Contents

1 Some properties of the normal distribution

The normal distribution is parameterised by two things: the distribution's **mean** and **variance**. Commonly we denote them with μ and σ^2.

Say we have a random variable that follows the normal distribution, say $X \sim N(\mu, \sigma^2)$. This variable has mean μ and variance σ^2, or equivalently, standard deviation σ.

We can shift the mean left or right (negatively or positively) by adding or subtracting a fixed constant to X. Consider $(X + c)$, which will have mean $\mu + c$. The variance remains unchanged.

The variance can be changed by multiplying a fixed constant to X. (cX) will have variance $c^2\sigma^2$. The mean is also scaled accordingly to $c\mu$.

So, to transform this variable $X \sim N(\mu, \sigma^2)$ into the standard normal distribution $Z \sim N(0, 1)$, we have to subtract the mean μ and divide by σ.

$$X \sim N(\mu, \sigma^2) \iff (X - \mu) \sim N(0, \sigma^2) \iff \frac{X - \mu}{\sigma} \sim N(0, 1)$$

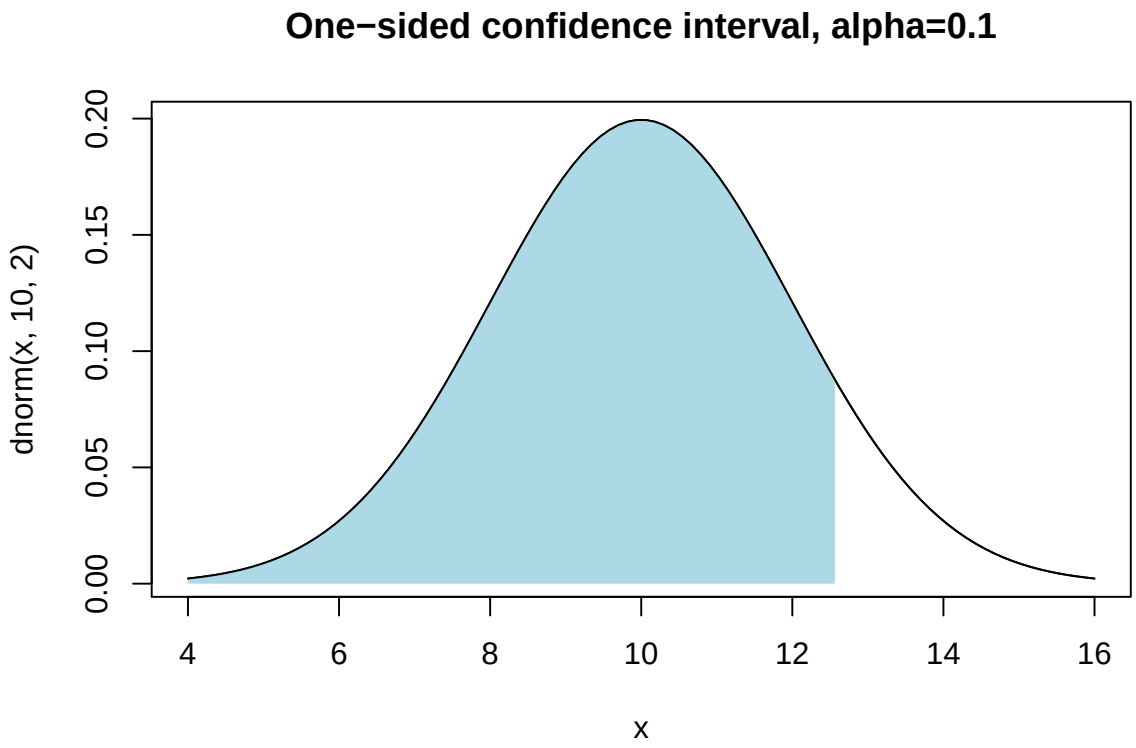
This actually corresponds with the properties of expectation and variance, i.e. $E(aX+b) = aE(X)+b$ and $Var(aX + b) = a^2Var(X)$.

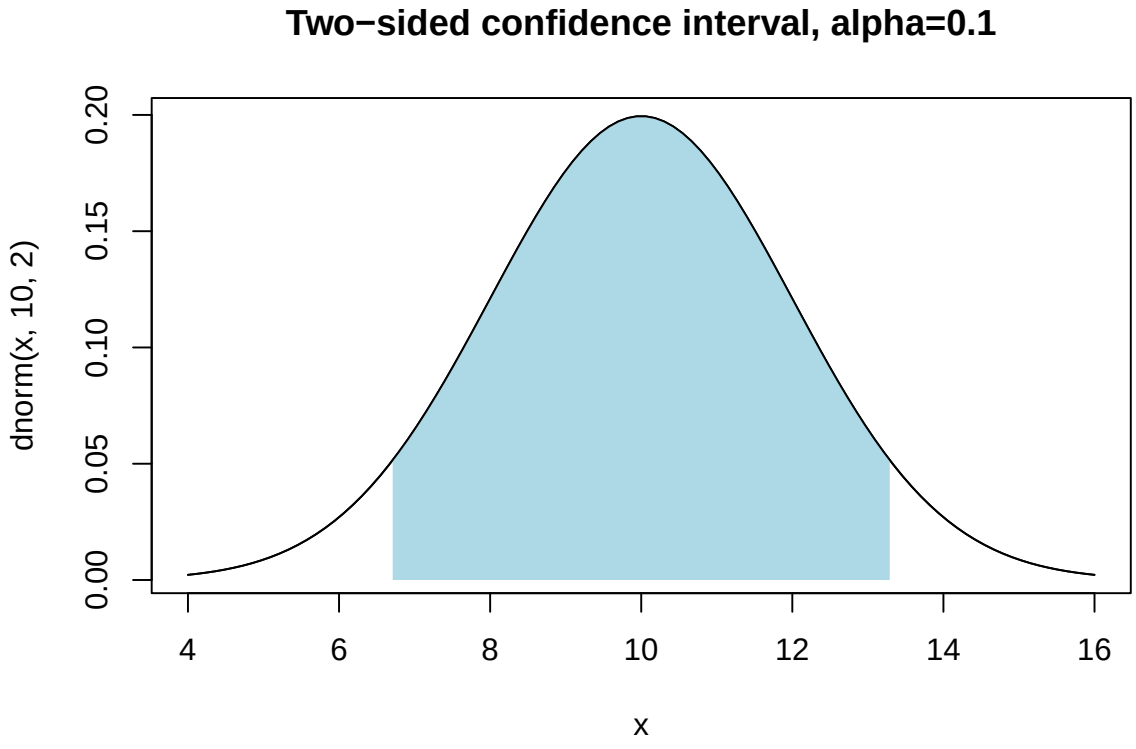
2 Confidence intervals

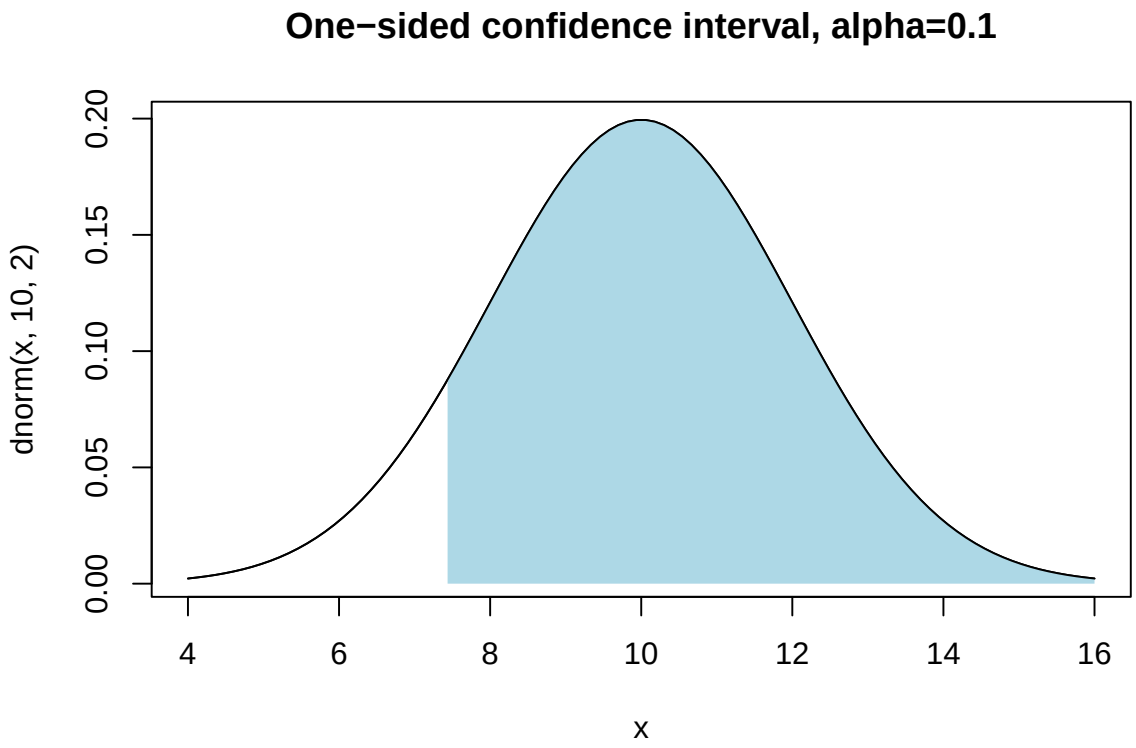
Confidence intervals are expressed with some value between 0 and 1. In the context of estimating the sample mean, for example, a 0.9-confidence interval represents an interval that has a 90% probability of containing the true population mean.

Here, what we really want is the population mean - but we since that is typically infeasible to calculate, we use the sample mean as an estimate. Typically, the sample mean has a nice distribution (normally distributed or t-distributed), thanks to the central limit theorem.

Say we find that our sample mean $\bar{X} \sim N(10, 4)$. To find a 90% confidence interval, we want to find an interval such that the area under the p.d.f. is 0.9. There are infinitely many - typically it is chosen to be one-sided or two-sided. All of the graphs below represent a valid 0.9-confidence interval - all of them have an area under curve of 0.9.


One−sided confidence interval, alpha=0.1
dnorm(x, 10, 2)
0.00
0.05
0.10
0.15
0.20
4
6
8
10
12
14
16
x


Two−sided confidence interval, alpha=0.1
dnorm(x, 10, 2)
0.00
0.05
0.10
0.15
0.20
4
6
8
10
12
14
16
x



These intervals are constructed with the quantile functions. Recall that the quantile function Q is the inverse c.d.f.

If I take the interval $(Q(0), Q(0.9))$, this gives exactly a (one-sided) 0.9-confidence interval. Likewise, the interval $(Q(0.05), Q(0.95))$ also gives a (two-sided) 0.9-confidence interval.

It is common to parameterise this confidence interval, usually with α as the significance level in hypothesis testing, thus forming $(1-\alpha)$-confidence intervals, which are $(Q(0), Q(1-\alpha))$, $(Q(\alpha), Q(1))$, and $(Q(\alpha/2), Q(1-\alpha/2))$.

For example, if $\alpha = 0.05$, we get 0.95-confidence intervals.

These confidence intervals can be manually derived using a z-table (or t-table) if we standardise the distributions to have mean 0 and variance 1.

Continuing on with the example that our sample mean $\bar{X} \sim N(10, 4)$, we instead consider the standardised sample mean $((\bar{X} - 10)/2) \sim N(0, 1)$ since it is a well-known distribution with pre-computed quantiles. We get the quantiles of interest from the z-table - for a 90% confidence interval, say ± 1.645.

$$\begin{aligned} 0.9 = P\left(-1.645 \leq \frac{\bar{X} - 10}{2} \leq 1.645\right) &= P(-3.29 \leq \bar{X} - 10 \leq 3.29) \\ &= P(6.71 \leq \bar{X} \leq 13.29) = 0.9 \end{aligned}$$

From which we can see a 0.9-confidence interval is $(6.71, 13.29)$, which lines up with the two-sided

confidence interval graphed above.

3 Confidence intervals for proportions

The confidence intervals for proportions may seem randomly defined but has an easy enough derivation, but requires some background knowledge.

3.1 Bernoulli distribution

The Bernoulli distribution can only take on the values 0 or 1, and has pdf $f_X(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$. It takes on value 1 with probability p and 0 with probability $1 - p$.

The expectation (mean) and variance can be easily calculated: $E(X) = p$ and $Var(X) = p(1-p)$.

So, suppose that we have $X_1, \ldots, X_n$ representing polls of an audience - 1 and 0 representing the responses (whatever they may be). We model them as identically and independently distributed (i.i.d.) Bernoulli variables. The sample mean $\bar{X}$ thus represents the proportion of respondents to the poll who responded positively (recorded as 1).

Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, with big enough n, we apply the central limit theorem, so we say $\bar{X}$ is normally distributed. To find the parameters of the normal distribution, then we simply take our best guess of the mean - the sample mean of our data - and variance - calculated based on the data.

3.2 Variance of the sample mean

Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the variance on both sides is equal, i.e. $Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i)$.

Since all the X_i are i.i.d., we have that $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$. We know $Var(X_i) = p(1-p)$, which holds for all $i = 1, \ldots, n$, so $\sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$.

We can conclude that $Var(\bar{X}) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$, and therefore the standard deviation is $\sqrt{\frac{p(1-p)}{n}}$. After applying the central limit theorem, we thus have $\bar{X} \sim N(\mu, \frac{p(1-p)}{n})$ where μ is the sample mean of our observed data.

Once we have the distribution of $\bar{X}$, we can get the confidence intervals in the exact same manner: standardise it to $N(0, 1)$ and we can use the z-scores!