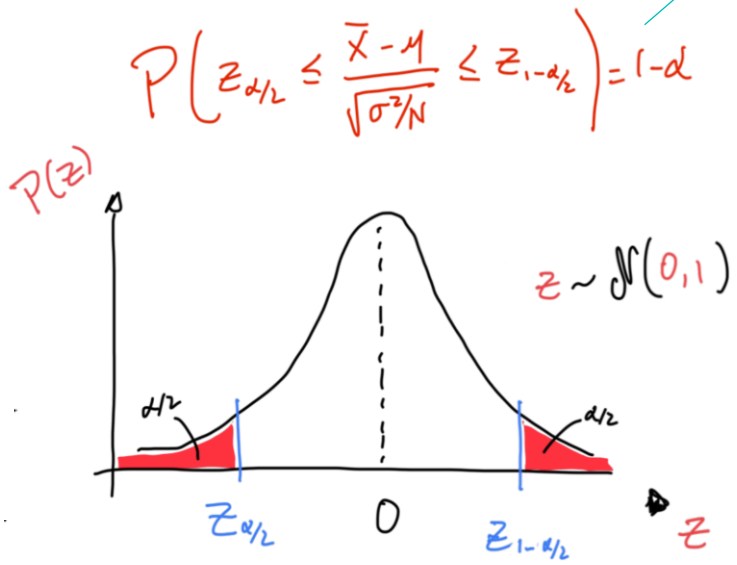
CI on the mean of a normal variable, known σ^2



Note that there are infinitely many normal distributions, but only one standard normal, $z \sim \mathcal{N}(0,1)$, which has known, well-defined **quantiles**.

Cumulative probability (p)	Quantile (z_p)
0.005	$z_{0.005} = -2.576$
0.025	$z_{0.025} = -1.960$
0.05	$z_{0.05} = -1.645$
0.1	$z_{0.01} = -1.282$
0.5	$z_{0.5} = 0$
(1-0.1) = 0.9	$z_{0.9} = 1.282$
(1-0.05) = 0.95	$z_{0.95} = 1.645$
(1 - 0.25) = 0.975	$z_{0.975} = 1.96$
(1 - 0.005) = 0.995	$z_{0.995} = 2.576$

CI on the mean of a normal variable, known σ^2



We can manipulate this expression to get:

$$P(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}) = 1 - \alpha$$

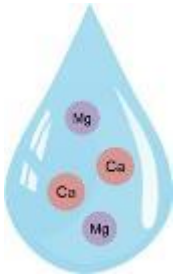
Since $z_{\alpha/2} = -z_{1-\alpha/2}$, we have that the interval

$$CI_{1-\alpha} = \left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$$

$$= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{N}} = \bar{X} + z_{\alpha/2}se_{\bar{X}}$$

captures the true mean with confidence $1 - \alpha$.

Example



You get data related to the daily concentration of Calcium in a local water supply. Your data represents the readings (in mg/litre) of 10 sensors, located in the city reservoir. Historical data provides a **known population standard deviation** $\sigma = 0.14$mg/l. There is also good evidence that the distribution of the readings is approximately normal.

On a given day, the 10 readings result in a sample mean of 0.37 mg/l. Based on this data, a $CI_{0.99}$ on the mean Ca concentration on that day can be calculated as:

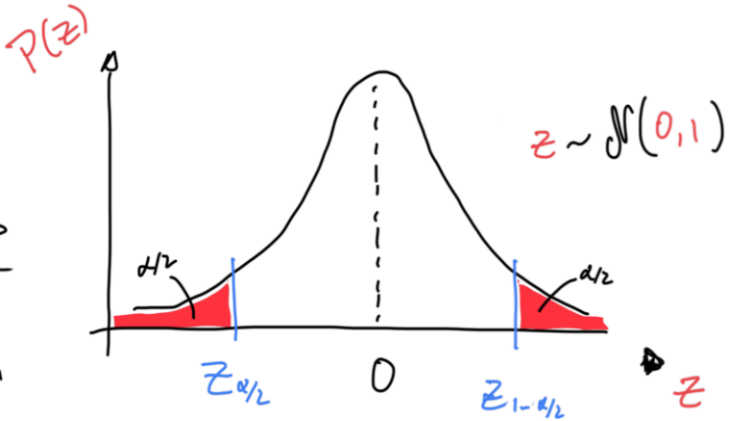
$$CI_{0.99} = 0.37 \pm z_{0.005}\frac{0.14}{\sqrt{10}}$$

$$= (0.255, 0.483)$$

```
> x
 [1] 0.270 0.464 0.539 0.434 0.248 0.320 0.366 0.451 0.373 0.227
> mean(x)
[1] 0.3692
> mean(x) + qnorm(c(0.005, 0.995)) * 0.14/sqrt(10)
[1] 0.2551632 0.4832368
>
```

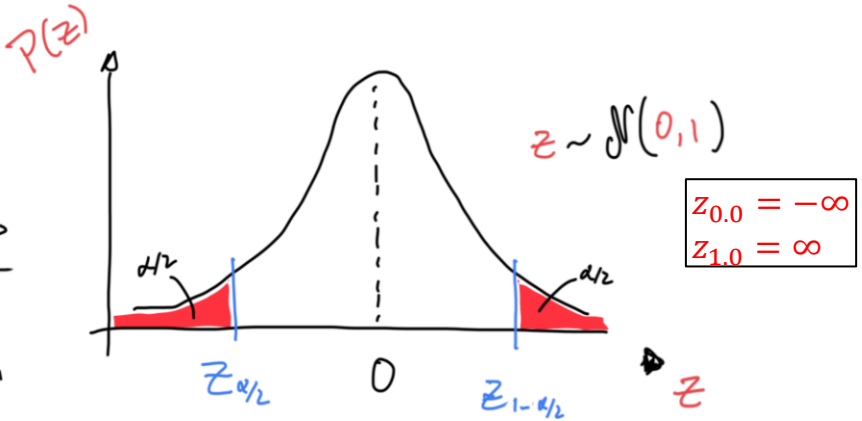
Confidence level and interval width

Notice that the confidence level is a user-defined parameter. Common choices are 0.95 (95%) or 0.99 (99%). Why don't people go for 100% confidence?



Confidence level and interval width

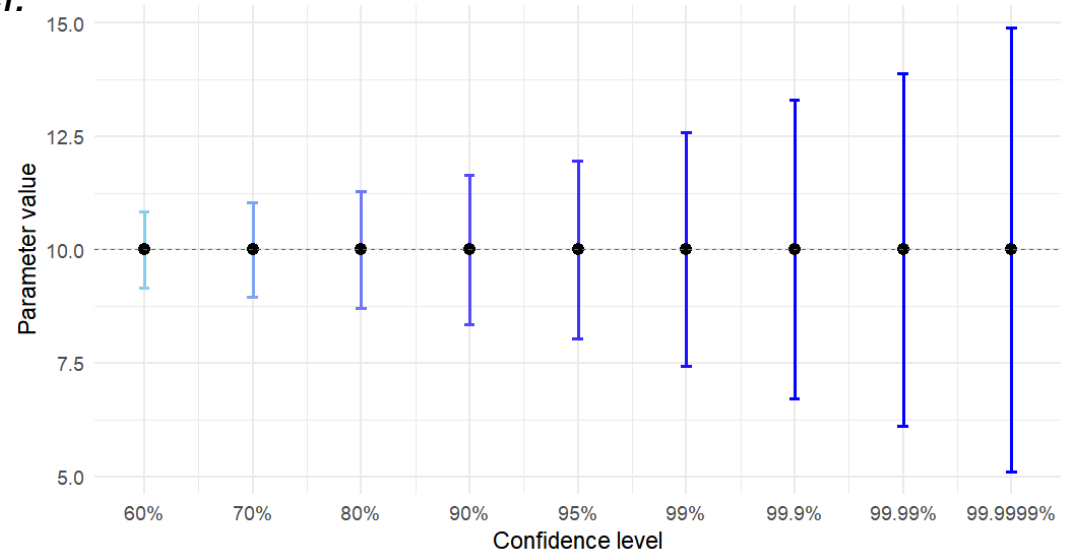
Notice that the confidence level is a user-defined parameter. Common choices are 0.95 (95%) or 0.99 (99%). Why don't people go for 100% confidence?



Confidence level and interval width

Notice that the confidence level is a user-defined parameter. Common choices are 0.95 (95%) or 0.99 (99%). Why don't people go for 100% confidence?

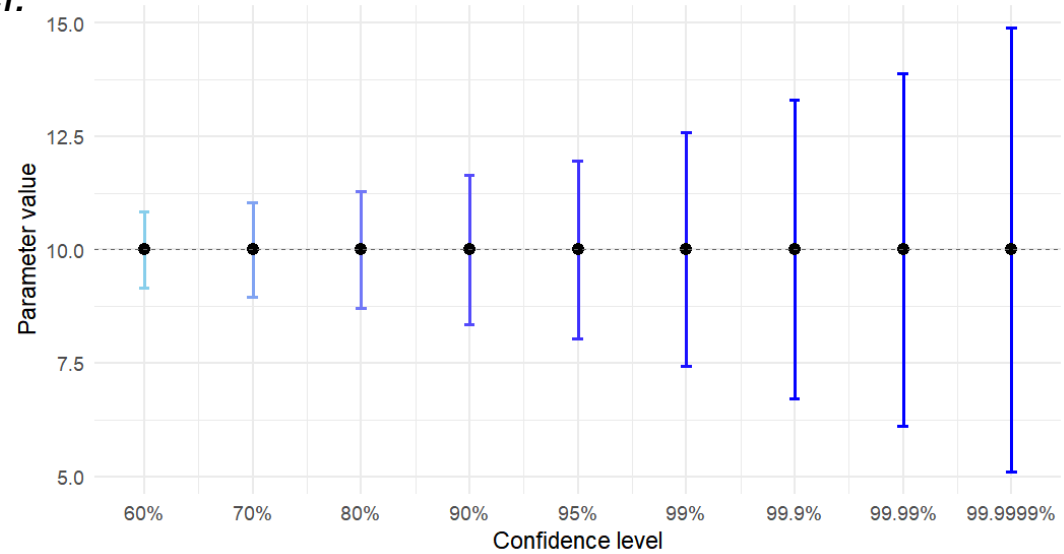
Given a sample, the width of the confidence interval increases as we increase the desired level of confidence - i.e., there is a trade-off between the *precision* of the interval and the *confidence level.*



Confidence level and interval width

Notice that the confidence level is a user-defined parameter. Common choices are 0.95 (95%) or 0.99 (99%). Why don't people go for 100% confidence?

Given a sample, the width of the confidence interval increases as we increase the desired level of confidence - i.e., there is a trade-off between the *precision* of the interval and the *confidence level.*



How can we set both the confidence level and the precision of an estimated interval?

Sample size calculation

Recall that the confidence interval can be expressed as $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$, which means that the interval half-length (sometimes called the “error” of the estimate) is given by:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

If we want a $CI_{1-\alpha}$ with a predefined half-length $E = e$, we can easily determine the necessary sample size as:

$$N = \left(\frac{z_{\alpha/2}\, \sigma}{e} \right)^2$$

Example

A health-tech startup wants to estimate the **average daily step count** of its app users with a **margin of error** of 500 steps at 95% confidence. From previous studies, the **standard deviation** is known to be approximately $s = 2000$ steps. How many users does it need to sample?

$$N = \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2 \rightarrow N = \left(\frac{z_{0.025} \times 2000}{500}\right)^2 \approx 62$$

To get these levels of precision and confidence, the company would need to sample at least 62 users.

(*note that data needs increase quadratically with reductions in the error*)

```
> sigma <- 2000
> e <- 500
> alpha <- 0.05
> (N <- (qnorm(0.025) * sigma/e)^2)
[1] 61.46334
> 
```

CI for the mean of a normal variable, unknown σ^2

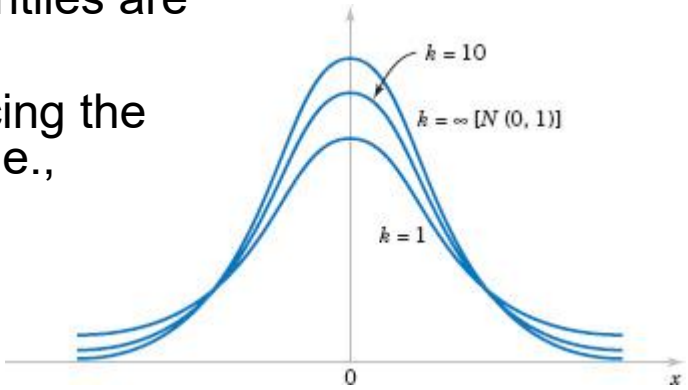
If the variance of the data-generating distribution is unknown, the standard error of the mean is estimated as $s/\sqrt{N}$ and the standardisation operation becomes:

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/N}} \sim t^{(k)}$$

which is distributed according to a t *distribution* with $\mathrm{k} = N - 1$ *degrees-of-freedom.* The t distribution is a bell-shaped distribution like the normal, but with *heavier tails* – meaning that the quantiles are located farther away from the centre.

A confidence interval can be calculated simply replacing the normal quantiles by their corresponding t quantiles, i.e.,

$$CI_{1-\alpha} = \left(\bar{X} + t_{\alpha/2}^{(N-1)} \frac{\sigma}{\sqrt{N}}, \bar{X} - t_{\alpha/2}^{(N-1)} \frac{\sigma}{\sqrt{N}} \right)$$



CI for the mean of a normal variable, unknown σ^2

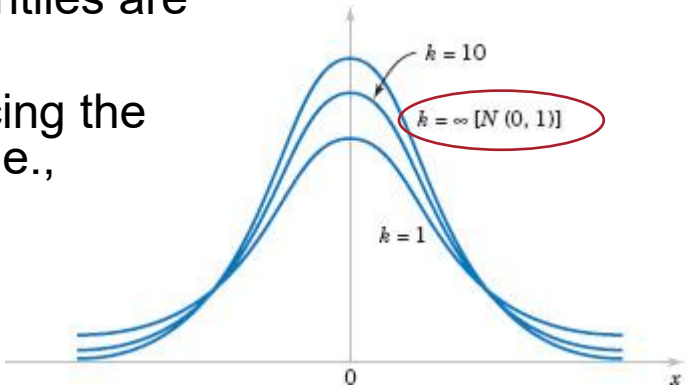
If the variance of the data-generating distribution is unknown, the standard error of the mean is estimated as $s/\sqrt{N}$ and the standardisation operation becomes:

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/N}} \sim t^{(k)}$$

which is distributed according to a t *distribution* with $\mathrm{k} = N - 1$ *degrees-of-freedom.* The t distribution is a bell-shaped distribution like the normal, but with *heavier tails* – meaning that the quantiles are located farther away from the centre.

A confidence interval can be calculated simply replacing the normal quantiles by their corresponding t quantiles, i.e.,

$$CI_{1-\alpha} = \left(\bar{X} + t_{\alpha/2}^{(N-1)} \frac{\sigma}{\sqrt{N}}, \bar{X} - t_{\alpha/2}^{(N-1)} \frac{\sigma}{\sqrt{N}} \right)$$

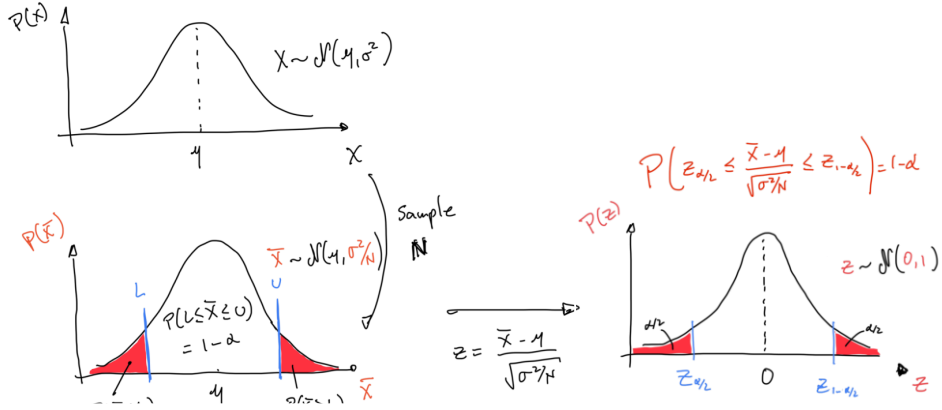


Large-sample CI for μ, any distribution.

So far, we have assumed that the data-generating distribution is normal.

However, the connection between the data-generating distribution and the confidence interval formula is mediated by the *sampling distribution of the means,* and the CLT tells us that, under i.i.d. sampling, the sampling distribution of the means converges to a normal distribution *regardless of the original data distribution*, as long as the mean and variance are defined and finite.

Also, for large N, the t distribution converges to the standard normal.



Large-sample CI for μ, any distribution.

So far, we have assumed that the data-generating distribution is normal.

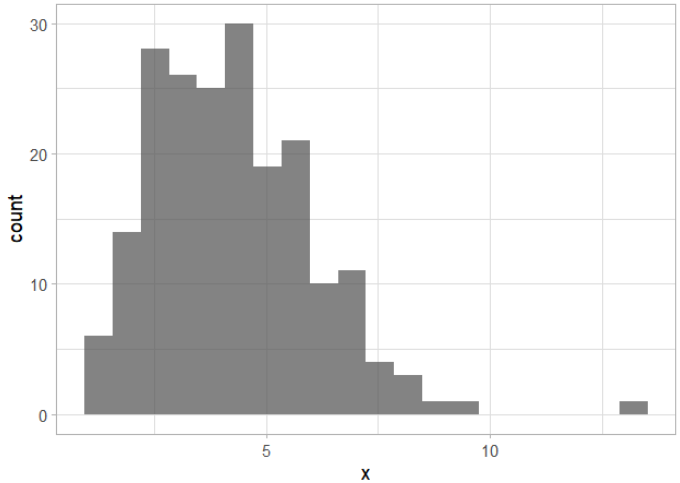
However, the connection between the data-generating distribution and the confidence interval formula is mediated by the *sampling distribution of the means,* and the CLT tells us that, under i.i.d. sampling, the sampling distribution of the means converges to a normal distribution *regardless of the original data distribution*, as long as the mean and variance are defined and finite.

Also, for large N, the t distribution converges to the standard normal.

This means that, for large N, the quantity $(\bar{X} - \mu)/(\sigma/\sqrt{N})$ has an approximate standard normal distribution and, consequently, we can use

$$CI_{1-\alpha} = \left(\bar{X} + z_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{X} - z_{\alpha/2}\frac{s}{\sqrt{N}}\right)$$

Example



An e-commerce platform measures the **delivery time (in days)** for 200 randomly chosen orders. The sample mean is $\bar{X} = 4.27$ days, and the sample standard deviation is $s = 1.79$ days. Delivery times are known to be skewed, but for $N = 200$ the sampling distribution of the means is highly likely to be normal.

A large-sample approximate $CI_{0.95}$ on the mean delivery time can be easily calculated as:

$$CI_{0.95} = 4.27 \pm z_{0.025} \frac{1.79}{\sqrt{200}}$$

$$= (4.03, 4.52).$$

```
> mean(x) + qnorm(c(0.025, 0.975)) * sd(x)/sqrt(length(x))
[1] 4.028988 4.524274
>
```

CIs for other parameters/distributions

If your data-generating process follows a known probability distribution, there are often specific formulas to calculate confidence intervals. For large sample sizes, these are often based on the approximate interval template:

$$CI_{1-\alpha} = \left(\widehat{\Theta} + z_{\alpha/2} se_{\widehat{\Theta}}, \widehat{\Theta} - z_{\alpha/2} se_{\widehat{\Theta}}\right).$$

For smaller sample sizes, there are distribution-specific formulas or procedures that can usually be easily found – e.g., the exact CI on the variance of a normal distribution,

$$CI_{1-\alpha} = \left(\frac{(N-1)s^2}{\chi^2_{\alpha/2,N-1}}, \frac{(N-1)s^2}{\chi^2_{1-\alpha/2,N-1}}\right).$$

CIs for other parameters/distributions

The more general case, in which the data-generating distribution cannot be easily described as a known / canonical random variable, requires *distribution-free methods.*

There are a few such methods based on sophisticated statistical results, but one of the most common (and most powerful) is called the *bootstrap.* This method is based on repeated resampling of the available sample and the distribution of an empirical estimate of the sampling distribution of any statistic of interest.

We will discuss the bootstrap in this week's lab session.

In this lecture we discussed...

- What confidence intervals represent and how they express uncertainty around sample estimates of parameters.
- How to derive and compute confidence intervals for a population mean under different scenarios.
- The influence of sample size, variability, and confidence level on interval width and precision.

Confidence intervals express estimation uncertainty as a range with a quantified degree of certainty - helping you make decisions based not only on your maximum likelihood value for the parameter, but on the range of plausible values (including quantified plausible best/wors cases).

Further reading

DC Montgomery, GC Runger, *Applied statistics and probability for engineers, 5th ed.*
[Chapters 8.1-8.5]

- *Read the chapter and try to solve the worked examples by yourself*.