$z-$ or $t-$ distribution

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

$z -$ ή $t -$ formulas

$$\bar{x} \pm z\frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

Examples

Example 1

- We randomly select ten people aged 65 and over (sample) of millions of people in this age group (population of interest)
- Find the mean diastolic blood pressure
- In our paper or report we say «The average diastolic blood pressure is …»

Example 2

- We randomly select five students from a high school class
- Find the mean height
- In our paper or report we say «The average height is …»

These examples sound problematic

The limits of data in research

- Our primary focus while conducting quantitative research or analysis is typically on large populations
  - The prevalence of depression among people aged 65 or older
  - The average height of students
- However, we almost always use sample data to represent the larger population due to cost and time limitations
- However, sample data is always an estimate or approximation of the whole population that it was taken from

The limits of data in research

- We are less confident that our sample size n is representative of our entire population as it gets smaller. The risk of error is greater
- Additionally, we often know nothing about the population· its mean, variance, or standard deviation

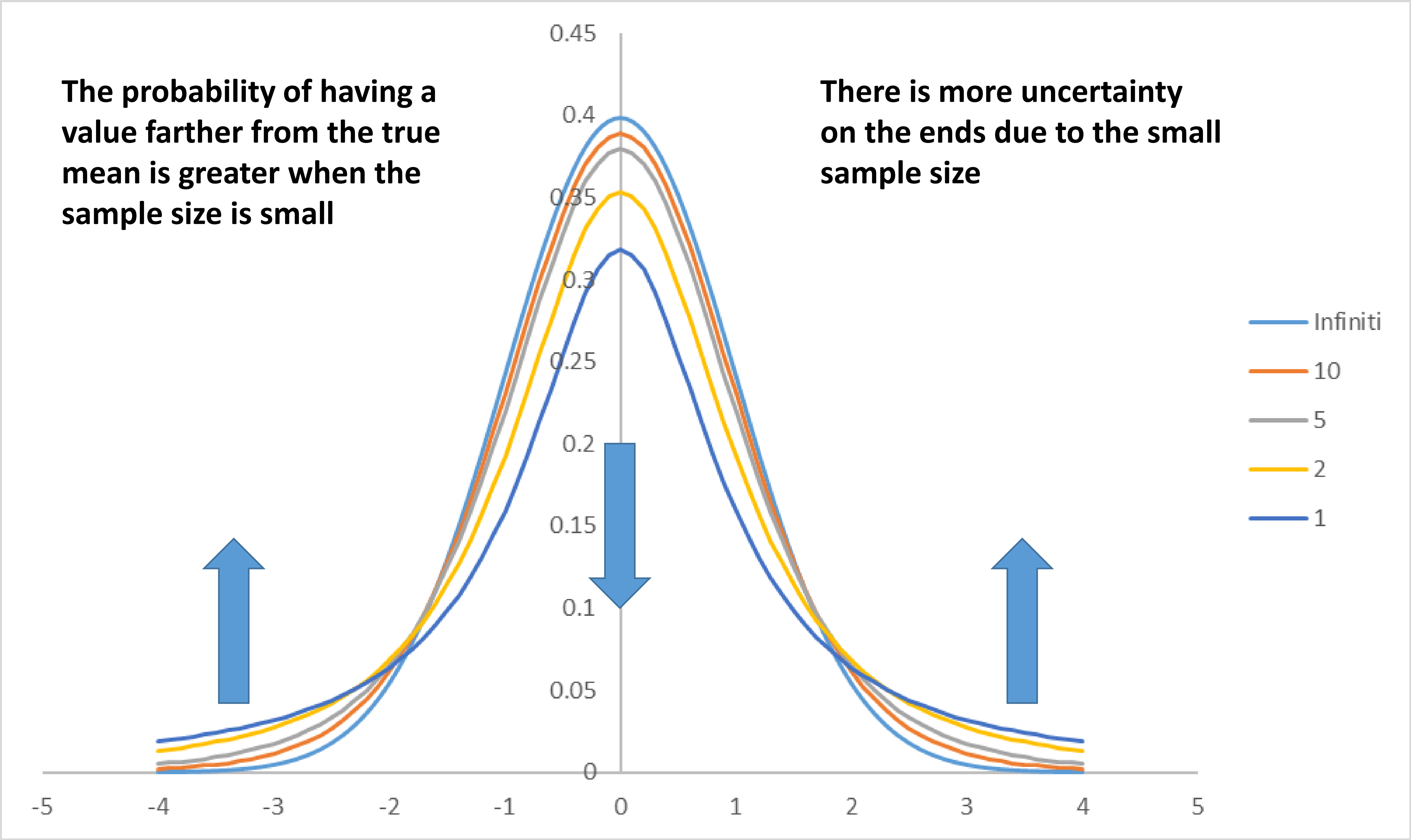
Thus, we are using sample data from a larger population about which we have little or no information

Sample Size

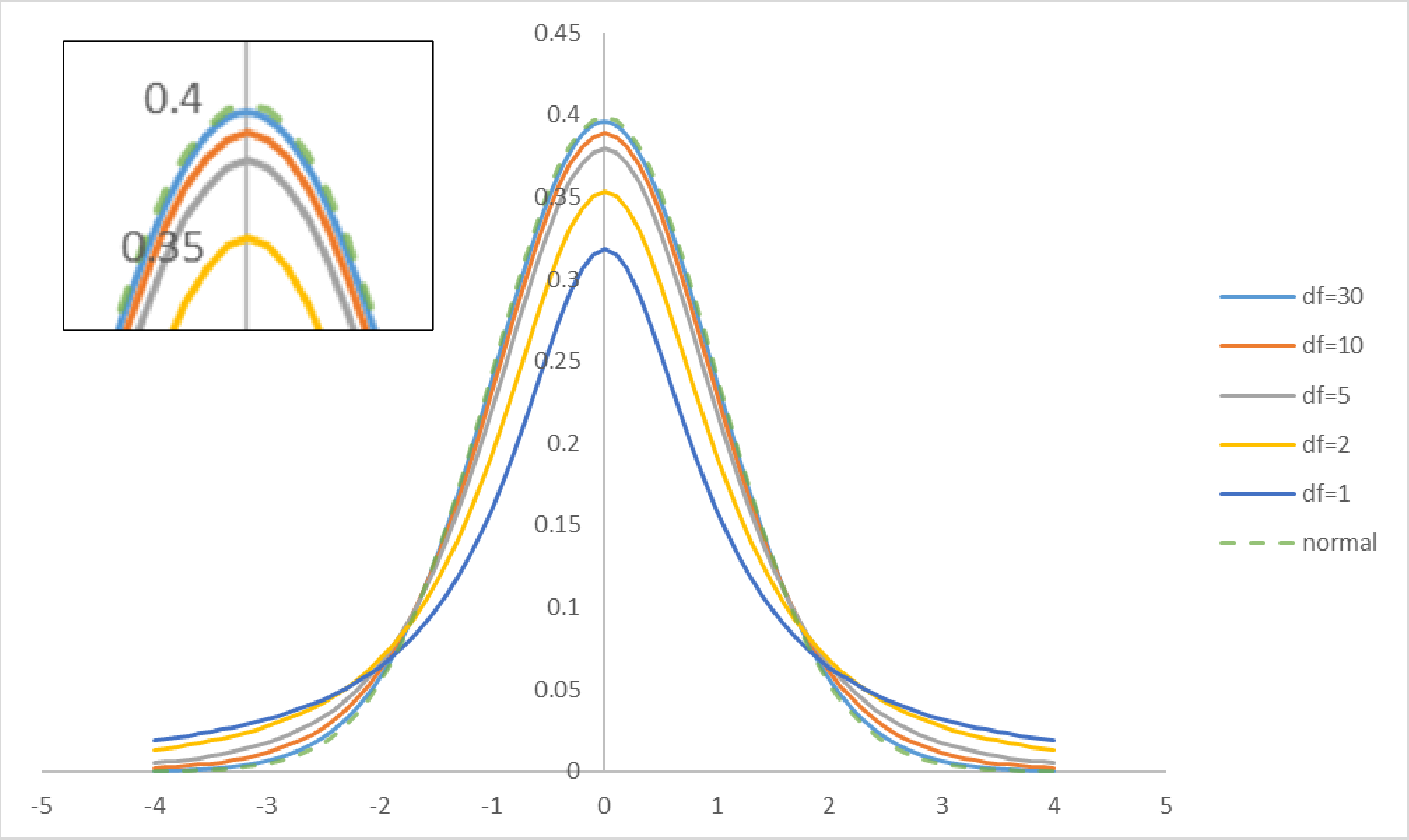
- Therefore, we understand that the larger the sample, the more confident we are that it represents the entire population
- In a large sample, we are more likely to capture the natural variation of the data in the population
- However, there is a point where increasing the sample size further offers no additional statistical benefits

$t - Distribution(s)$

- When the sample size is $n \leq 30$, or when the variance or standard deviation of the population is unknown, we use the t-distribution instead of the z-distribution (standard normal distribution)
- What does this allow us to do?;
  - The t-distribution allows us to work with small samples $(n \leq 30)$
  - However, we sacrifice some certainty in our calculations, as the margin of error increases
  - The t-distribution adjusts for the sample size by using $n - 1$ degrees of freedom. The t-distribution is different for each sample size
  - The bell curve of the t-distribution is "pulled" down in the middle and "fatter" on the ends compared to the z-distribution. The smaller the sample, the more the distribution is pulled down in the center and fatter at the tails
  - As n increases $n > 30$ and especially $n \geq 100$, the t-distribution and the z-distribution become similar and eventually hard to tell apart



The probability of having a value farther from the true mean is greater when the sample size is small
There is more uncertainty on the ends due to the small sample size
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Infiniti
10
5
2
1

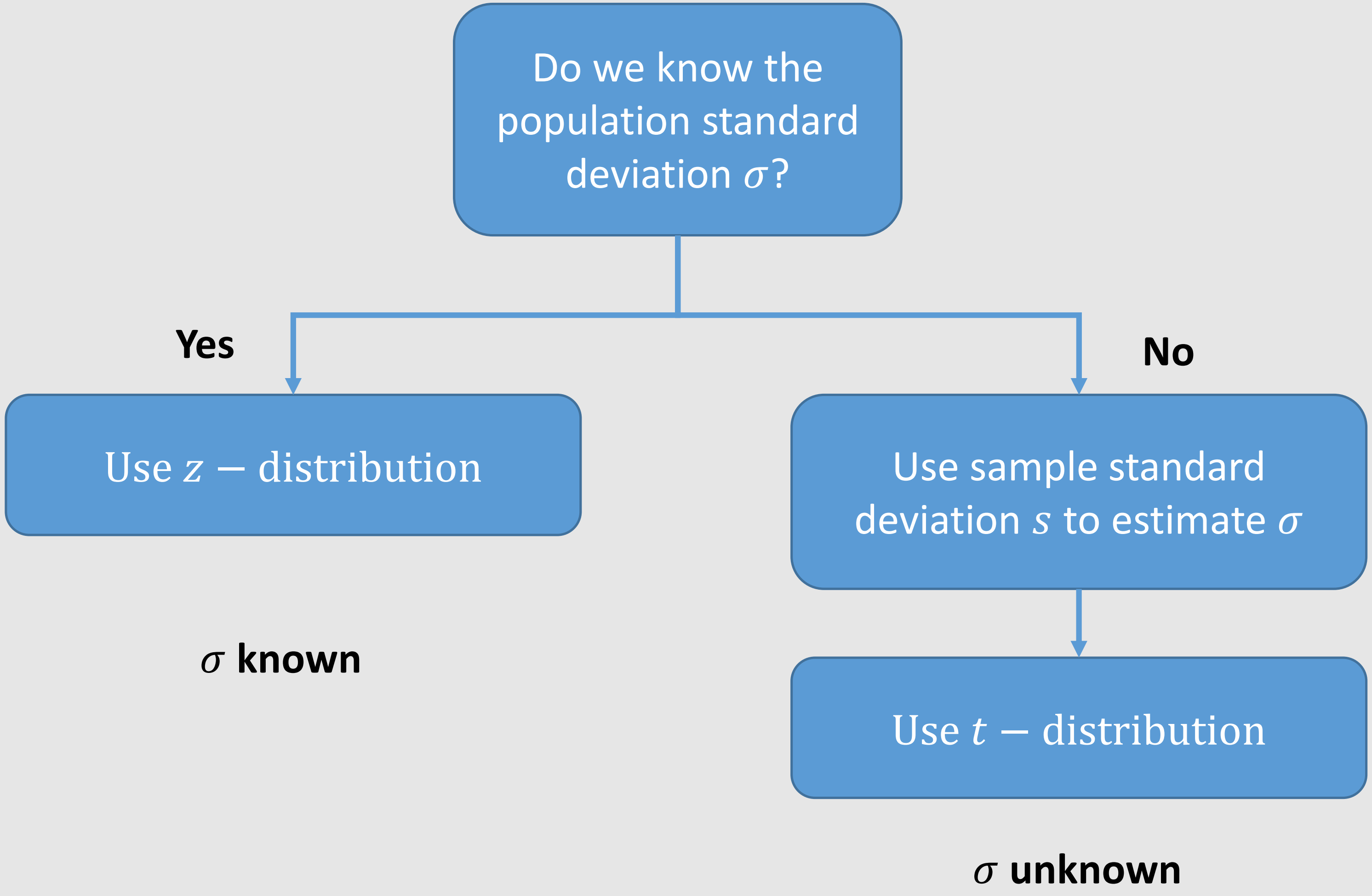


0.4
0.35
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
df=30
df=10
df=5
df=2
df=1
normal

Degree of Freedom

- Degrees of freedom are simply an adjustment to the sample size, typically $(n-1)$ or in some cases $(n-2)$ or more
- It is related to the idea that we estimate something for the larger population, most often the population variance or standard deviation
- It results in a slightly larger margin of error in the estimate
- It is a statistical adjustment

Decision Tree



Decision Tree