Normal Distribution, Central Limit Theorem, Confidence Intervals and Hypothesis Testing

Konstantinos I. Bougioukas

2021-02-19

Contents

Objectives

- Characteristics of Normal Distribution
- The concept of Central Limit Theorem (CLT)
- The concept of confidence intervals
- The steps of hypothesis testing

1 Characteristics of Normal Distribution

Historical background

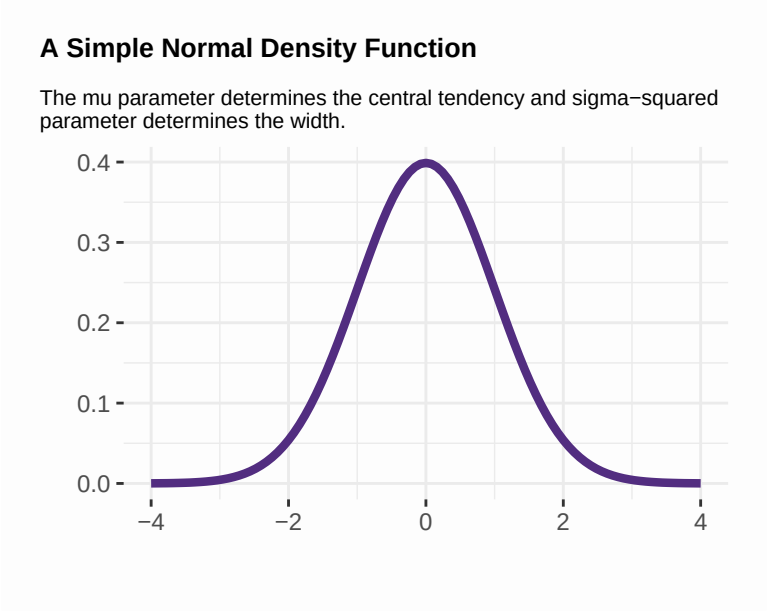
There are several important probability distributions in statistics. However, the normal distribution might be the most important. A normal distribution is the familiar "bell curve" and it's a way of formalizing a distribution where observations cluster around some central tendency. Observations farther from the central tendency occur less frequently. First, Galileo informally described a normal distribution in 1632 when discussing the random errors from observations of celestial phenomena. However, Galileo existed before the time of differential equations and derivatives. We owe its formalization to Carl Friedrich Gauss, which is why the normal distribution is often called a Gaussian distribution. A very familiar example is the height for adult people that approximates a normal distribution very well.

The mathimatical type

Gauss' normal distribution, technically a density function, is a distribution defined by two parameters, mean μ and variance σ^2. The mean , μ, is a "location parameter", which defines the central tendency. The variance, σ^2 is the "scale parameter", which defines the width of the distribution and how short the distribution is. It's formally given as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

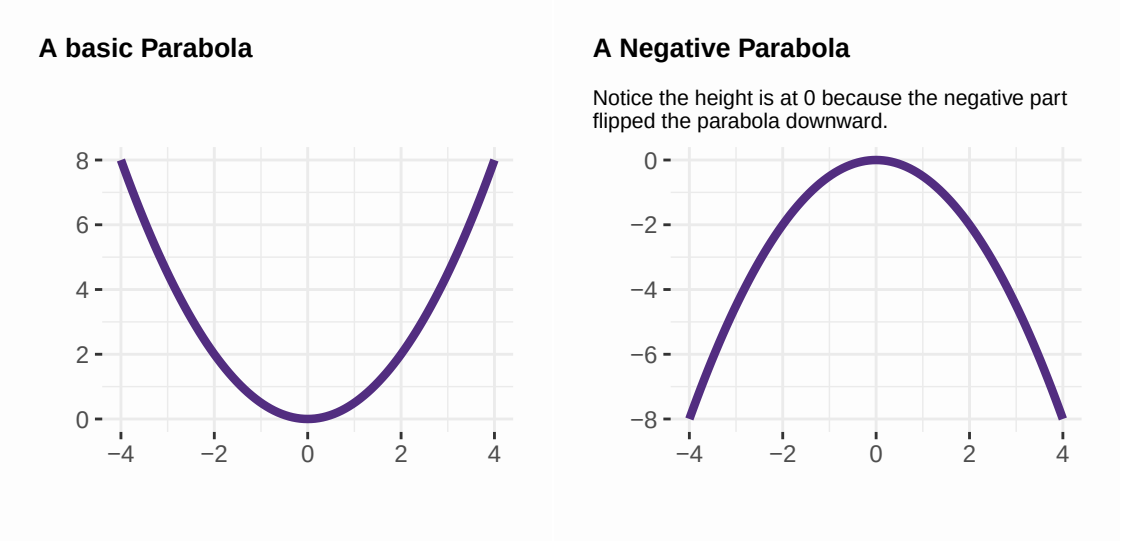
The ensuing distribution will look like this in a simple case where μ is 0 and σ^2 is 1.



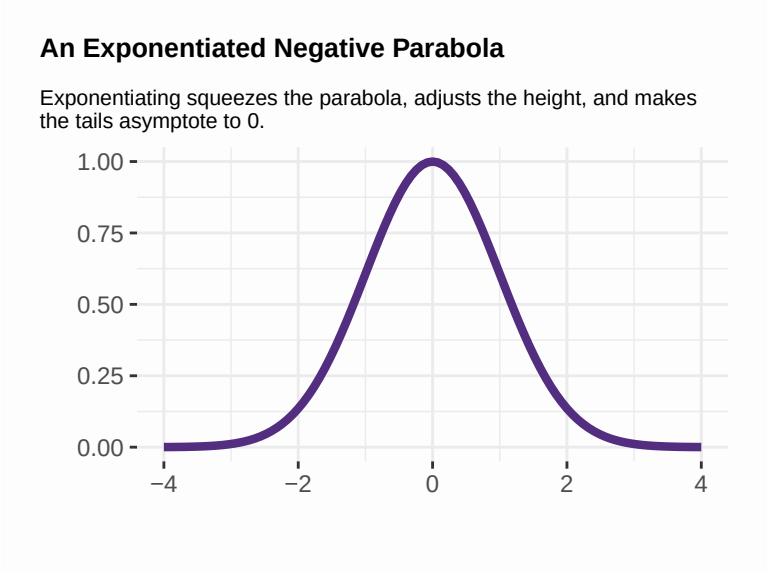
Individual components of a normal distribution

We can break down individual components of a normal distribution and explain them until they seem more accessible.

First, the "kernel" is the part inside the exponent of the above equation (i.e. $-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$). Observe, for a simple case where μ is 0 and σ^2 is 1 that this part becomes $-\frac{1}{2}x^2$ that is a negative parabola (notice the square term). The minus sign just flips the basic parabola $\frac{1}{2}x^2$ downward.



Second, exponentiating the negative parabola ($e^{-\frac{1}{2}x^2}$) makes it asymptote to 0.



Notice the tails in the above graph are asymptote to 0. "Asymptote" is a fancier way of saying the tails approximate 0 but never touch or surpass 0. One way of thinking about this as we build toward its inferential implications is that deviations farther from the central tendency are increasingly "unlikely".

Third, and with the above point in mind, it should be clear that $\frac{1}{\sigma\sqrt{2\pi}}$ will scale the height of the distribution. Observe that in our simple case where μ is 0 and σ^2 is 1, the height of the exponentiated parabola is at 1. That gets multiplied by $\frac{1}{\sqrt{2\pi}}$ to equal about 0.398. Some basic R code will show this as well.

```
# Are these two things identical?
identical(1/sqrt(2*pi), dnorm(x = 0, mean = 0, sd = 1))
```

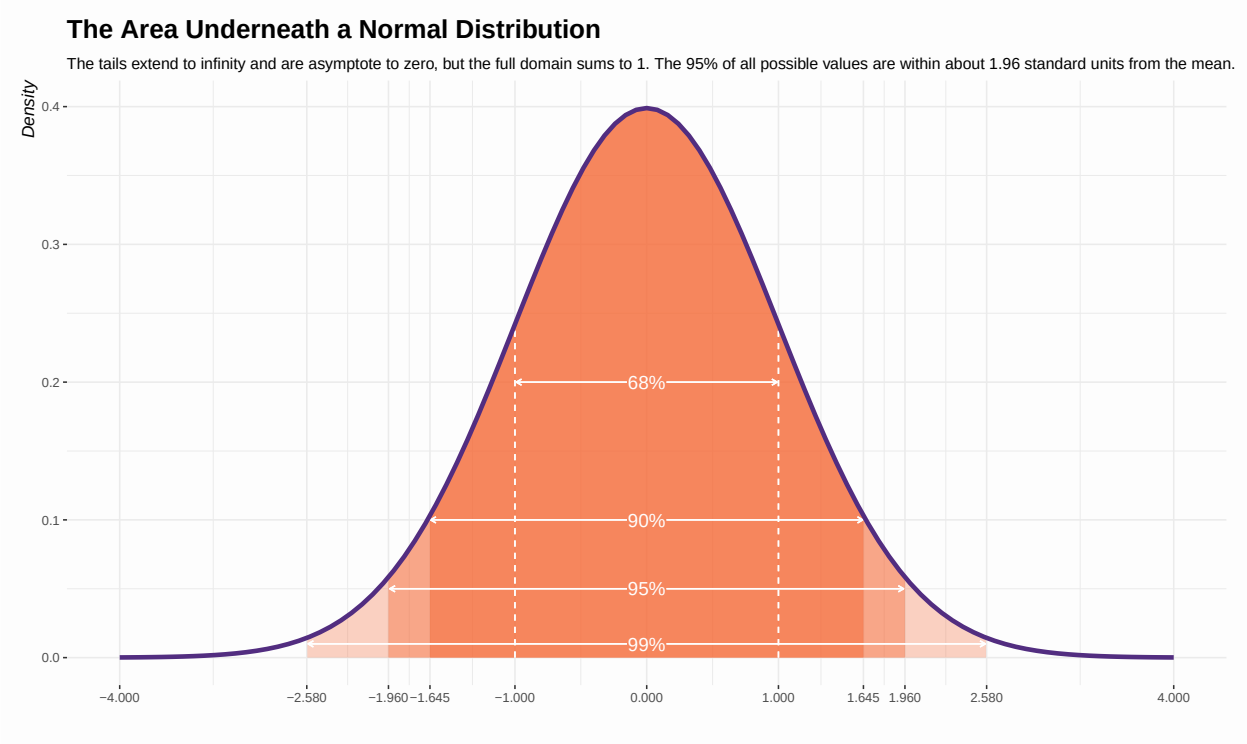
```
## [1] TRUE
```

Fourth, the normal distribution is perfectly symmetrical. The mean, μ determines the location of the distribution as well as its central tendency. All three measures of central tendency, the mode (most frequently occurring value), the median (the middlemost value), and the mean (the statistical average), will be the same. It also means a given observation of x will be as far from μ as -x. Additionally, the statistical moments of skewness and excess kurtosis are zero.

Fifth, we noted the normal distribution as a function and not a probability because the probability of any one value is effectively zero.

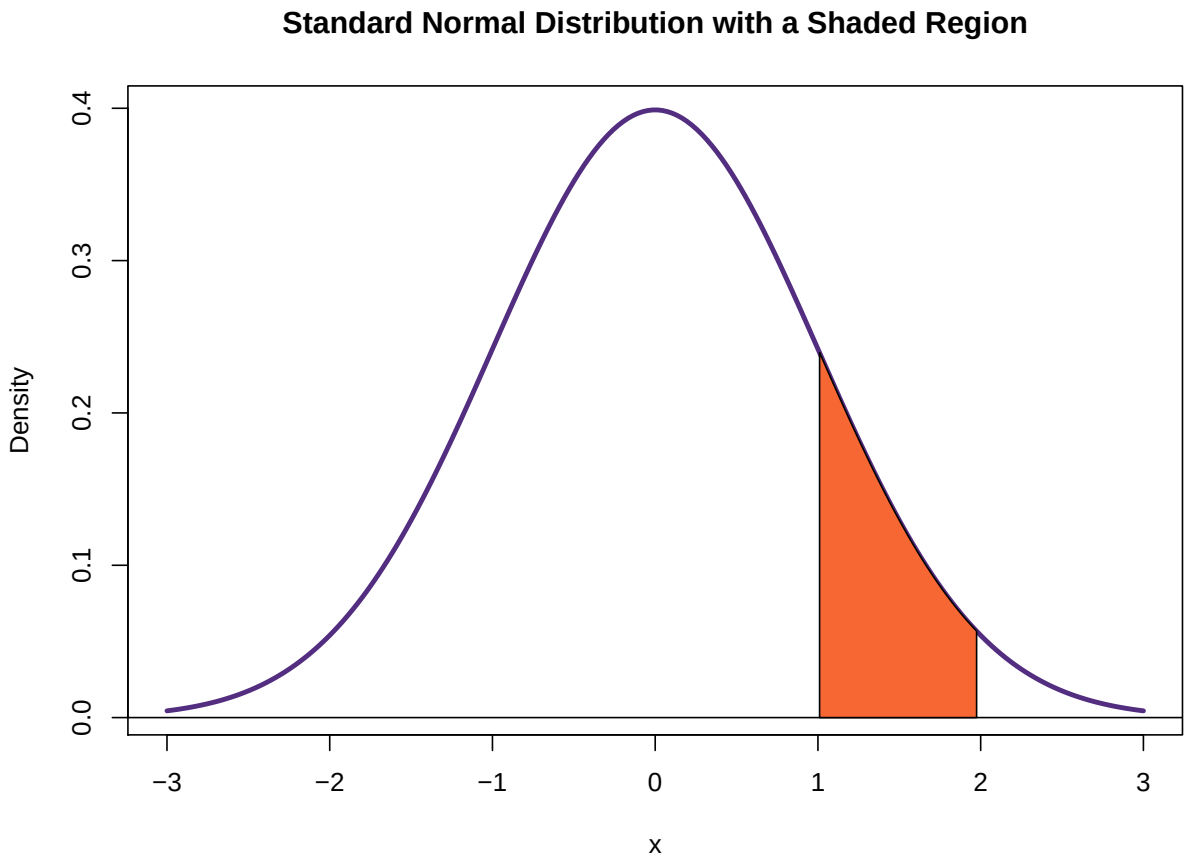
Normal Density Plot with Shaded Regions

Importantly, around 68% of the distribution is between one standard unit of μ. Around 90% of the distribution is between 1.645 standard units on either side of μ. Around 95% of the distribution is between about 1.96 standard units on either side of μ. About 99% of the distribution is between 2.58 standard units on either side of μ. So, the probability that x is between 1 on either side of the μ of 0 is effectively 0.68. The ease of this interpretation is why researchers like to standardize their variables so that the mean is 0 and the standard deviation (i.e. the scale parameter) is 1.



The normal distribution appears as a foundation assumption for a lot of quantitative approaches to frequentist statistics. It is the foundation for ordinary least squares (OLS) regression and even for some generalized linear models. Importantly, central limit theorem, itself a foundation of a lot of classical statistical testing, states that sampling distributions are effectively normal as well.

In summary, the normal density function is technically unbounded. It has just the two parameters that define its location and scale and the tails are asymptote to 0 no matter what the values of μ and σ^2 are. This makes the distribution continuous since x can range over the entire line from $-\infty$ to $+\infty$. Thus, the function does not reveal the probability of x (the probability of any one value is effectively 0). However, the area under the curve is the full domain of the probability space and sums to 1. The probability of selecting a number between two points on the x-axis equals the area under the curve between those two points.



To find the area between x = 1 and x = 2, we must subtract the area to the left of x = 1 from the area to the left of x = 2, that corresponds to the following integral:

$$E(x) = \int_1^2 f(x)dx$$

In R we can calculate this area using the `pnorm` command:

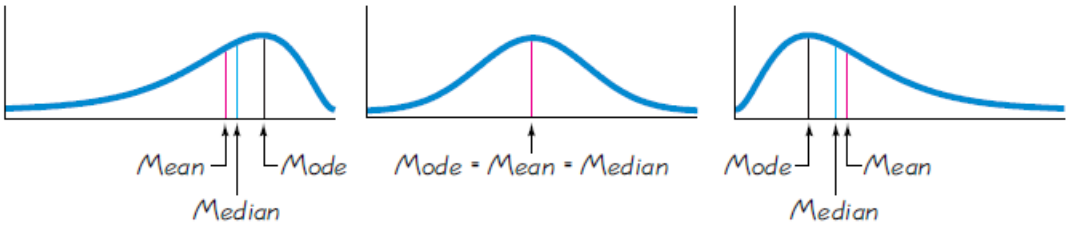
```
pnorm(2, mean = 0, sd = 1) - pnorm(1, mean = 0,sd = 1)
```

```
## [1] 0.1359051
```

Properties of an aproximately normal distribution

In an approximately **bell-shaped (normal)** distribution:

- the mean, the median and the mode have very close values
- the histogram is symmetric about the mean
- "nearly all" values (99.7%) are within -3 and +3 standard deviations of the mean
- the measure of skewness takes values close to zero (symmetric). Values below -3 or above +3 strongly indicate non-normality (Figure 1).
- the measure of excess kurtosis is close to 0 (mesokurtic). Distributions with positive excess kurtosis are called `leptokurtic` and with negative excess kurtosis `platykurtic` (Figure 2).



(a) Skewed to the Left (Negatively Skewed): The mean and median are to the *left* of the mode.

(b) Symmetric (Zero Skewness): The mean, median, and mode are the same.

(c) Skewed to the Right (Positively Skewed): The mean and median are to the *right* of the mode.

Figure 1: Skewness

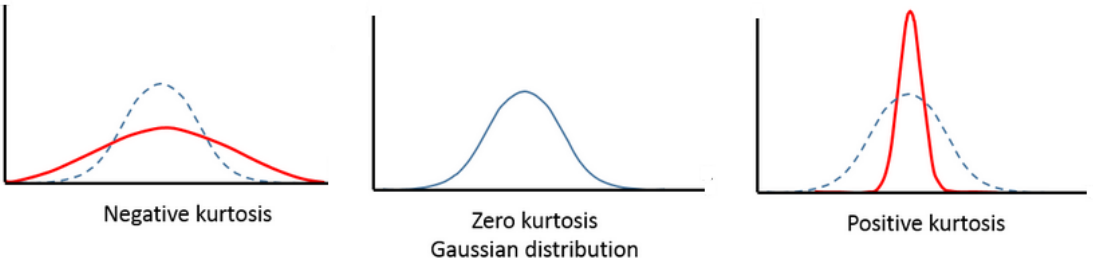


Figure 2: Kurtosis

2 Central Limit Theorem (CLT)

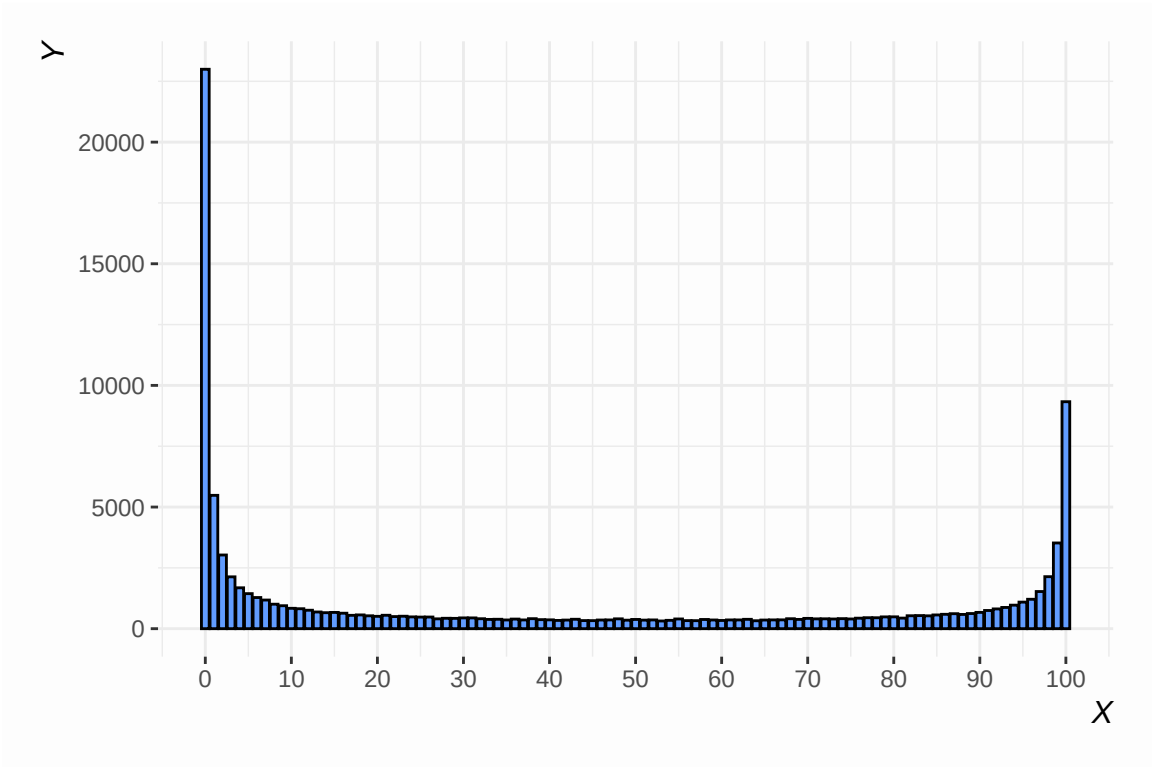
Definition of Central Limit Theorem

The central limit theorem is a novel proposition in statistical testing that proposes a description of sampling distributions from a known population. In plain English, central limit theorem's five important points are:

1. infinity samples of any size `n`
2. from a population of `N` units (where `N>n`) will
3. have sample means ($\bar{x}$) that are normally distributed.
4. The mean of sample means converges on the known population mean (μ) and
5. random sampling error would equal the standard error of the sample mean ($\frac{\sigma}{\sqrt{n}}$).

A hypothetical population

The importance of central limit theorem is that it works no matter the underlying distribution of the population data. The underlying population data could be noisy and central limit theorem will still hold. To illustrate this, let's draw some data (100,000 observations):



Here are some descriptive statistics to show how ugly these data are:

```
##    vars     n  mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 1e+05 40.16 40.31     24   37.71 35.58   0 100   100 0.39    -1.56
##      se
## X1 0.13
```

```
## [1] 40.16112
```

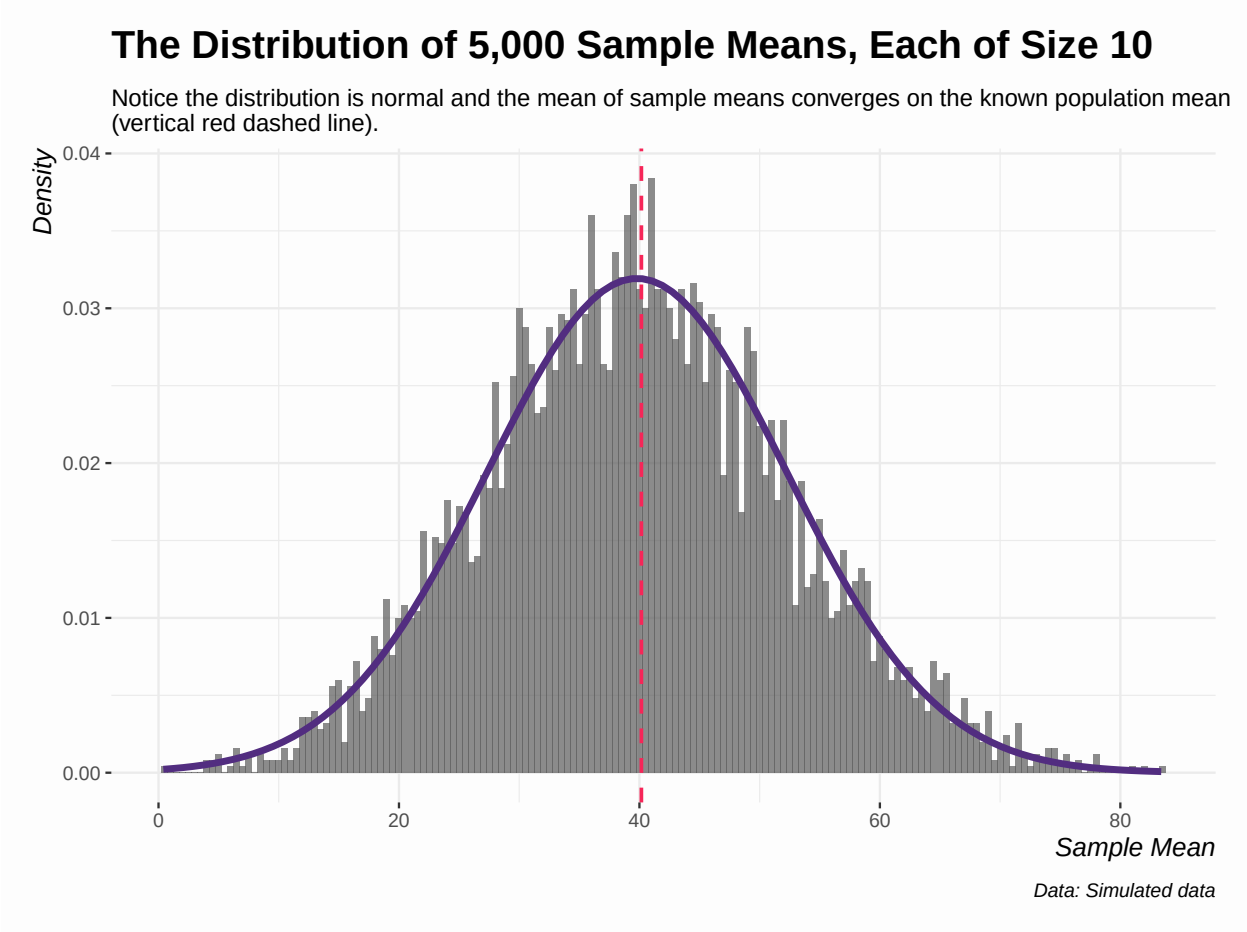
If we knew nothing else from the data other than the descriptive statistics above, we would likely guess the data would look anything other than "normal" no matter how many different values there are. There is a clear bimodality problem in these data. Namely, that "average" (i.e. the mean) doesn't look "average" at all.

The data, we have just created above, will serve as the entire population (N=100,000) of data from which we can sample.

Sampling Distribution

Now, what if we get 5,000 samples, each sample consisting of just 10 observations, save the means of those samples, and draw their histogram?

The distribution of sample means (as a density plot) converges on a normal distribution where the provided location and scale parameters are from 5,000 sample means. Further, the center of the distribution is converging on the known population mean. The true population mean 40.161 (red dashed line) is very close to the mean of the 5,000 sample means.

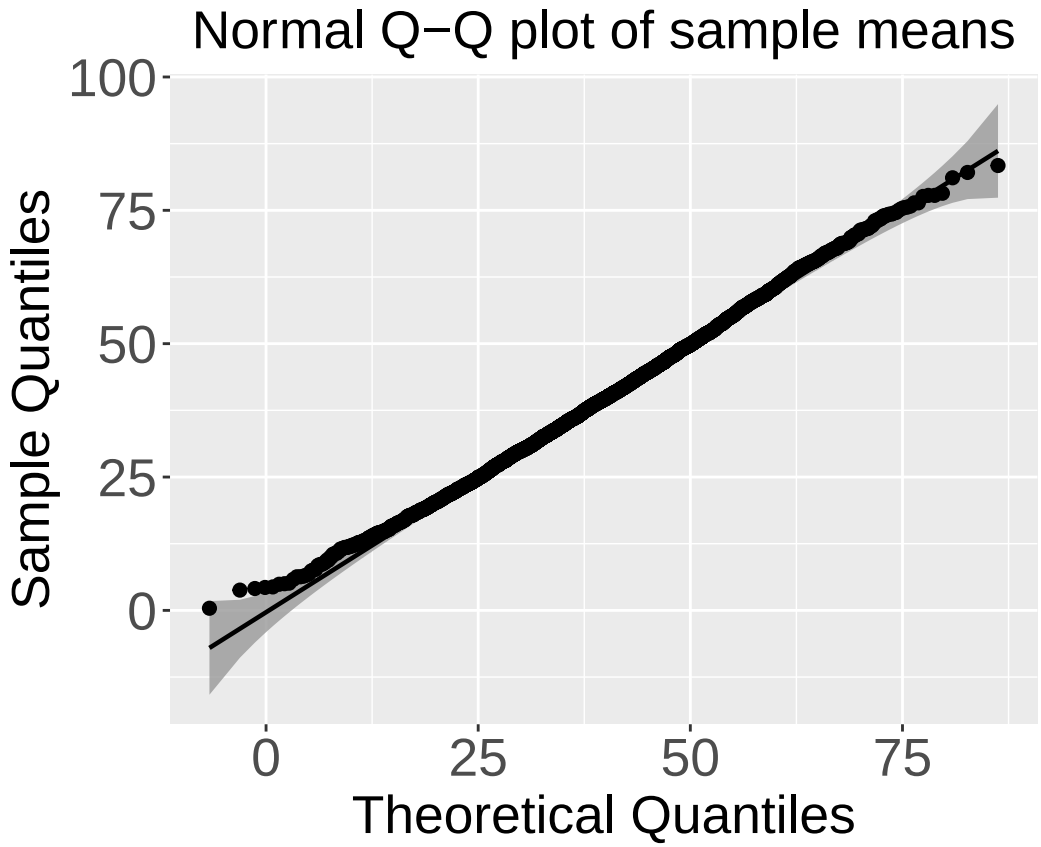


Are you surprised that a variable with a non-normal distribution in the population can have a sampling distribution that is approximately normal? This discovery is probably the single most important result presented in introductory statistics courses. Central limit theorem, says that for large samples, the sampling distribution of sample means is approximately normal. This theorem is important! Inference procedures, such as hypothesis tests and confidence intervals, are based on a normal model for the sampling distribution. The central limit theorem assures us that we can use a normal probability model for sample means without knowing anything about the shape of the distribution of the variable in the population. All we have to do is collect large samples.

Q-Q plot

In statistics, a Q–Q (quantile-quantile) plots are graphs on which quantiles from two distributions are plotted relative to each other.

On a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line).



3 The confidence intervals

We will base the definition of confidence interval on three ideas:

1. Our point estimate (e.g., mean from the sample) is the most plausible value of the actual parameter, so it makes sense to build the confidence interval around the point estimate.
2. The plausibility of a range of values can be defined from the sampling distribution of the estimate. As measure of unreliability can be used the standard error.
3. The Central Limit Theorem states that the sampling distribution is normal.

In the case of mean, and in order to define an interval, we can make use of the well-known result from probability that applies to normal distributions: roughly 95% of the distribution

is between about 1.96 standard deviations on either side of the mean. Note that for the sampling distribution, the standard deviation is actually the standard error of the mean.

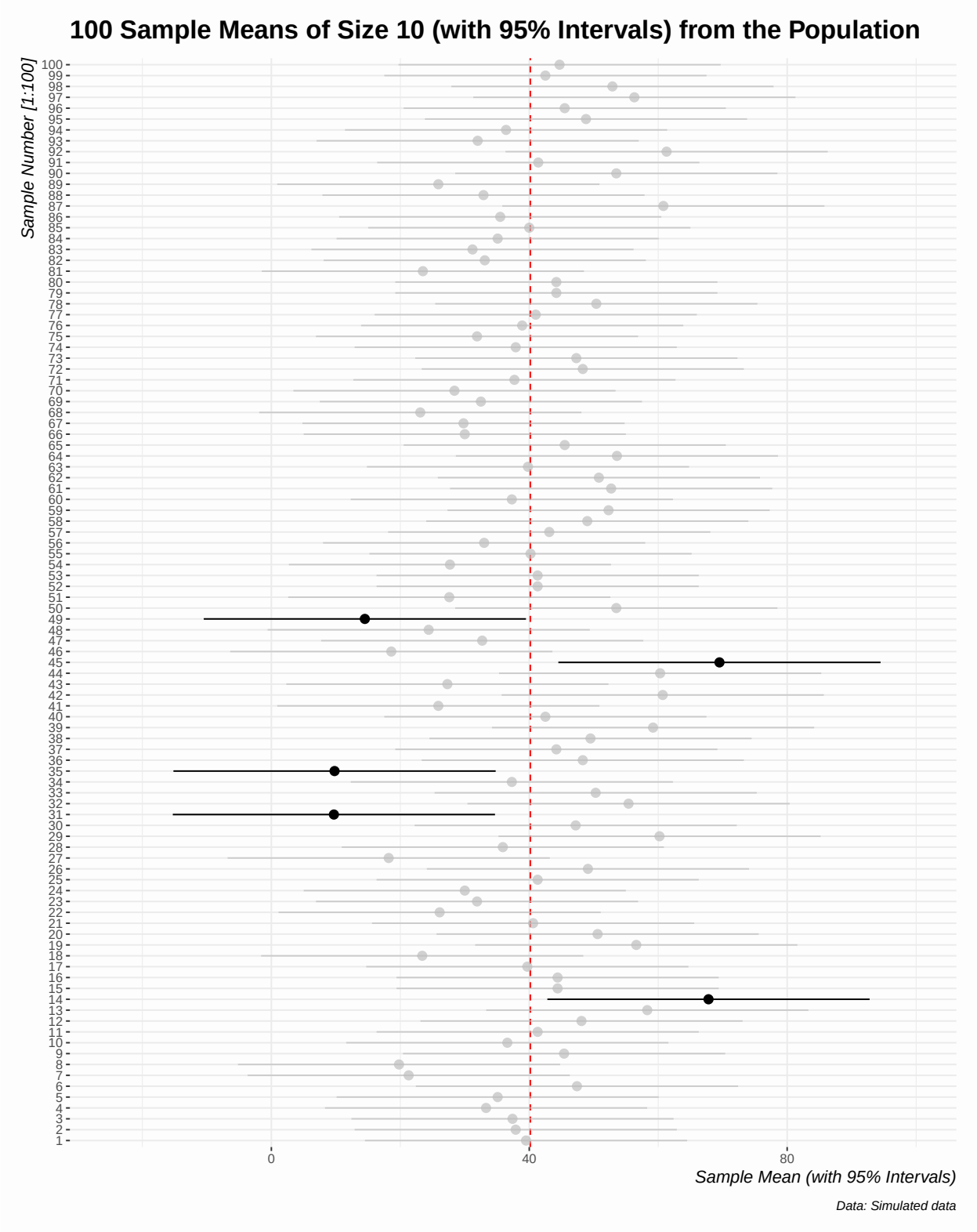
$$95\%\ Confidence\ Interval\ = mean \pm 1.96 * standard\ error$$

However, the real meaning of "confidence" is not evident and it must be understood from the point of view of the generating process.

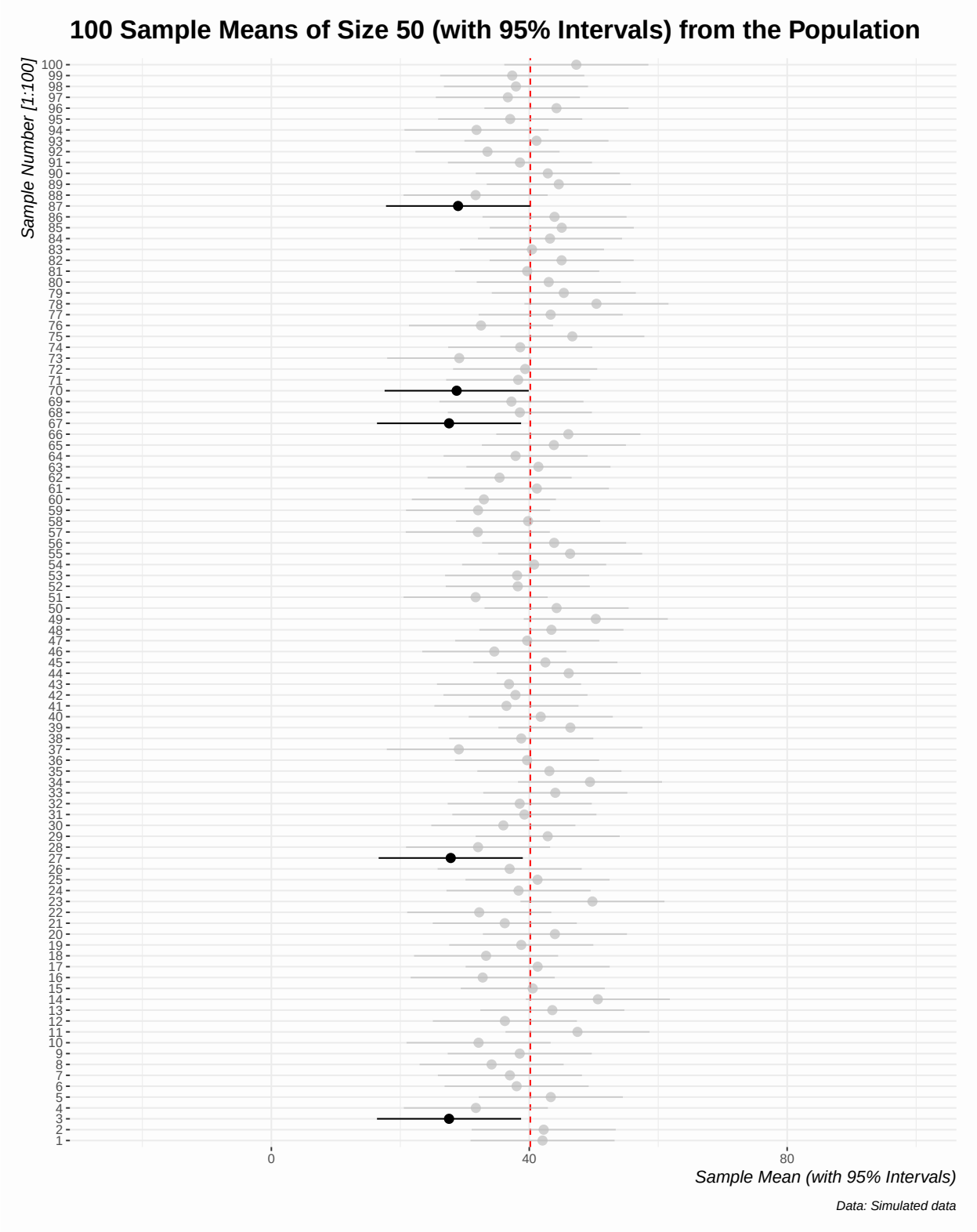
Explanation Suppose we took many (infinite) samples from a population and built a 95% confidence interval from each sample. Then about 95% of those intervals would contain the actual parameter.

Caution! It is **not exactly accurate** to say that 'there is a 95% probability that the true value of the mean falls within the confidence interval' using a frequentist approach.

Now, we can present the confidence intervals of 100 randomly generated samples of size = 10 from our population. Each horizontal bar is a confidence interval, centered on a sample mean (point). The intervals all have the same length, but are centered on different sample means as a result of random sampling from the popoulation. The five bold confidence intervals do not cover the true population mean (the vertical red dashed line $\mu = 40.161$). This is what we would expect using a 95% confidence level– approximately 95% of the intervals covering the population mean.



Next, we will create the confidence intervals of 100 randomly generated samples of size = 50 from our population:



Increasing sample size not only converges the sample statistic on the population parameter (red dashed line) but decreases the uncertainty around the estimate.

4 Hypothesis Testing

Statistical hypothesis testing involves a model for data:

- Parametric tests have very specific models
- Nonparametric tests have semi-specific models without a distribution assumption
- Permutation tests make an assumption about the best data summarization (e.g., mean, median)

A statistical hypothesis test is a method of statistical inference. Most medical statistics are based on the concept of hypothesis testing and therefore an associated p-value is usually reported.

In hypothesis testing, a **'null hypothesis'** (H_0) is first specified, that is a hypothesis stating that there is no difference, for example, there is no difference in the summary statistics of the study groups (placebo and treatment). The null hypothesis assumes that the groups that are being compared are drawn from the same population.
An **'alternative hypothesis'** (H_1), which states that there is a difference between groups, can also be specified.

Definition The **p-value** is, the probability of obtaining a difference as large as or larger than the one observed between the groups, assuming the null hypothesis is true (i.e., no difference between groups).

`General method for hypothesis testing step-by-step`

1. From the research question, determine the appropriate null hypothesis, H_0, and the alternative, H_1 (ususally two-sided).
2. Set the level of significance, α (usually 0.05).
3. Identify the appropriate test statistic and calculate the observed test statistic from the data.
4. Using the known distribution of the test statistic and calculate the p-value. Compare the p-value to significant level α. If $p - value < \alpha$, reject the null hypothesis. If $p - value \geq a$, do not reject the null hypothesis.
5. Interpret the results.

The Figure 3 shows what the two-tailed rejection region looks like for α (usually 0.05). The α is divided evenly between the left tail (also called the lower tail) and the right tail (the upper tail).

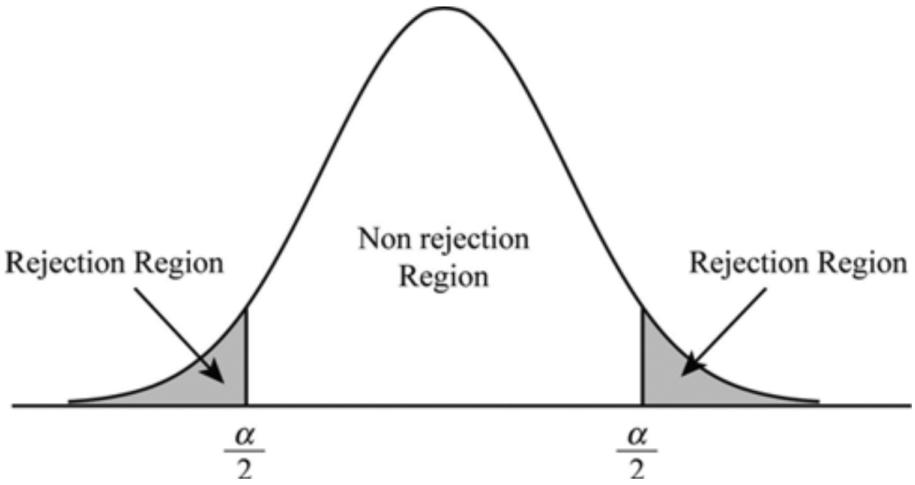


Figure 3: The two-tailed rejection region for a (usually 0.05)

Rejecting the null hypothesis when the null is true represents a Type I error (which is the α), while a Type II error refers to failing to reject the null hypothesis when the null is false.

Choosing a significance level

Reducing the error probability of one type of error increases the chance of making the other type. As a result, the significance level (α) is often adjusted based on the consequences of any decisions that might follow from the result of a significance test.

By convention, most scientific studies use a significance level of $\alpha = 0.05$; small enough such that the chance of a Type I error is relatively rare (occurring on average 5 out of 100 times), but also large enough to prevent the null hypothesis from almost never being rejected. If a Type I error is especially dangerous or costly, a smaller value of α is chosen (e.g., 0.01). Under this scenario, very strong evidence against H0 is required in order to reject H0. Conversely, if a Type II error is relatively dangerous, then a larger value of α is chosen (e.g., 0.10). Hypothesis tests with larger values of α will reject H0 more often.

For example, in the early stages of assessing a drug therapy, it may be important to continue further testing even if there is not very strong initial evidence for a beneficial effect. If the scientists conducting the research know that any initial positive results will eventually be more rigorously tested in a larger study, they might choose to use $\alpha = 0.10$ to reduce the

chances of making a Type II error: prematurely ending research on what might turn out to be a promising drug.

A government agency responsible for approving drugs to be marketed to the general population, however, would like to minimize the chances of making a Type I error— approving a drug that turns out to be unsafe or ineffective. As a result, they might conduct tests at significance level 0.01 in order to reduce the chances of concluding that a drug works when it is in fact ineffective.

The US FDA and the European Medical Agency (EMA) customarily require that two independent studies show the efficacy of a new drug or regimen using $\alpha = 0.05$, though other values are sometimes used.