Statistical inference

Daniel Hammarström

2019-10-14

Statistical inference

- What are Rønnestad and Vikmoen talking about here?

Received: 8 April 2019 | Revised: 31 July 2019 | Accepted: 7 August 2019

DOI: 10.1111/sms.13536

ORIGINAL ARTICLE

WILEY

A 11-day compressed overload and taper induces larger physiological improvements than a normal taper in elite cyclists

Bent R. Rønnestad[1] | **Olav Vikmoen**[1,2]

- Specifically, are they talking about the results in their study, or the "real world" it represents?

Statistical inference

- A scientific study is often designed to answer questions about the "real world" based on some collected data.
- This is because we are unable to measure all possible cases.
- The data are collected under controlled circumstances so that claims about the "real world" are unbiased
- Researchers aim to describe e.g. a difference between two training protocols outside the laboratory, given some set of assumptions and based on a sample of data.
- Based on our collected data, we want to infer what the *true* real-world value is.

Population and sample

- The population (in statistical terms) are all the **possible values** a **variable** can take.
- We can therefore talk about a population of values, e.g. height of all men in Norway, or all possible test results of a group of athletes.
- The sample is a **subset** of values from the population.
- The problem we face when trying to infer about the *population* based on a *sample* is that we will estimate with some *error*.
- If we have designed our experiment in a bad way, the error will to a large extent consist of **bias**.
- If the experiment is well designed, without bias, we are still left with **sampling error**

A simple example

- We ordered blue and red pills from China. We messed up the order and the factory forgot to document how many blue pills there were.
- We need to know the proportion of blue pills as blue and red pills has different prices.
- The total number of pills in our containers is $N = 10000000$, we wont be able to count them all, we can only use a sample.

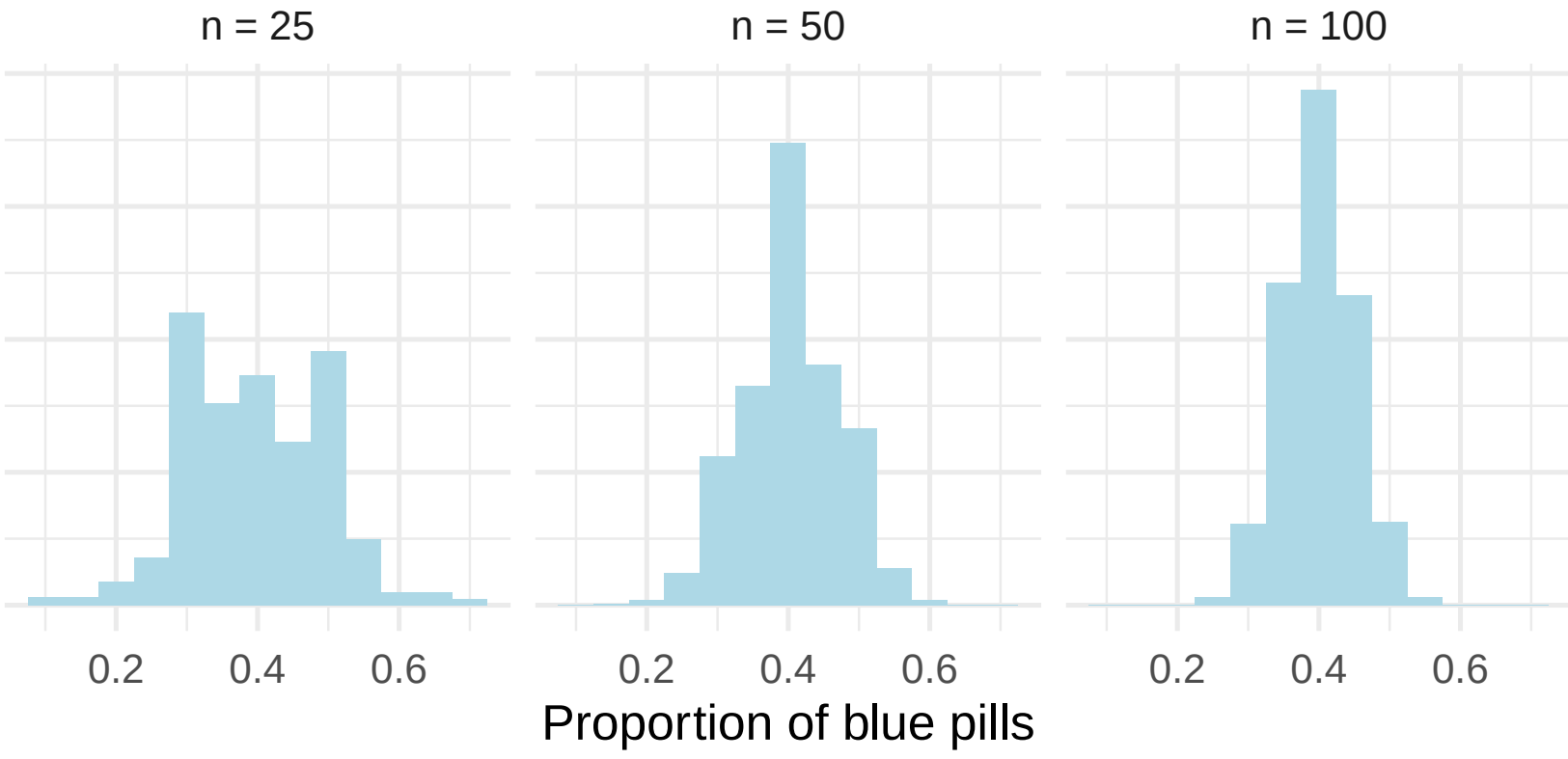
A simple example

- We open a box and count 10 pills, there are 2 blue and 8 red pills (20% blue).
- Another 10 pills are drawn there are 4 blue and 6 red pills (40% blue).
- As long as all containers/boxes are unbiased, the differences in samples are due to sampling error.
- Your boss tells you that you only need to count 50 pills to know the true proportions of pills.
- You think it makes sense that if the sample size is bigger we are more confident that the sample represents the actual population.
- To prove you point, you spend the night counting pills...

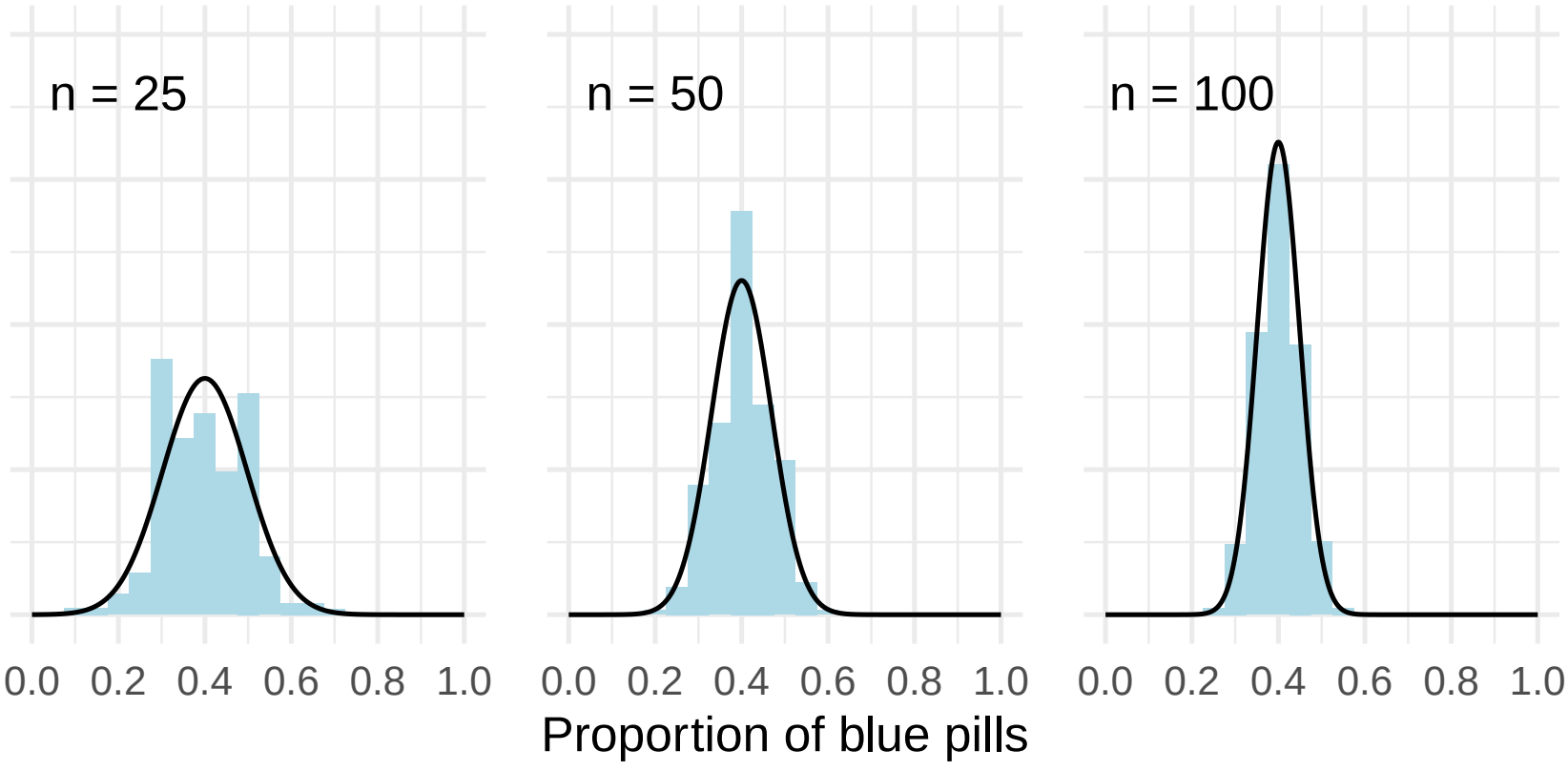
A distribution of samples

- When we draw a sample from a population, it will give us a best guess of the true proportion.
- The true, unknown population proportion can be called p, our best guess can be called $\hat{p}$
- When we calculate the proportion of blue pills in a sample it might be $\hat{p} = 0.45$ (45%). In another draw, it might be $\hat{p} = 0.55$.
- If we would have drawn many samples we would have created a **sampling distribution**.
- To prove to your boss that sample sizes matters, you draw random samples of 25, 50 and 100 and count the number of blue pills.

A distribution of samples



A sampling distribution can be modeled using the normal distribution



The sampling distribution of proportions

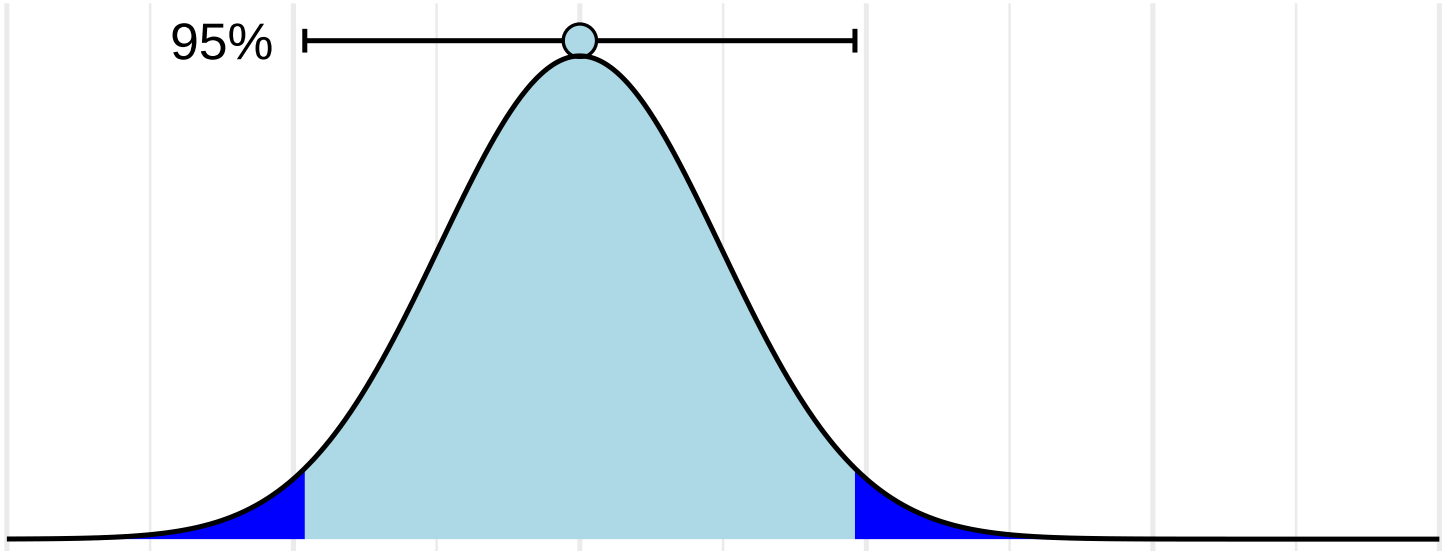
- You have spent all night counting pills and proven to your boss that the sample size matters in the sense that bigger samples more often gives a value close to your best guess.
- It turned out that the sampling distribution looked like a normal distribution created with the the p as center and a larger spread with smaller sample sizes.
- After reading up on some statistics, you find out that the spread or standard deviation of the sampling distribution for proportions could be calculated as

$$= \sqrt{\frac{p(1-p)}{n}}$$

- This is called the **standard error**.
- If we only take one sample, **we can estimate** the mean and spread of the sampling distribution

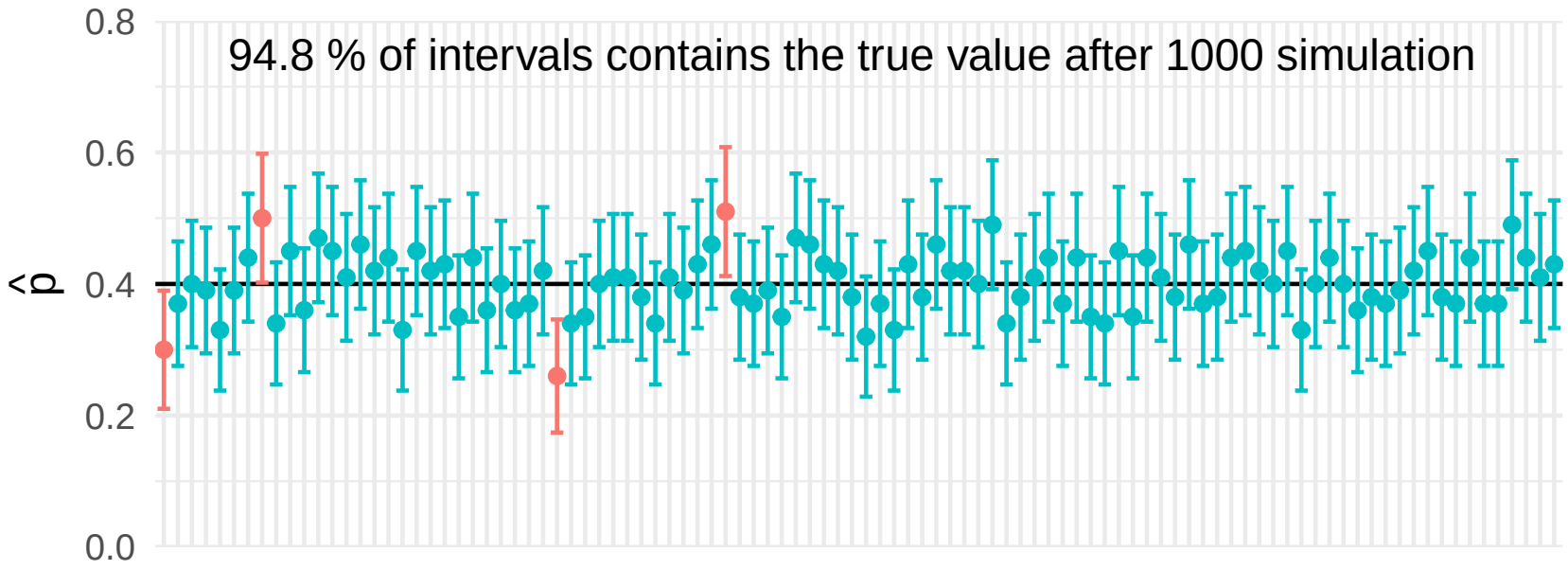
The normal distribution

- Given some assumptions, a sample can be used to estimate a sampling distribution.
- This estimated sampling distribution is then used to draw inference about the real-world.
- A property of the normal distribution is that 95% of the data fits under $\pm 1.96\times$ the standard deviation
- As the standard deviation of the sampling distribution can be estimated using the standard error (taking sample size into account), we can estimate a range a plausible values of the sampling distribution



The confidence interval

- A confidence interval contains 95% of the data of the estimated sampling distribution.
- The confidence interval thus has an interpretation: **upon repeated sampling, the** 95% **confidence interval will contain the true value** 95% **of the time**
- To prove this point, we can make a small simulation: (1) Using the blue and red pills, draw a sample and calculate $\hat{p}$ and a 95% confidence interval. (2) repeat the process to check how many intervals contains the true value ($p = 0.4$).



The confidence interval for hypothesis testing

- The confidence interval can be used to perform hypothesis testing as the interval contains plausible values of true parameter.
- A statistical hypothesis test consists of to competing positions of denoted as H_0, the null-hypothesis and H_A, the alternative hypothesis.
- By convention, we test against H_0. Why?

The confidence interval for hypothesis testing

- Let's say that if everything is done at random, our Chinese factory will produce equally many blue and red pills.
- This can be our H_0: There is no difference in the number of blue and red pills ($p = 0.5$).
- An alternative hypothesis could be that H_A: Manufacturing is biased in some direction, $p \neq 0.5$.
- We can use the confidence interval to test these hypotheses. As we would test against the H_0 we would create confidence intervals a check if they contained the H_0 or not.

Sample size	$\hat{p}$	CI lower bound	CI upper bound
25	0.240	0.073	0.407
50	0.420	0.263	0.577
100	0.430	0.318	0.542
1000	0.404	0.396	0.438

Numerical data Population and sample

- Up to now, we have used proportions as examples.
- The same principles are relevnt for other types of data such as categorical and numerical data.
- When we use data on the interval or ratio scale, the sample mean is used to estimate the population parameter (the population mean).

The population mean:

$$\mu = \frac{\sum X_i}{N}$$

The sample mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

The mean is a measure of central tendency, *example*:

$$X_1 = 90, X_2 = 95, X_3 = 100, X_4 = 105, X_5 = 110$$

Measures of central tendency

- The mean
- The median
- The mode

Measures of dispersion

- Central tendency captures e.g. the "center of gravity" in the data, however, we might also want to know something about its variation.
- The population variance:

$$\sigma^2 = \sum_{i=1}^{N} (X_i - \mu)^2 / N$$

- As the population parameters are unknown, we estimate them with our sample:

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$$

Variance

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
103.0	99.6	-3.4	11.6
105.7	99.6	-6.1	37.2
84.9	99.6	14.7	216.5
91.2	99.6	8.4	71.0
113.0	99.6	-13.4	179.1
99.6	99.6	0.0	0.0
100.5	99.6	-0.9	0.8
88.3	99.6	11.3	128.3
112.9	99.6	-13.3	176.7
96.9	99.6	2.7	7.2

Sum of deviations $= \sum x_i - \bar{x} = 0.$

Sum of the square deviations $= \sum (x_i - \bar{x})^2 = 828$

Average square deviations $= \frac{\sum (x_i - \bar{x})^2}{n-1} = 92 = \textit{variance}$

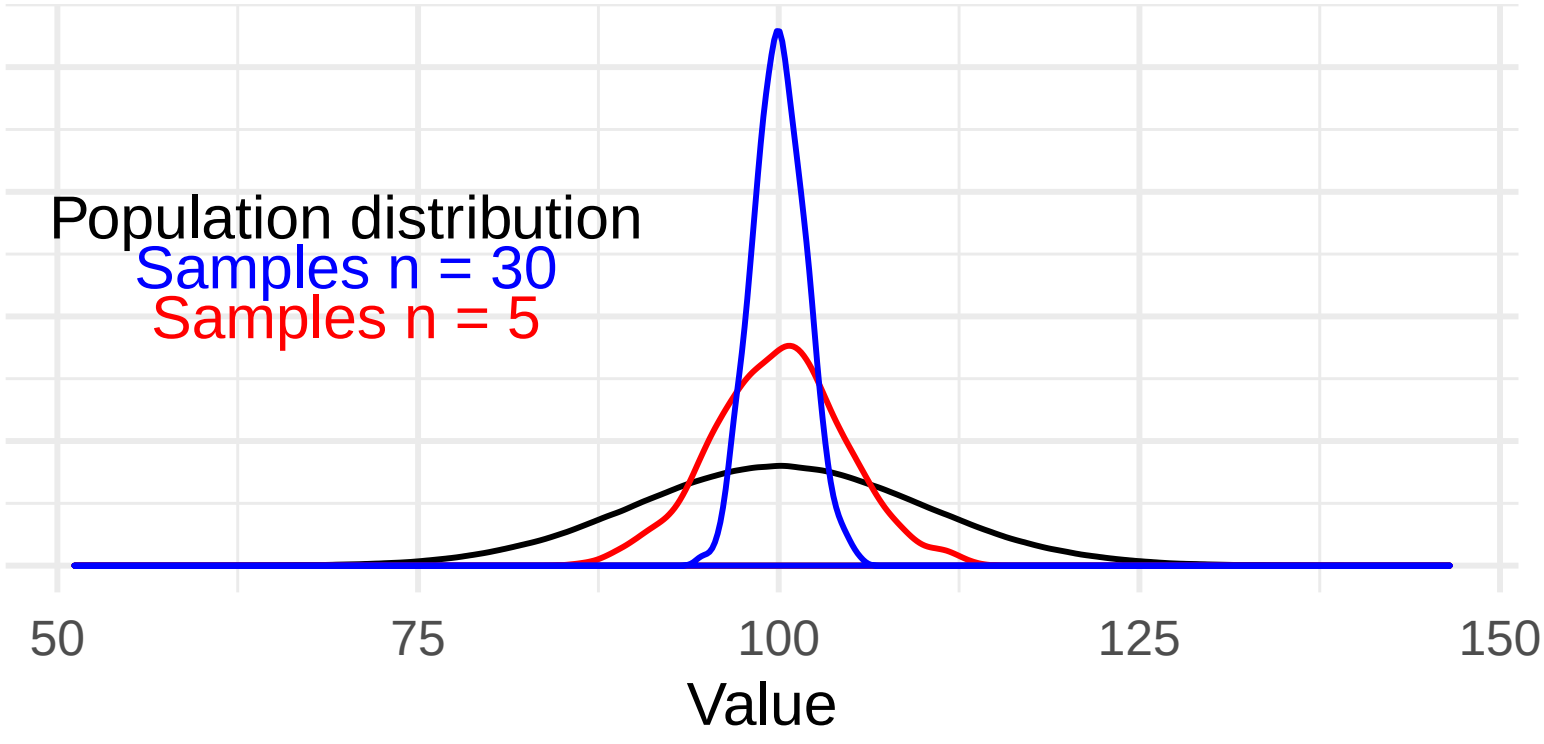
Variance and the standard deviation

- The variance is the average squared deviation from the mean
- The standard deviation is the square root of the variance, thus on the same scale as the mean

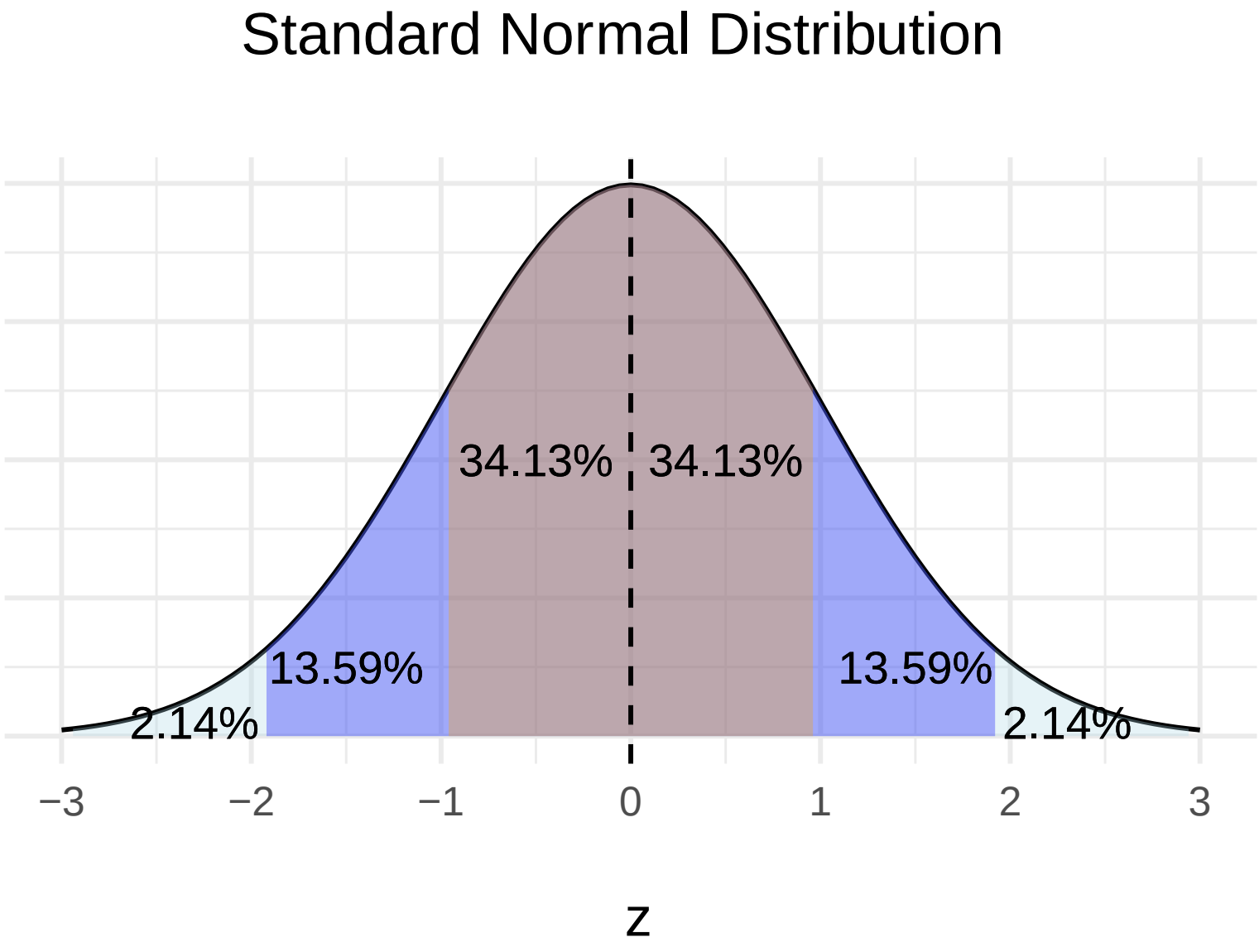
$$\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = SD$$

Sampling distributions

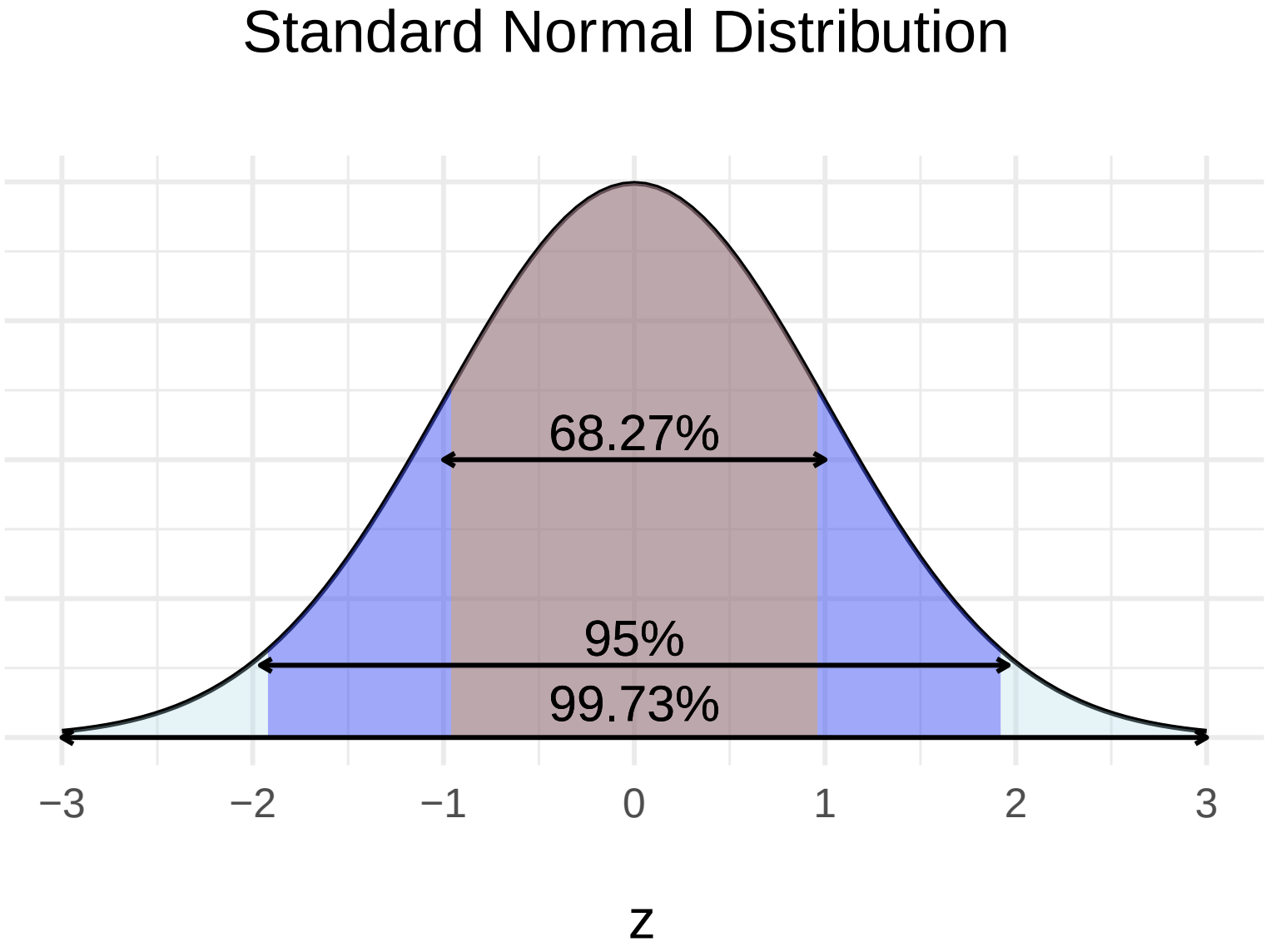
- Any *statistic* can be calculated from a sample and used as an estimation of the population parameter.
- As an example, the sample mean ($\bar{x}$) is an unbiased estimator of the population mean, we know this because the average of repeated samples from a population will be close to the population mean (μ).



The Normal Distribution



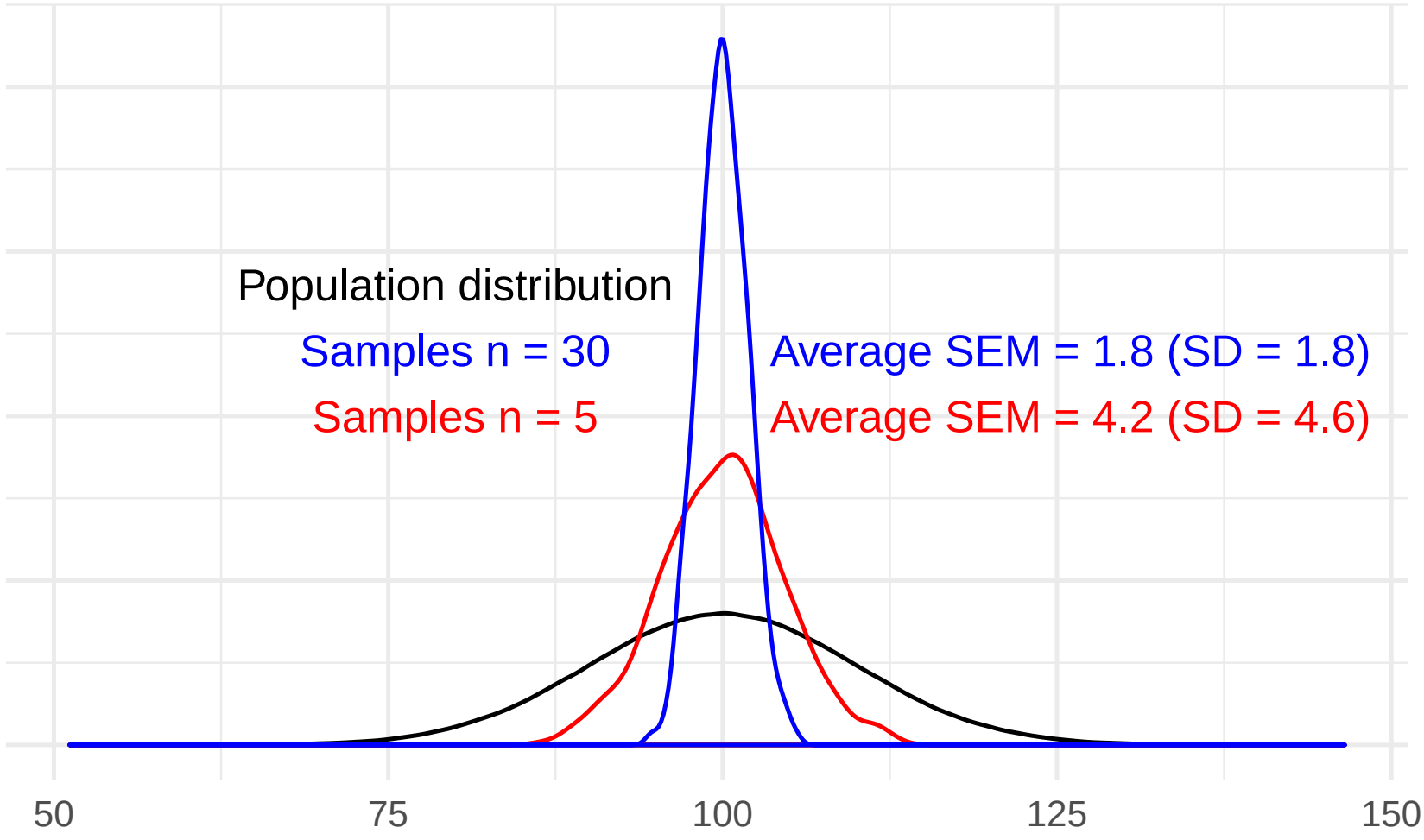
The Normal distribution



Sampling distributions

- The variation of a distribution of averages is affected by the sample size, the normal distribution does not account for sample sizes
- This variation can be estimated from samples, this is known as the standard error.
- As with proportions, the **sample standard error** is an estimator of the **standard deviation of the sampling distribution**!

Sampling distributions



Hypothesis testing

- As with a proportion, based on the estimate of the sampling distribution we can device a test, to test if a value exists within specified range.
- 95% of all values lies within $\pm 1.96 \times \sigma$ from the mean in a normal distribution, this leaves us with an uncertainty of 5%.
- However, due to problems with *proving a theory or hypothesis*, we instead test against a null-hypothesis. Thus we try to *disprove the hypothesis*.
- The null hypothesis H_0 is constructed to contain scenarios not covered by the alternative hypothesis H_A

Hypothesis tests - a two sample scenario

- The null hypothesis is that the mean of group 1 is similar to group 2

$$H_0 : \mu_1 - \mu_2 = 0$$

- To reject this hypothesis, we need to show that

$$\mu_1 - \mu_2 \neq 0$$

- We want to do this with some specified uncertainty, usually 5%
- We can calculate a 95% confidence interval of the difference

A 95% confidence interval for small samples

Upper bound:

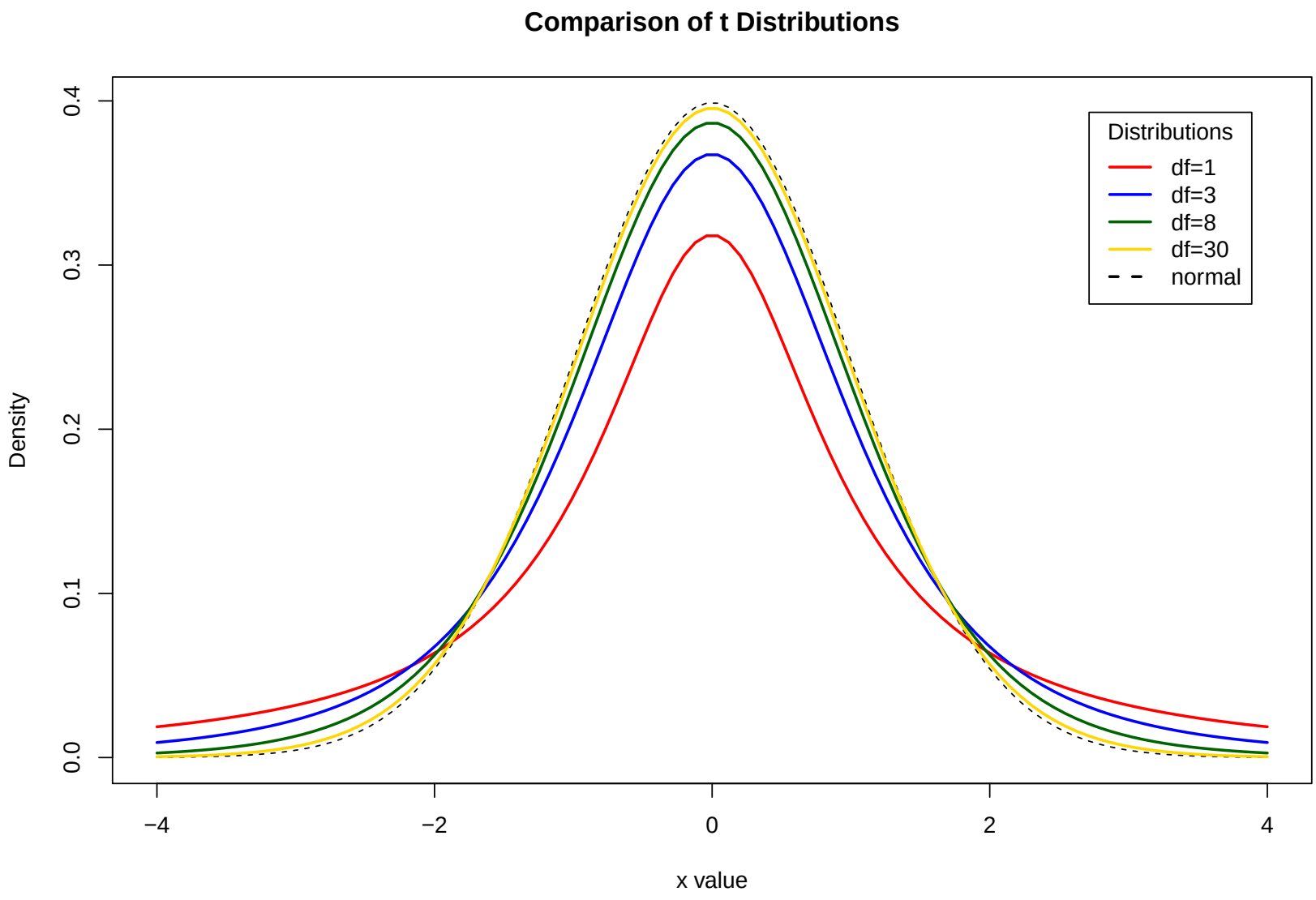
$$\bar{x} + t_{1-\alpha/2} \times SE$$

Lower bound:

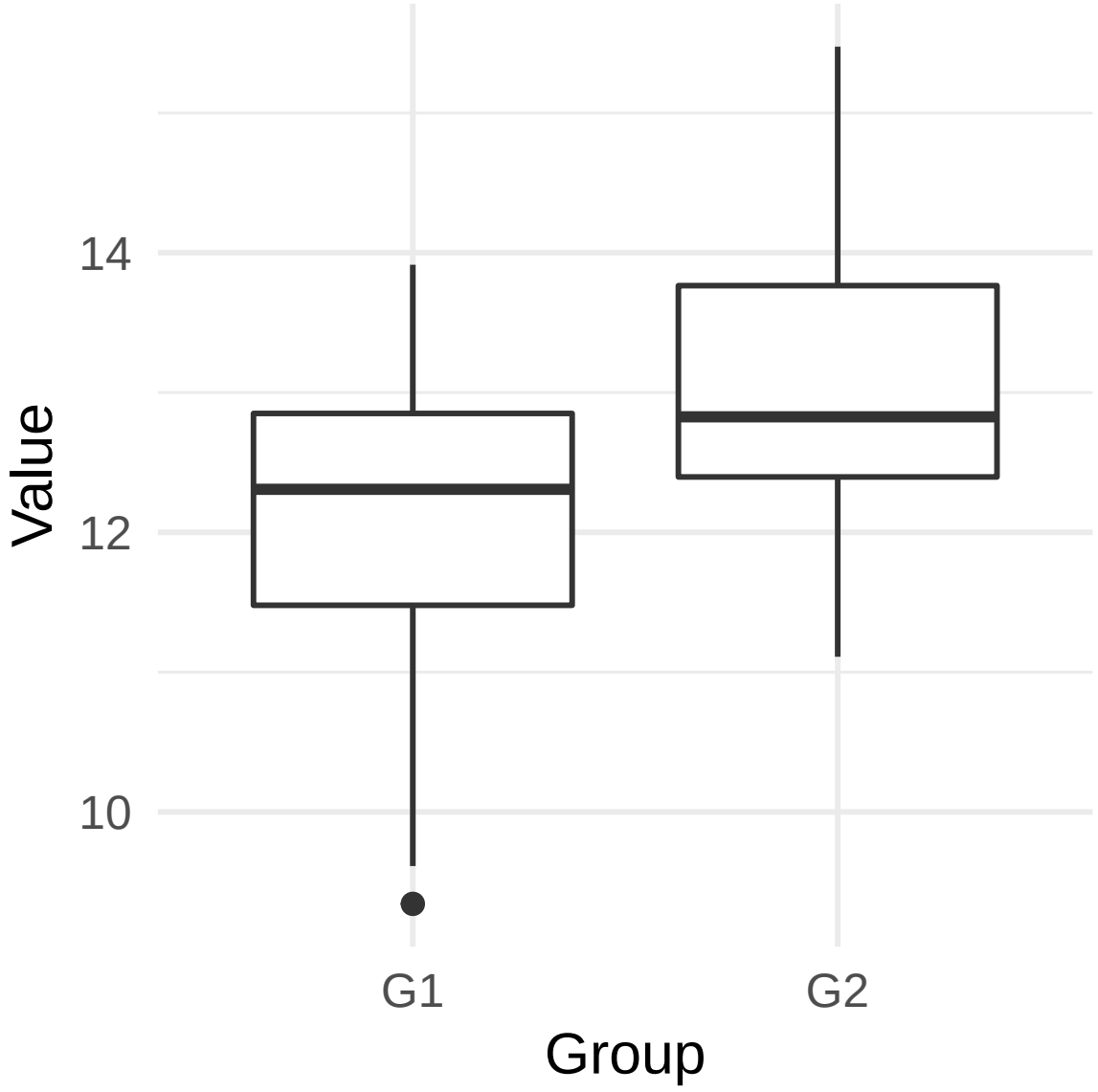
$$\bar{x} - t_{1-\alpha/2} \times SE$$

- $\bar{x}$ is the difference in means between groups.
- The standard error (SE) estimates the standard deviation of the sampling distribution
- $t_{1-\alpha/2}$ represents the area under probability distribution curve containing 95% of all values.
- The t-distribution is used instead of the normal distribution since it can take sample-size into account.

The t-distribution



A two sample case



A 95% confidence of the difference in means

- Two groups are compared, the H_0 is that there is no difference between the groups:

$$H_0 : \mu_1 = \mu_2$$

- The difference between the groups are estimated to $\mu_2 - \mu_1 = 0.97$
- The 95% confidence interval is

$$m_2 - m_1 \pm t_{\alpha/2} \times SE(m_2 - m_1)$$

where the

$$SE(m_2 - m_1)$$

is the standard error of the difference.

$$0.97 \pm 2 \times 0.28$$

Key points so far

- We can *estimate* population *parameters* using a **random** sample from the population
- The calculated sample standard error is an estimate of the standard deviation of a sampling distribution
- Using a *probability density function* like the t- or z-distribution, we can estimate a range a plausible values of a population parameter (e.g. mean).
- We can test if a estimated interval contains the null hypothesis, if not we can reject H_0.